Order number:

Dissertation submitted to the

UNIVERSITY OF MOHAMED BOUDIAF – MSILA

FACULTY OF MATHMATICS AND COMPUTER SCIENCE
DEPARTEMENT OF COMPUTER SCIENCE

In partial fulfillment of the requirements for the degree of

Master in **Computer science**

Option : Informatique Décisionnelle et Optimisation

By

Bouti Hadil

Zerrouak Fatma Zohra

Title of the dissertation

Monitoring food safety risks using machines learning techniques

Composition of the jury

Dr. Noureddine Amraoui	University of M'sila	President
Dr. Tahar Mehenni	University of M'sila	Supervisor
Dr. Amel Meliouh	University of M'sila	Examiner

*June, **2024***

DEDICATIONS

"We would like to dedicate this dissertation to our beloved families and dear friends, who have been an unwavering source of support and motivation throughout our studies. Their steadfast presence and continuous encouragement have had a profound impact on our success. We are grateful for their enduring love and sacrifices, and we thank them for all the support and encouragement they have provided us. This dissertation would not have been possible without their presence and boundless support. We dedicate this work to them with all our love and gratitude."

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to ***Dr. Taher Mehenni***
our esteemed supervisor, for the invaluable guidance and unwavering support he provided us throughout our journey in preparing this research and writing this dissertation. His contribution has been crucial in enriching our knowledge and guiding us on the right path towards achieving our academic and research goals.

We also wish to express our appreciation to the members of the dissertation committee, ***Mr. Noureddine Amraoui*** *and Mrs.* ***Meliouh Amel****, for their tremendous efforts in reading and evaluating our research, and for providing, us with valuable and constructive feedback that helped us improve and enhance our work.*

We extend our gratitude to Mohammed Boudiaf University – M'sila, for providing us with the necessary resources and facilities that assisted us in successfully completing this research.

We express our profound gratitude and sincere thanks to everyone who contributed to the success of this work and helped us achieve our academic goals.

With love and appreciation ***Zerrouak Fatma Zohra, Bouti*** **hadil**

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Liste of Abbreviations

- **UNOCHA** - United Nations Office for the Coordination of Humanitarian Affairs
- **WFP** - World Food Programmer
- **IPCC** - Intergovernmental Panel on Climate Change
- **FAO** - Food and Agriculture Organization
- **IMF** - International Monetary Fund
- **NASA** - National Aeronautics and Space Administration
- **UNSDG** - United Nations Sustainable Development Goals
- **IFDRI** - International Food Policy Research Institute
- **NCBI** - National Center for Biotechnology Information
- **DKI APCSS** - Daniel K. Inouye Asia-Pacific Center for Security Studies
- **IFAD** - International Fund for Agricultural Development
- **PCI** - Peripheral Component Interconnect
- **DMV**-Washington, Maryland, and Virginia
- **RBF**-Radial Basis Function
- **SVM**- Support Vector Machine
- **DMV**- District of Columbia, Maryland, Virginia
- **CNN-** Convolutional Neural Network
- **LSTM-** Long Short-Term Memory

GENERAL INTRODUCTION

The food sector in Algeria and many countries around the world faces multiple and complex challenges, ranging from environmental pollution and climate change to fluctuations in production and distribution processes. These challenges increase the complexities of maintaining food safety and mitigating health and environmental risks. These issues are of utmost importance, as food is the cornerstone of human survival and health.

This study aims to develop models for predicting food crises and providing effective solutions to address them. By collecting comprehensive data encompassing a diverse range of products and influential factors, we will be able to create an analytical tool that allows us to identify potential factors that may lead to an increase or decrease in the price of a specific product, or estimate food safety risks.

This research includes several chapters, each reviewing an important aspect of analyzing and predicting food safety risks. In the first chapter, we will provide an overview of food security, explaining its importance, influencing factors, and types. We will also review some examples of food security crises and their impact on society.

In the second chapter, we will analyze and review some of the models used in predicting food crises, providing an in-depth look at usage methods, potential benefits, and expected challenges. These models will help us understand the complex dynamics that control the food market and provide reliable analytical tools for decision-making.

In the third chapter, the proposed models will be applied to real market data, and the results will be analyzed to provide practical and effective recommendations. These recommendations will help improve policies and procedures for dealing with food crises, contributing to enhanced food security.

Finally, in the fourth chapter, the results will be reviewed, and the performance of the proposed models will be evaluated, focusing on the effectiveness of these tools in predicting food safety risks and their application in reality. We will discuss how these tools can be used to enhance the ability to predict food crises and mitigate their effects, contributing to greater food sustainability.

This study aims to provide scientific and practical contributions that help improve the management of food crises, enhance food safety, and reduce associated risks, benefiting society as a whole.

CHAPITER 1

FOOD SECURITY

1. Introduction

Food security is a very important determinant of whether people can lead an active and healthy life, because it determines their access to foods required to meet nutrient needs. This reviews the definition of food security, the indicators used to measure food security depending on the level at which it is studied, how it links to nutrition and health as well as to livelihoods, what it is affected by, the consequences of food insecurity, and measures that are taken to mitigate these causes and consequences. Special attention will be paid to why it is important that food security assessments also include an estimate of the extent to which nutrient needs are being met, and the approaches and indicators, which can be used for that purpose. [1]

2. Definition

Food security is the community's ability to provide food that is safe and nutritious enough to eat. It also includes physical and economic access to food, physical and economic access to food, and the safe and proper use of food to ensure an active and healthy life. Food availability addresses the "supply side" of food security and is determined by the level of food production, stock levels and net trade. Food security also refers to achieving food stability and reducing hunger and malnutrition in the world.

⟶1974 the term food security appears at the World Food Summit:

<< Ability to supply the world with commodities at all times, to support food consumption growth, while controlling fluctuations and prices. >>

The definition focused on the term food supply; ensuring food availability and price stability at national and international level

→1983 FAO definition:

<<To ensure that every person at all times has physical and economic access to the food they need >>

The studies initiated by FAO have been based on the balance between food demand and supply.

→1986 World Bank definition improved:

The revision of the definition could add the concept of food security at the individual level. The report published by the World Bank after long studies allowed the emergence of the notion of food insecurity and the distinction between chronic and transient food insecurity. This new definition was supplemented by the notion of famine.

→1996 matching the Global Definition by the World Food Summit: Food Security means that food is available at all times, that it is accessible to all, that it is nutritionally appropriate, in quantity, quality and variety, and that they are culturally acceptable.

When all these conditions are met, and only then, can we consider that a population has achieved security food. [2]

3. Importance

Food security is the state in which all people have access to enough safe and nutritious food to meet their dietary needs for an active and healthy life, at all times. It is important because it ensures that people have access to the food they need to lead a healthy and prosperous life. Food security is linked to economic stability, long-term health, women's empowerment, and the environment. Many countries are facing the double burden of hunger and under nutrition alongside overweight and with one in three people across the globe currently suffering from some form of malnutrition. Food security faces a number of challenges across both production and consumption which research is addressing. These challenges include understanding how to re-design the food system to be healthy, sustainable, and more resilient to climate change, helping to meet both the Sustainable Development Goals and the Paris Agreement.

4. Factors

Global food supply is not even. Some places produce more food than others do. Physical factors (such as climate, soil quality and gradient) and human factors (such as technology) have historically controlled the quantity and type of food produced in any location. Today, many other factors explain why some countries produce more food than others do:

- **Climate**: Global warming is increasing temperatures by around 0.2°C every 10 years. Rainfall is increasing in some places, but decreasing in others. Higher temperatures and unreliable rainfall make farming difficult, especially for those farming marginal lands, who already struggle to survive. Even advanced countries (ACs) can be affected by drought. Countries such as Russia and Australia are huge exporters of wheat and barley respectively. When they suffer drought there is less food available globally and global food prices increase, leaving the poor most vulnerable.
- **Technology:** Improvements in technology have increased the amount of food available. Technology can overcome temperature, water and nutrient deficiencies in the form of greenhouses, irrigation and fertilizers. This can incur an economic or environmental cost. ACs import food from across the globe, all year round.
- **Loss of farmland:** The growth of the biofuel market is taking up valuable farmland which is then not used for food.
- **Pests and diseases**: Pesticides have increased crop yields. Farmers in ACs can afford pesticides, whereas most farmers in low-income developing countries (LIDCs) cannot afford them.
- **Water stress:** Irrigation systems provide water for countries with unreliable or low rainfall. Irrigation can double crop yields, but it is expensive to put these systems in place. Water can be taken either from underground aquifers or directly from rivers. Both have environmental consequences.
- **Conflict:** War forces farmers to flee their land or to fight in conflict. Food can be used as

a weapon, with enemies cutting off food supplies in order to gain ground. Crops can also be destroyed during fighting. Food shortages have caused riots and conflict. The South Sudan region has faced conflict for years, with 4 million people facing food insecurity. In the Darfur area, conflict has lasted years because of disagreement over land and grazing rights.

- **Poverty:** When people have less money, they cannot afford food and they become unable to work. Families in developing countries spend much of their income on food. [3]

5. Types

Food security can be divided into three main types, based on the availability, access and healthy utilization of food:

- **Full food security:** All individuals, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life.
- **Relative food security:** All individuals, most of the time, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life.
- **Food insecurity:** Not all individuals sometimes or always have physical and economic access to sufficient, safe and nutritious foods that meet their dietary needs and food preferences for an active and healthy life. [4]

6. Indicators of food security crises

Indicators of food security crises play a crucial role in assessing and tackling food insecurity, with key indicators including:

- **Prevalence of Food insecurity:** Measurement of the percentage of the population experiencing moderate or severe food insecurity. UNOCHA estimates reveal that approximately 811 million individuals worldwide are grappling with malnutrition. The escalation of acute hunger is becoming more pronounced in both scope and severity. The WFP issues a warning that without immediate life-saving assistance; about 41 million individuals are at risk of experiencing famine or conditions akin to famine on a global scale.

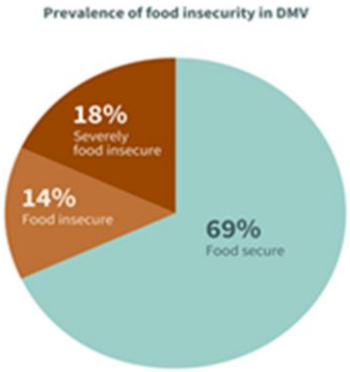

Figure 1. 1:Prevalence of Food insecurity in DMV1 *[87]*

A recent Hunger Report by the Capital Area Food Bank reveals that one in three residents, or 32%, experienced uncertainty about their next meal between May 2022 and April 2023. This statistic remains relatively consistent with the 33% reported in the 2022 survey, emphasizing the ongoing challenge of food insecurity. [5]

- **Climate change:** Increased frequency and severity of climate shocks, regional conflicts, and pandemics disrupting food production and distribution. Climate change poses a significant threat to the availability, accessibility, and quality of food. Changes in temperature, precipitation patterns, extreme weather events, and water scarcity can diminish agricultural productivity. The IPCC underscores the heightened risks to food security due to climate change, particularly affecting vulnerable countries and populations. [6]
- **Price Volatility and Trade Disruptions:** Sudden spikes in food prices, export restrictions, and trade disruptions indicating a worsening food security situation. Price volatility has a substantial impact on food security, influencing household incomes and purchasing power. The welfare consequences for consumers are exacerbated by higher prices, while producers stand to benefit. Effectively addressing both price spikes and volatility is crucial for overall welfare [7]. IMF estimates project that higher import costs for food and fertilizer will impose an additional $9 billion [8] in balance of payments pressures on highly exposed countries in 2022 and 2023.

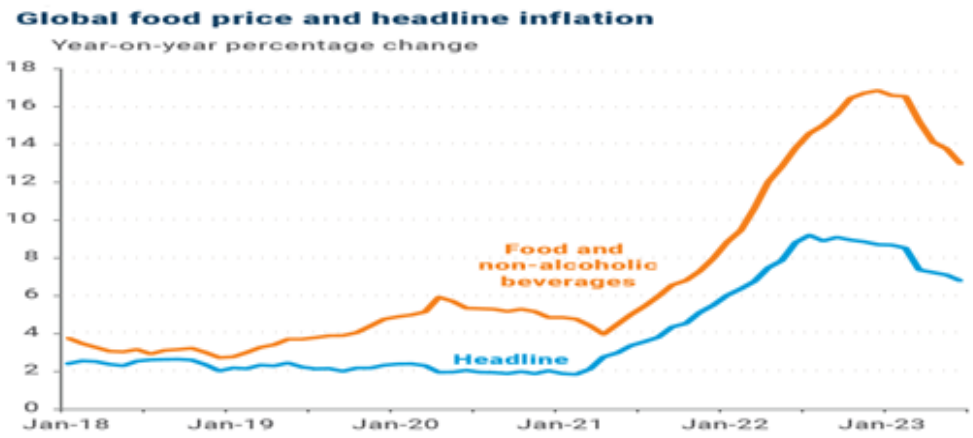

Figure 1. 2: Global food Price and headline inflation *[9]*

- **Armed Conflicts:** Impact of armed conflicts on food security, especially in countries heavily reliant on imports from conflict zones. The global repercussions of food shocks are evident, with 48 countries, many of which heavily depend on imports from conflict-affected areas like Ukraine and Russia, facing severe economic challenges. The World Bank notes a substantial increase in acute food insecurity from 135 million in 2019 to 345 million in 82 countries by June 2022, driven by the war in Ukraine, disruptions in supply chains, and the ongoing repercussions of the COVID-19 pandemic. [10]

7. Global Crisis Chronicles

- **Drought in Somalia (2010-2012):** The drought in Somalia from 2010 to 2012 was one of the most severe in the region's history, leading to a devastating famine and significant loss of life. [11]
- **Floods in Pakistan (2010):** The floods in Pakistan in 2010 were caused by extremely high rainfall in the Indus River watershed during July and August [12]. The floods affected approximately 20 million people, destroyed homes, crops, and infrastructure, leaving millions vulnerable to malnutrition and waterborne diseases. [13]
- **Locust Plague in East Africa (2020):** A plague of locusts descended on East Africa in 2020, devastating crops, trees, and pasture as they moved. [14]

Figure 1. 3: Locust Plague in East Africa [14]

- **War in Syria (2011 - present):** The war in Syria erupted in 2011, resulting in widespread destruction of infrastructure and large-scale displacement of populations. This conflict has had a significant impact on humanitarian and security conditions in Syria and neighboring regions. [15]
- **COVID-19 Pandemic (Starting from 2019):** The COVID-19 pandemic has significantly affected food security worldwide. Millions of people face hunger and malnutrition due to reduced incomes and disrupted food supply chains. Global food insecurity has increased in almost every country, with the number of severely food-insecure people doubling from before the pandemic to 276 million. [16]
- **the Russia-Ukraine Conflict:** The Russia-Ukraine conflict has had a significant impact on global food security. Ukraine, known as the "breadbasket of Europe," is a major grain exporter, and the conflict has raised fears of a global food crisis, further exacerbating existing food security challenges worldwide. [17]

8. Risks

The food system encompasses various risks, including:

- **Environmental Impact**
  - ✓ The food system has significantly contributed to environmental challenges such as climate change, environmental degradation, natural resource overexploitation, and air, water, and soil pollution. [6]
  - ✓ Key environmental impacts of food production include the sector accounting for 26% of global greenhouse gas emissions, with half of habitable land worldwide dedicated to agriculture. [18]
  - ✓ The U.S. food system has unintentionally resulted in environmental issues like pollution, greenhouse gas emissions, and water quality deterioration. [19]
  - ✓ Industrial agriculture contributes to environmental harm through air, soil, and water pollution, with livestock emissions responsible for 14.5% of global greenhouse gas emissions. [20]
  - ✓ Unsustainable food production and consumption practices have led to environmental degradation and biodiversity loss, emphasizing the urgency for sustainable food systems. [21]
- **Health Effects**
  - ✓ Food insecurity can result in various adverse health effects, including increased illness, reduced productivity, and the development of chronic conditions.
  - ✓ The impact of food insecurity on health extends to both physical and mental well-being, leading to conditions such as cardiovascular disease, type 2 diabetes, cancer, and obesity. [22]
  - ✓ Additionally, food insecurity is associated with impaired learning for both children and adults, increased healthcare needs, and decreased productivity. [23]
- **Social and Economic Consequences**
  - ✓ Food insecurity is linked to low wages, adverse social and economic conditions, limited access to healthy foods, and residential segregation, contributing to socioeconomic disparities in health and well-being. [24]
  - ✓ The complexity of food insecurity involves economic root causes influenced by factors such as low wages, adverse social and economic conditions, and limited access to healthy foods, residential segregation, and a lack of affordable housing. [24]
  - ✓ Notably, food insecurity is not exclusive to impoverished, unhealthy, and socially isolated families; many reporting it have incomes above the official poverty line. [23]
  - ✓ The extensive social and economic implications of food insecurity affect overall health, productivity, and community well-being. [25]

- **Political Influence**
  - ✓ The impact of political risk and institutions on food security is substantial, with government stability, the rule of law, investment profile, and democratic accountability affecting food supply and security.
  - ✓ Political instability, weak institutions, and poor governance can result in high-level corruption, rapidly decreasing food security.
  - ✓ Food accessibility is more dependent on political, economic, and social factors than on availability.
  - ✓ Political risk and environmental degradation can adversely affect food security.
  - ✓ Addressing food insecurity requires comprehensive strategies considering political, economic, social, and environmental dimensions, promoting good governance, strengthening institutions, and ensuring democratic accountability to improve food security. [26] [27]

9. Ways to Prevent Food Insecurity

- **Invest in Robust Food Storage Systems:** Allocate resources to develop resilient food storage systems capable of withstanding extreme weather events.
- **Diversify Food Sources and Agricultural Techniques:** Mitigate risk by diversifying both food sources and agricultural production techniques.
- **Adopt Water Management Systems:** Implement systems for water management to reduce crop damage caused by floods or droughts.
- **Promote Sustainable Farming Practices:** Advocate for sustainable farming methods, including no-till agriculture, agroforestry, and cover crops.
- **Support Smallholder Farmer:** Empower smallholder farmers through access to credit and essential services, fostering economic empowerment.
- **Raise Public Awareness:** Increase awareness among the public about the challenges to food security posed by climate change**.**
- **Enhance Soil Resilience:** Increase organic carbon in the soil to enhance water retention and boost resilience to drought.
- **Promote Education on Food Preservation:** Educate communities on effective food preservation techniques, such as refrigeration and dehydration.
- **Develop Early Warning Systems:** Establish early warning systems for extreme weather events, utilizing technologies like data analytics, insights, and predictive AI to facilitate adaptive food production.
- **Invest in Research and Development:** Allocate resources for research and development focused on creating climate-resilient food crops. [28]

10. Global Organizations

Prominent international organizations dedicated to advancing food security include. These organizations play pivotal roles in addressing global food security challenges.

- **World Food Programmer (WFP):** The UN's food-assistance branch, focused on alleviating hunger and promoting food security.

Figure 1. 4: WFP

- **Food and Agricultural Organization (FAO):** A specialized UN agency leading global initiatives to combat hunger.

Figure 1. 5:FAO

- **International Fund for Agricultural Development (IFAD):** Another UN specialized agency, addressing food insecurity and poverty in rural areas.

Figure 1. 6: IFAD

- **World Bank :** Providing financial and technical assistance to support nations in developing agriculture and food security programs.

Figure 1. 7: The World Bank

- **Care:** A global organization committed to eliminating poverty and upholding human dignity, striving for a world where poverty is overcome, and everyone lives with dignity and security**.**

Figure 1. 8: Care

- **Feeding American:** Leading the fight against hunger in the United States, with a mission to feed the nation's hungry through a nationwide network of member food banks.

Figure 1. 9: Feeding American

- **UNICEF:** One of the largest UN agencies dedicated to supporting children in need worldwide. UNICEF is actively involved in nutrition programs and is part of the Scaling Up Nutrition initiative, a global effort focusing on nutrition in numerous countries.

Figure 1. 10: Unicef

- **Project Concern International (PCI):** A global development organization employing innovation to address hunger, improve health, overcome hardships, and empower women and girls across Asia, Africa, and the Americas. PCI has positively affected the lives of nearly ten million people.

Figure 1. 11: PCI *[29] [30]*

11. Conclusion

Food security is a vital issue that directly affects the health and well-being of individuals and communities. This chapter addressed the definition of food security, its importance, the factors affecting it, and global food crisis indicators, along with examples of severe food crises. Achieving sustainable food security requires international cooperation, technological advancements in agriculture, and promoting sustainability to face increasing challenges.

In the next chapter, we will discuss the Machine Learning models and how they can be used for predicting and analyzing data to improve food security and enhance the ability to forecast future crises.

CHAPTER 2

PREDICTIVE MODELING

1. INTRODUCTION

Predictive modeling involves using statistical techniques and machine learning algorithms to analyze historical data and make forecasts about future outcomes. It is widely applied in various fields such as finance, healthcare, and marketing to identify patterns and predict trends. By leveraging these models, organizations can make informed decisions and improve strategic planning.

This chapter offers an in-depth analysis of various models, thoroughly defining each one and reviewing the different types and dimensions they encompass. Emphasis is placed on explaining the role of each model in the analysis and prediction process. By exploring the algorithms and techniques used in each model, the chapter aims to provide readers with a comprehensive understanding of their characteristics and applications, guiding them in selecting the most suitable model for specific research objectives. [31]

2. Linear Regression

2.1. Definition

Linear regression is a statistical algorithm employed in data science and machine learning to establish a linear connection between an independent variable (predictor) and a dependent variable (outcome). In this supervised learning approach, the predictor remains constant amid changes in other variables, influencing fluctuations in the dependent variable.

The regression model, at its core, predicts the value of the dependent variable, offering insights into various fields such as stock market forecasting, portfolio management, and scientific analysis. It particularly excels in scenarios with at least two variables, providing valuable predictions for continuous or numeric variables like sales, salary, age, or product prices.

Visualized as a sloped straight line (Fig. 2.1) , the linear regression model serves as a powerful tool for predictive analysis, enhancing our understanding of relationships between variables and facilitating predictions for future events.

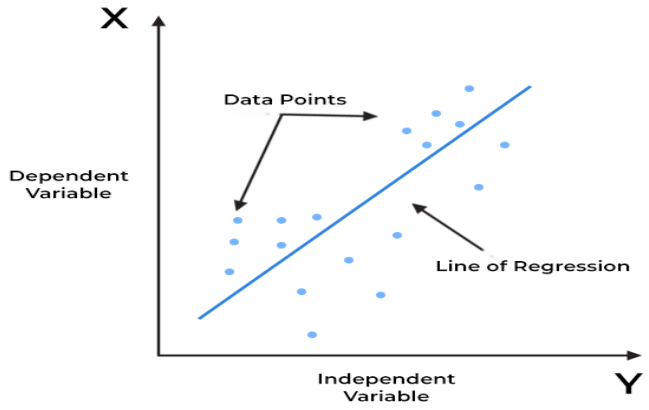

Figure 2. 1 :Best-Fit Line for a Linear Regression Model *[32]*

2.2. Types

Linear Regression is generally classified into two types:

2.2.1. Simple Linear Regression

Finding the link between a single independent variable (input) and a matching dependent variable (output) is the goal of simple linear regression. A straight line can be used to represent this.

One possible rewrite of the same line equation is:

$$Y = \beta o + \beta_1 X + €.$$

- Y. denotes the dependent variable, or output.
- The unknown constants ß0 and 31, respectively represent the intercept and coefficient (slope).
- The error term is ε (Epsilon).

An example of a simple linear regression model's graph is shown in Fig. 2.2:

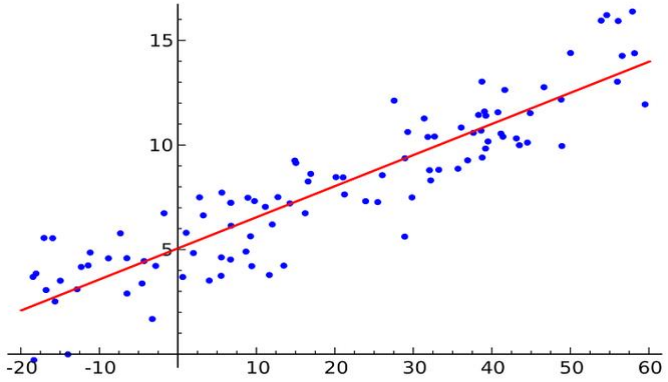

Figure 2. 2: Graph of Simple Linear Regression Model

2.2.2. Applications of Simple Linear Regression include

- Predicting crop yields based on the amount of rainfall: Yield is dependent variable while the amount of rainfall is independent variable.

- The student's mark according to the number of hours studied (ideally): In this case, the amount of study hours is independent and the grades obtained are dependent.
- Estimating an individual's salary based on years of experience: this makes experience an independent variable and salary a dependent variable. [33]

2.2.3. Polynomial Regression

Polynomial Regression aims to model the relationship between the independent variable X and the dependent variable Y as an n degree polynomial. Its general equation is of the format:

$Y=\beta0+\beta1X+\beta2X2+\ldots+\beta nXn+e.$

Where:

- $\beta0, \beta1\ldots, \beta n$ are the coefficients
- X is the independent variable
- Y is the dependent variable
- e is the error term

The goal of Polynomial Regression is to find the best estimates for the coefficients using the method of least squares. [34]

2.3. Properties

The following characteristics apply to the regression line where the regression parameters $\beta0$ and $\beta1$ are defined

- The regression line decreases the total of squared discrepancies between observed and predicted values
- The regression line crosses the mean value of the X and Y variables
- The linear regression's y-intercept equals the regression constant $\beta0$
- The slope of the regression line is the regression coefficient, or b_1. For every unit change in the independent variable (X), its value is equal to the average change in the dependent variable (Y) [35]

2.4. Importance

A regression line is used to describe the behavior of a set of data, a logical approach that helps us study and analyze the relationship between two different continuous variables. Which is then enacted in machine learning models, mathematical analysis, statistics field, forecasting sectors, and other such quantitative applications. Looking at the financial sector, where financial analysts use linear regression to predict stock prices and commodity prices and perform various stock valuations for different securities. Several well-renowned companies make use of linear regressions for the purpose of predicting sales, inventories, etc. [35]

2.5. Algorithm

Linear Regression Algorithm Implementation Steps:

1. Data Reading and Understanding:
    - Import essential libraries like pandas, numpy, seaborne, and matplotlib.
    - Clean and manipulate the data, addressing null values, updating formats, changing data types, and removing unwanted rows or columns.
2. Exploratory Data Analysis (EDA)
    - Visualize numerical variables using scatter or pair plots for meaningful

business/domain insights.
    - Employ bar plots or boxplots for visualizing categorical variables and extracting insights.
3. Data Preparation
    - Convert categorical variables into dummy variables, ensuring numerical representation for model building.
4. Data Splitting
    - Divide the data into training and test sets, typically with a 70:30 or 80:20 ratios.
    - Rescale the model to normalize numerical variable ranges, addressing varying magnitudes.
5. Linear Model Building
    - Explore feature selection methods like Forward Selection, Backward Selection, or Recursive Feature Elimination (RFE).
    - Construct a linear model using selected features to establish a relationship with the target variable.
6. Residual Analysis
    - Evaluate residuals on the training data to assess the distribution of errors.
    - A well-centered mean around 0 indicates a good residual analysis.
7. Prediction and Evaluation
    - Predict outcomes on the test dataset by applying the trained model.
    - Divide the test set into X_test and y_test and calculate the **r2_score**.
        - Aim for a similar **r2_score** between the train and test sets, with an acceptable difference of 2–3%. [36]

2.6. Advantages

- Easy to interpret: The coefficients of a linear regression model represent the change in the dependent variable for a one-unit change in the independent variable, making it simple to comprehend the relationship between the variables.
- Robust to outliers: Linear regression is relatively robust to outliers meaning it is less affected by extreme values of the independent variable compared to other statistical methods.
- Can handle both linear and nonlinear relationships: Linear regression can be used to model both linear and nonlinear relationships between variables. This is because the independent variable can be transformed before it is used in the model.
- No need for feature scaling or transformation: Unlike some machine learning algorithms, linear regression does not require feature scaling or transformation. This can be a significant advantage, especially when dealing with large datasets. [36]

2.7. Disadvantages

- Assumes linearity: Linear regression assumes that the relationship between the independent variable and the dependent variable is linear. This assumption may not be valid for all data sets. In cases where the relationship is nonlinear, linear regression may not be a good choice.

- Sensitive to multicollinearity: Linear regression is sensitive to multicollinearity. This occurs when there is a high correlation between the independent variables. Multicollinearity can make it difficult to interpret the coefficients of the model and can lead to overfitting.
- May not be suitable for highly complex relationships: Linear regression may not be suitable for modeling highly complex relationships between variables. For example, it may not be able to model relationships that include interactions between the independent variables.
- Not suitable for classification tasks: Linear regression is a regression algorithm and is not suitable for classification tasks, which involve predicting a categorical variable rather than a continuous variable [31]

3. Decision Trees

3.1. Definition

A decision tree presents a flowchart-style structure where internal nodes represent features, branches depict rules, and leaf nodes indicate the algorithm's outcomes. This versatile supervised machine-learning algorithm serves for both classification and regression tasks, showcasing its robustness. Additionally, it plays a pivotal role in Random Forest, contributing to its potency by training on diverse subsets of data, thereby establishing Random Forest as one of the most formidable algorithms in machine learning [37]. AN example of Decision tree model is shown in Fig. 2.3

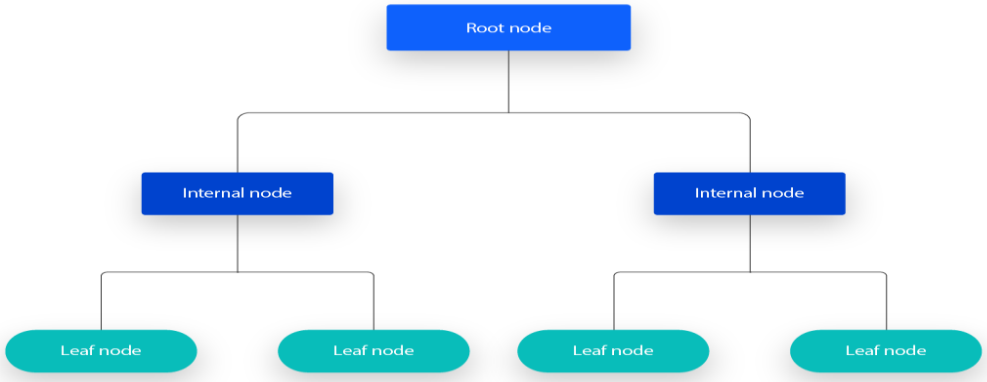

Figure 2. 3:Example of Decision Tree model [38]

3.2. Types

There are two main types of decision trees: classification trees and regression trees. These branches further split into various algorithms, utilizing nodes and branches for comprehensive decision-making. It is crucial to select the one aligning with the specific goal.

Selecting the appropriate type is straightforward. The example questions provided for each type can guide analysts in determining whether their objective is best suited for a classification tree, dealing with "yes" or "no" questions, or a regression tree, focused on predicting continuous values. It is nonetheless that visualizing the decision tree through diagrams can enhance communication.

- **Classification Trees:** ideal for real-world problems with binary outcomes, such as food safety having two outcomes Yes or No.
- **Regression Trees:** designed to predict continuous values, regression trees are constructed from historical data. Examples include forecasting products sales for the next quarter of the year.

3.3. Decision tree nodes: Decision trees are made of a few simple parts, which can be applied, to any type of tree as well as all of the algorithms.

- *Decision tree root node:* Root nodes (also known as parent nodes, meaning they have nodes under them in a parent-child structure) are the highest level of a decision tree, a starting point from which the rest of the tree grows. This node represents the question, task or problem the tree stems from.
- *Decision tree internal node:* An internal node inside of a decision tree in which the preceding node branches out into two or more variables.
- *Decision tree leaf node:* Also known as external nodes or terminal nodes, these are the endpoints and have no child nodes. It is always found furthest from the root node and is where the answer or solution is found.
- *Decision tree pruning:* Pruning is the process of slimming down variables by removing nodes leaving only the most critical nodes and potential outcomes.
- *Decision tree splitting:* The opposite of pruning — this divides nodes into two or more variables in the system.
- *Decision tree sub-tree or branch:* This is a specific section of a decision tree. It contains multiple internal nodes and potentially some leaf nodes depending on the specific branch in question. [39]

3.4. Attribute selection measures

When a dataset contains N attributes, choosing which attribute to place at the root node or as an internal node at different tree levels can be complex. Even randomly selecting a node as the root does not solve the problem. Using random methods also produces poor results with very low accuracy. To solve this attribute selection problem, we need to apply the following solution:

- entropy
- information gain
- Gini index
- gain ratio
- reduce variance

These solutions help in calculating the value of each attribute. We can sort the values and put them into a tree, with higher values leading to the root node and lower values leading to child nodes. Note that when using information gain we must treat the attributes as categorical. On the other hand, if we use the Gini index, we should consider it continuous.

3.4.1. Entropy

Entropy is a measure of the randomness of processing information. The higher the entropy, the more difficult it is to solve that information. For example, when we toss a coin, we cannot be sure

of the outcome. All we do is perform random operations that lead to random results.

In ID3, branches with zero entropy are called leaf nodes. Those with entropy greater than zero must be divided. The formula for attribute entropy is:

$$E(s) = \sum_{i=1}^{e} -p_i \, log_2 \, p_i \tag{1}$$

Here P_i is the probability of event i in state S.

3.4.2. Information gain

Information gain is a statistical characteristic that measures how well an attribute divides training instances according to target type. When building a decision tree, the attributes that provide the best information gain and lowest entropy are found.

Information gain is a reduction in entropy. It calculates the difference between the entropy of the records before splitting and the average entropy of the records after splitting based on the specified attribute value. The formula is as follows:

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{K} Entropy(j, after) \tag{2}$$

Where before is the data set before splitting, (j, after) is the subset j after splitting, and K is the set of subsets created by splitting.

3.4.3. Gini index

The Gini index is a measure of purity or impurity used in building decision trees in the CART algorithm. Properties with a lower Gini index can only be compared with properties with a higher Gini index. Indexes can only create binary splits, which are used by the CART algorithm to create the same splits.

We can use a cost function to evaluate segmentation in a data set given using the Gini index. We can calculate this by subtracting the sum of squared probabilities for each category from 1. It favors large partitions and is very easy to implement, but IG will win fewer partitions with unique values. The Gini index calculation formula is as follows:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{3}$$

3.4.4. Gain ratio

Selecting attributes with higher values as root nodes affects information gain. It prefers properties with higher and unique value. C4.5 is an evolution of ID3, and the gain ratio is an information gain modification that reduces the impact, making it the best choice.

Gain ratio solves the problem of exploiting information gain by taking into account the number of branches that occurred before the split. It corrects the information gain by preserving the inherent information of the split.

The gain ratio is calculated as follows:

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j log_2 w_j} \quad (4)$$

Where before is the record before splitting. (j, after) is the subset adjacent to the partition, and K is the set of subsets resulting from the partition.

3.4.5. Reduce variance

Variance reduction is an algorithm used in regression problems or continuous target variables. It uses the usual variance formula to select perfect splits. The split with lower variance is the criterion for splitting the population. The formula is:

$$Variance = \frac{\sum(X - \bar{X})^2}{n} \quad (5)$$

Where X-bar is the average of the values, n is the total value, and X is the actual value. [40]

3.5. The role of decision trees in data science

We have mostly focused on the use of decision trees in choosing the most effective course of action in business, but this type of informational mapping also has practical applications in data mining and machine learning.

In this context, decision trees are not used to manually determine some optimal course of action, but rather as a predictive model to automatically make observations about a given dataset. These algorithms take in enormous amounts of information and use a decision tree to derive accurate predictions about new data points. For example, consider using the medical data of thousands of hospital patients to predict the likelihood of a person developing a disease. [37]

3.6. Algorithm

Creating a decision tree algorithm involves the following steps:

- Data Pre-processing step.
- Fitting a Decision-Tree algorithm to the Training set.
- Predicting the test result.
- Test accuracy of the result (Creation of Confusion matrix).
- Visualizing the test set result. [40]

3.7. Advantages

Decision trees hold several advantages for data scientists.

- *Interpretability*: Their interpretability stands out as a key strength, presenting model rules in a flow chart-like manner for clear communication with stakeholders. This transparency builds confidence and provides detailed insights into predictions.
- *Less Data Preparation:* A notable benefit is the minimal need for data preparation. Unlike other algorithms, decision trees do not necessitate steps like normalization or outlier treatment, simplifying the modeling process and making it a preferred choice for data scientists.

- *Non-Parametric*: Being a non-parametric algorithm distinguishes decision trees from others like linear regression or naïve Bayes. There are fewer assumptions to fulfill, offering flexibility and ease of implementation.
- *Versatility*: Versatility is another asset, as decision trees excel not only in predictions but also in data exploration and as a baseline model for assessing data quality. They adeptly handle both regression and classification problems, and their variants can address segmentation challenges.
- *Non-Linearity*: The ability to create complex decision boundaries allows decision trees to tackle non-linear problems effectively. This distinctive feature, coupled with interpretability, sets them apart in the realm of algorithmic solutions. [41]

3.8. Disadvantages

Decision trees come with notable drawbacks that may limit their utility in certain scenarios. Here are key disadvantages:

- *Overfitting*: Decision trees exhibit a high variance, making them prone to overfitting. Their lack of an inherent stopping mechanism can lead to complex decision rules. While tuning parameters or pruning can mitigate this, their effectiveness might be constrained.
- *Feature Reduction & Data Resampling*: The training phase of a decision tree can be time-consuming, especially with numerous continuous independent variables. Imbalanced class datasets can bias the model toward the majority class. Addressing these issues involves reducing features to streamline time complexity and potentially duplicating or removing rows to handle class imbalances.
- *Optimization*: Decision trees operate with a greedy approach, seeking pure nodes at each level without considering the broader impact on subsequent splits. This heuristic nature enhances interpretability but does not guarantee globally optimal results. Variables with high significance or causing data leakage can influence the process, and while using an ensemble of decision trees can address some issues, it sacrifices interpretability. [41]

4. Random Forests

4.1. Definition

Random Forests, also known as Random Decision Forests, constitute an ensemble learning technique employed for classification, regression, and various tasks. This algorithm constructs multiple decision trees during training, and in classification, the final output is determined by the majority vote of the trees. For regression, it is the mean or average prediction of individual trees. Characterized by the creation of numerous small decision trees termed estimators, Random Forests generate predictions, combining their outputs to enhance overall accuracy.

Initially developed in 1995 by Tin Kam Ho using the random subspace method, the algorithm was later expanded by Leo Breiman and Adele Cutler. Random Forests find wide applications in tasks involving large datasets with high dimensionality and diverse feature types, excelling in mitigating overfitting and demonstrating robust generalization to unseen data, even with missing values.

The term "forest" refers to an ensemble of decision trees, typically trained using the bagging

method, addressing the tendency of decision trees to over fit their training set. [42] [31]

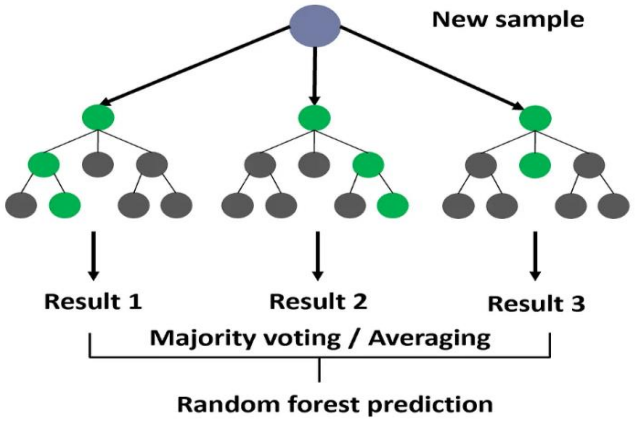

Figure 2. 4: Example of Random Forests

Hyper parameters in Random Forests are used to enhance performance and predictive power or to make the model faster. Key hyper parameters include:

- *n_estimators*: Number of trees built by the algorithm before averaging the products.
- max_features: Maximum number of features the random forest uses before considering splitting a node.
- *min_sample_leaf:* Determines the minimum number of leaves required to split an internal node.
- *n_jobs:* Conveys how many processors are allowed to use. If the value is 1, it can use only one processor, but if the value is -1, there is no limit.
- *random_state:* Controls randomness of the sample, ensuring consistent results with the same hyper parameters and training data.
- *oob_score:* Out of Bag (OOB) is a random forest cross-validation method, using one-third of the sample to evaluate performance. [43]

4.2. Random Forest Applications

The random forest algorithm finds applications across various domains, for instance:

- *Finance***:** Preferred over other algorithms for evaluating credit risk, detecting fraud, and options pricing.
- *Healthcare:* Applied in computational biology for gene expression profiling, biomarker discovery, and sequence annotation, enabling estimates about drug responses to specific medications.
- *E-commerce:* Used in recommendation engines for cross-selling purposes [38]
- *Food safety:* Random Forest can be used to forecast food safety risks and enhance effectively early warning systems.

4.3. Important Terms

Understanding the Random Forest algorithm involves knowing key terms:

- *Entropy:* A measure of randomness or unpredictability in the dataset.
- *Information Gain:* A measure of the decrease in entropy after splitting the dataset.
- *Leaf Node:* Carries the classification or decision.

- *Decision Node:* A node with two or more branches.
- *Root Node:* The topmost decision node where all data is initially. [44]

4.4. Algorithm

The Random Forest algorithm is explained through the formula:

Random forest = tree bagging + feature sampling.

4.4.1. Tree Bagging: Bagging, short for "bootstrap aggregation," involves three main steps:

- Construction of n decision trees by randomly selecting n observation samples.
- Training each decision tree independently.
- Making predictions on new data by applying each of the n trees and taking the majority decision among the forecasts.

4.4.2. Feature Sampling

- Feature sampling is a process of randomly selecting variables (data columns), typically drawing Root n variables for a problem with n total variables.
- In the context of credit acceptance, each bank studies a loan application with limited access to customer information. Different banks may base their decisions on subsets of information, such as age, occupation, and annual income for one, while another may consider marital status, gender, and outstanding loan count.
- Feature sampling helps lower the correlation between trees, reducing the potential disruption in result quality. In statistical terms, it is said to reduce the variance of the created set. [45]

4.5. Advantages

Random Forests offer several advantages, differentiating them from other machine learning algorithms:

- Versatility in Handling Data Types: Proficiently manages both categorical and numerical data without the need for scaling or variable transformation.
- Implicit Feature Selection: Conducts feature selection implicitly, generating uncorrelated decision trees, advantageous for datasets with a substantial number of features.
- Robustness to Outliers: Demonstrates resilience to outliers by binning variables during the process.
- Handling Linear and Non-Linear Relationships: Adeptly manages both linear and non-linear relationships within the data.
- High Accuracy and Balanced Bias-Variance Trade-Off: Typically provides high accuracy while effectively balancing the bias-variance trade-off.
- Compensation for Overfitting: Excels in compensating for overfitting and exhibits robust generalization to unseen data, even with missing values.
- Flexibility and Ease of Use: Known for flexibility, ease of use, and suitability for both classification and regression tasks. [46] [47] [48]

4.6. Disadvantages

While Random Forests offer numerous advantages, it's crucial to consider their drawbacks:

- Computational Intensity for Large Datasets: Can be computationally intensive for extensive datasets, potentially resulting in slower performance.
- Lack of Interpretability: Models are not easily interpretable, posing challenges in understanding how the model reaches its predictions.
- Performance Issues on Small or Low-Dimensional Datasets: May not perform well on small or low-dimensional datasets as the randomness factor becomes significantly reduced.
- Potential for Overfitting with Noisy Data: Might over fit when dealing with data containing a high level of noise, although this risk can be mitigated through voting mechanisms.
- Black Box Nature with Limited Control: Operates like a black box, providing little control or insight into the inner workings of the model. [48] [49]

4.7. Random Forest vs. Decision Tree: The Table 2.1 illustrates the main differences between Random Forests and Decision Trees.

Aspect	Random Forest	Decision Tree
Nature	Ensemble of multiple decision trees	Single decision tree
Bias-Variance Trade-off	Lower variance, reduced overfitting	Higher variance, prone to overfitting
Predictive Accuracy	Generally higher due to ensemble	Prone to overfitting, may vary
Robustness	More robust to outliers and noise	Sensitive to outliers and noise
Training Time	Slower due to multiple tree construction	Faster as it builds a single tree
Interpretability	Less interpretable due to ensemble	More interpretable as a single tree
Feature Importance	Provides feature importance scores	Provides feature importance, but less reliable
Usage	Suitable for complex tasks, high-dimensional data	Simple tasks, easy interpretation

Table 2. 1: Overview of Random Forest vs Decision Tree *[50]*

5. Support Vector Machines (SVM)

5.1. Definition

A Support Vector Machine (SVM) is a machine-learning algorithm with applications in classification, regression, and outlier detection. It identifies an optimal line or decision boundary to separate classes, maximizing the margin between the hyperplane and the nearest data points of each category. SVMs excel in binary classification and can handle non-linear scenarios through the kernel trick, implicitly mapping inputs into higher-dimensional feature spaces.

Introduced in 1963 by Vladimir N. Vapnik and Alexey Ya. Chervonenkis, SVMs gained popularity in text categorization, image classification, and biological sciences. They extend to regression tasks, predicting continuous values with an optimal hyperplane and maintaining a tolerance margin. SVMs handle various data separation types, employing different kernel functions such as linear, polynomial, or radial basis function (RBF) kernels.

Despite computational costs and sensitivity to parameter tuning, SVMs remain popular for

their adaptability, flexibility, and effectiveness in real-world scenarios involving higher-dimensional spaces [51] [31]. Fig. 2.5 shows an example of SVM Model.

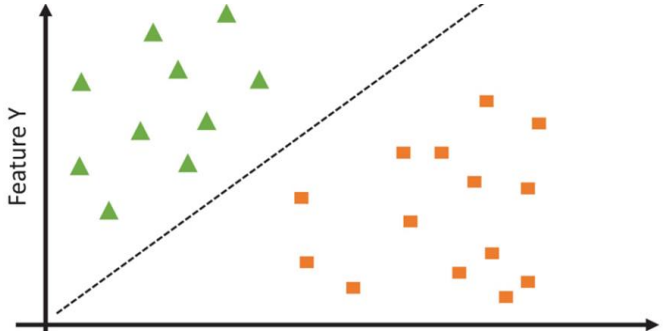

Figure 2. 5: Diagram-depicting-SVM-example-with-hyperplane-for-classification-problem *[52]*

5.2. Support Vector Machine (SVM) Terminology

- **Hyperplane:** The hyperplane serves as the decision boundary, separating data points of different classes within a feature space. In linear classifications, it is represented by the equation wx + b = 0. Optimal hyperplanes are selected based on maximizing the margin, which is the distance from the hyperplane to the nearest data point on each side.

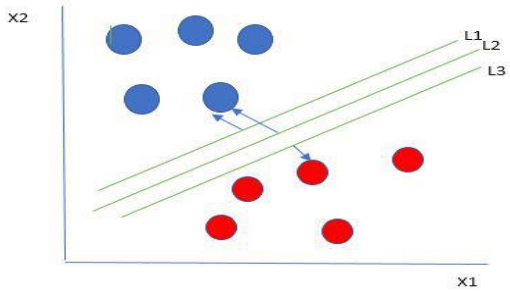

Figure 2. 6: Multiple hyperplanes in SVM separating data in two classes

- **Support Vectors:** These are the nearest data points to the hyperplane, playing a crucial role in determining the hyperplane and margin.

- **Margin:** The margin is the distance between the support vector and the hyperplane. The primary objective of the SVM algorithm is to maximize this margin, and a wider margin indicates better classification performance.

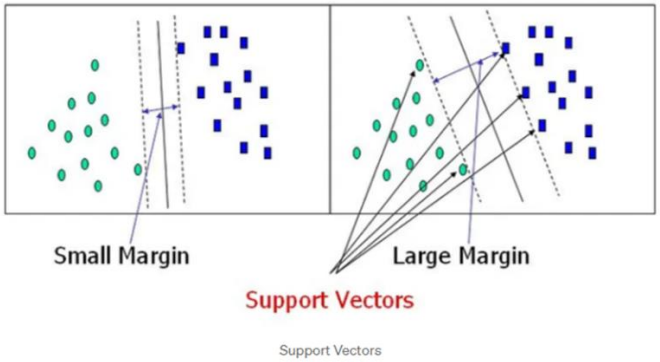

Figure 2. 7: Margin in SVM

- **Kernel:** Kernels are mathematical functions used in SVM to map original input data points into high-dimensional feature spaces. This transformation facilitates finding the hyperplane, even when data points are not linearly separable in the original input space. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid.

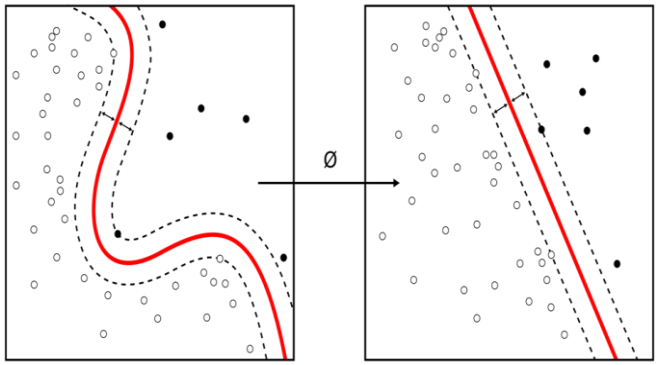

Figure 2. 8: Kernel machine in SVM

- **Hard Margin:** The hard margin hyperplane is the maximum-margin hyperplane that perfectly separates data points of different categories without any misclassifications.
- **Soft Margin:** In cases where data is not perfectly separable or contains outliers, SVM introduces a soft margin technique. Each data point has a slack variable, softening the strict margin requirement and allowing for certain misclassifications or violations. This approach finds a compromise between increasing the margin and reducing violations.
- Parameter **C:** The regularization parameter C in SVM balances margin maximization and misclassification fines. It determines the penalty for exceeding the margin or misclassifying data items. A higher value of C imposes a stricter penalty, resulting in a smaller margin and potentially fewer misclassifications.
- **Hinge Loss:** Hinge loss is a typical loss function in SVMs, punishing incorrect classifications or margin violations. The objective function in SVM frequently combines hinge loss with the regularization term.

- **Dual Problem:** The dual problem of the optimization requires locating Lagrange multipliers related to support vectors. Solving the dual problem enables the use of kernel tricks and more effective computation. [53]

5.3. Applications

SVMs find applications in various real-world scenarios, demonstrating their versatility in solving diverse problems:

- Text and Hypertext Categorization: SVMs play a crucial role in text and hypertext categorization. Their use significantly reduces the dependency on labeled training instances in both standard inductive and transductive settings. Additionally, methods for shallow semantic parsing often rely on support vector machines.
- Image Classification: SVMs are effective in classifying images. Experimental results indicate that SVMs achieve notably higher search accuracy compared to traditional query refinement schemes, particularly after just three to four rounds of relevance feedback. This effectiveness extends to image segmentation systems, including modified versions of SVM that incorporate the privileged approach as suggested by Vapnik.
- Satellite Data Classification: SVMs are employed in the classification of satellite data, including Synthetic Aperture Radar (SAR) data. The supervised SVM approach proves valuable in accurately categorizing diverse types of satellite information.
- Handwritten Character Recognition: SVMs are utilized for recognizing handwritten characters. Their ability to effectively distinguish and classify diverse handwritten symbols contributes to their application in character recognition systems.
- Biological and Scientific Applications: SVM algorithms have found widespread use in biological and other scientific fields. They have been successfully applied to classify proteins, achieving classification accuracies of up to 90%. Permutation tests based on SVM weights offer a mechanism for interpreting SVM models. The interpretation of support vector machine weights has been utilized in the past, and ongoing research focuses on posthoc interpretation to identify features used by the model for making predictions. This area of study holds special significance in the biological sciences. [31]

5.4. Types

There are two main types of Support Vector Machines (SVM)

5.4.1. Linear SVM

Linear SVM is applied to datasets characterized as linearly separable, implying that these datasets can be divided into two distinct classes using a sole straight line. When dealing with linearly separable data, a Linear SVM classifier is employed.

- To illustrate the functionality of the SVM algorithm in this context, consider a dataset containing two features, denoted as x1 and x2, and associated with two tags, green and blue. The objective is to create a classifier that can effectively categorize pairs of coordinates (x1, x2) into either the green or the blue class. This process involves finding the optimal hyperplane, which serves as the decision boundary, by maximizing the margin between the classes.

- Given the two-dimensional nature of the space, a straightforward separation of these two classes is achievable using a single straight line. However, it is important to note that multiple lines have the potential to separate these classes effectively.
- Thus, the SVM algorithm is designed to identify the optimal line or decision boundary, referred to as a hyperplane. The SVM algorithm determines the support vectors, which are the closest points from both classes to the hyperplane. The margin, defined as the distance between the support vectors and the hyperplane, is a crucial factor. The objective of SVM is to maximize this margin, ultimately leading to the identification of the hyperplane with the maximum margin, known as the optimal hyperplane

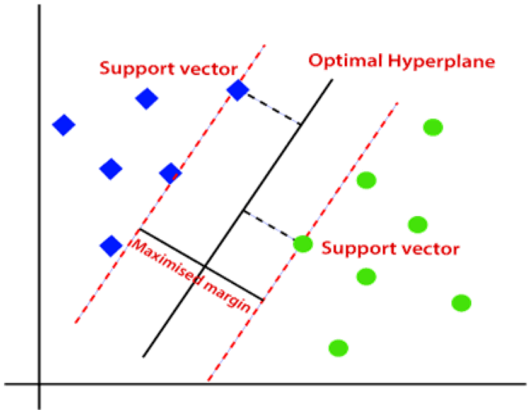

Figure 2. 9:Linear SVM

5.4.2. Non-Linear SVM

Non-Linear SVM, as its name suggests, is employed when dealing with non-linearly separated data. In cases where a dataset cannot be effectively classified using a straight line, denoting it as non-linear data, the appropriate classifier is referred to as the Non-linear SVM classifier.

- In scenarios where data is linearly organized, a straightforward line can be utilized for separation. However, when the data exhibits non-linear patterns, a single straight line becomes insufficient.
- To segregate these data points effectively, introducing an additional dimension is necessary. While linear two dimensions, denoted as x, characterize data and y, non-linear data requires the incorporation of a third dimension, represented as z.
- This dimension is calculated using the equation: $z = x^2 + y^2$. The inclusion of this third dimension transforms the sample space.
- At this point, the SVM algorithm will delineate the datasets into distinct .
- In a 3-dimensional space, it appears akin to a plane parallel to the x-axis. Upon conversion to a 2-dimensional space with z=1, the representation transforms into:

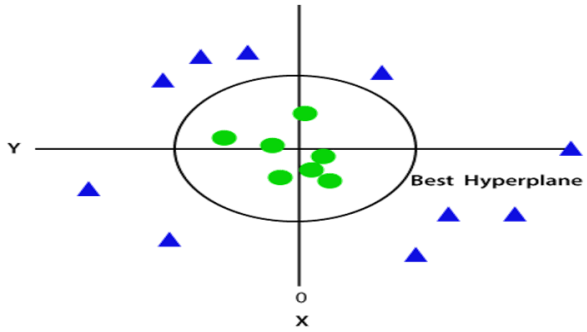

Figure 2. 10:Non-Linear SVM

Therefore, a circular boundary with a radius of 1 is obtained in the case of non-linear data. [54]

5.5. How Does function a Support Vector Machine?

- Optimal Hyperplane Determination :
  - SVM aims to find the hyperplane that best separates data points of different classes.
  - The chosen hyperplane maximizes the margin between classes, enhancing classification accuracy.
- Utilization of Support Vectors
  - Support vectors, which are the closest data points to the hyperplane, play a crucial role in determining the optimal hyperplane.
  - The distance between support vectors and the hyperplane, known as the margin, is a key factor in SVM.
- Kernel Transformation for Non-linearity
  - SVM employs kernel functions to transform data into higher-dimensional spaces.
  - This transformation is particularly useful for handling non-linear separable data points in the original space. [55]

5.6. Examples of Support Vector Machines (SVM)

- Addressing Geo-Sounding Problems
  - SVMs are extensively used to track the layered structure of the Earth, solving inversion problems in geo-sounding.
  - Linear functions and support vector algorithms assist in separating electromagnetic data, contributing to mapping the Earth's structure.
- Assessing Seismic Liquefaction Potential
  - SVMs play a crucial role in evaluating the potential for soil liquefaction during events like earthquakes.
  - Handling tests such as Standard Penetration Test (SPT) and Cone Penetration Test (CPT), SVMs use field data to determine seismic conditions, achieving high accuracy.
- Protein Remote Homology Detection

- In computational biology, SVMs contribute to protein remote homology by categorizing proteins based on amino acid sequences.
- Kernel functions in SVMs highlight commonalities between protein sequences, making them instrumental in computational biology.

- Data Classification
  - SVMs, especially smooth SVMs, are preferred for data classification, employing smoothing techniques to identify patterns by reducing data outliers.
  - Smooth SVMs utilize algorithms like the Newton-Armijo algorithm for handling larger datasets and optimizing problem solving.
- Facial Detection & Expression Classification
  - SVMs are used for classifying facial structures versus non-facial entities, creating decision boundaries based on pixel intensity.
  - Facial expression classification, encompassing emotions like happy, sad, angry, utilizes SVMs for accurate categorization.
- Surface Texture Classification
  - SVMs contribute to classifying surface images, determining the texture of surfaces in images, and categorizing them as smooth or gritty.
- Texte Categorization & Handwriting Recognition
  - SVMs are applied in text categorization, classifying data into predefined categories such as news articles or email types.
  - Handwriting recognition involves training SVM classifiers with sample data and later classifying handwriting based on score values.
- Speech Recognition
  - In speech recognition, SVMs process audio data by extracting features like Mel Frequency Cepstral Coefficients (MFCC) to recognize and separate words from speeches.
- Steganography Detection
  - SVMs are employed for detecting tampering or hidden data in digital images, aiding in security-related matters by analyzing pixel data. [56]

5.7. Advantages

Support Vector Machines (SVMs) present several advantages:

- *Effective in High-Dimensional Spaces:* SVMs perform well in high-dimensional spaces, where the number of features exceeds the number of observations. They efficiently handle such data, making them suitable for applications with numerous features.
- *Handling Nonlinear Data:* SVMs can implicitly manage non-linearly separable data using kernel functions. The kernel trick enables the transformation of the input space into a higher-dimensional feature space, facilitating the identification of linear decision boundaries.
- *Performance with Small Datasets:* SVMs demonstrate effectiveness with small datasets, proving valuable in situations where data availability is limited.

- *Effective in instances where the number of dimensions is larger than the number of specimens:* SVMs are effective in instances where the number of dimensions is larger than the number of specimens, showcasing their versatility.
- *Memory Efficiency:* SVMs exhibit comparably memory systematic behavior, making them memory-efficient compared to other machine learning algorithms. [55] [57] [58]

5.8. Disadvantages

Disadvantages of Support Vector Machines (SVMs) include:

- Computationally Intensive**:** SVMs can be computationally expensive, particularly with large datasets, leading to increased training time and memory requirements as the number of training samples grows.
- Sensitivity to Kernel Choice**:** The performance of an SVM is highly dependent on the choice of kernel, making it challenging to determine the most suitable kernel for a given dataset.
- Sensitivity to Parameter Choice**:** SVMs are sensitive to parameter choices, such as the regularization parameter, making it difficult to find optimal values for a given dataset.
- Not Suitable for Large Datasets: The SVM algorithm may not be suitable for handling large datasets efficiently.
- Performance with Overlapping Classes**:** SVMs may not perform well when datasets contain more noise, specifically when target classes overlap.
- Requirement of Complete Datasets**:** SVMs require complete datasets without missing values, as they cannot handle such cases. [57] [59]

6. Logistic regression

6.1. Definition

Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given data set of independent variables.

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\log(pi) = {}^{1}/_{(1 + \exp(-pi))} \tag{6}$$

$$In({}^{pi}/_{(1 - pi)}) = Beta\_0 + Beta\_1 * X\_1 + \cdots + B\_K * K\_k \tag{7}$$

In this logistic regression equation, log (pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log

likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than 0.5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit. [60]

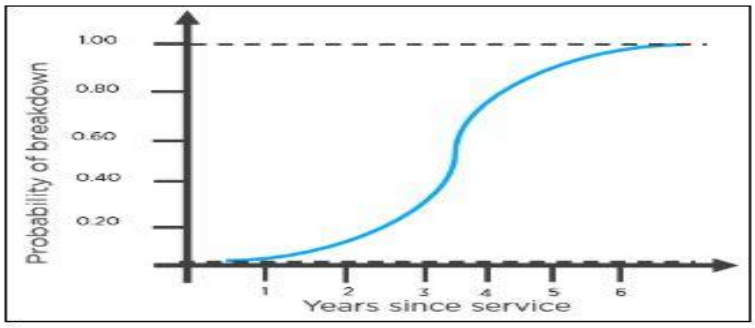

Figure 2. 11: Logistic regression

6.2. Logistic Regression in Predictive Analytics

Logistic Regression is a versatile statistical method used for binary classification and, with some modifications, for multiclass classification. Here are some common use cases for logistic regression

- Medical Diagnosis: Logistic regression is used in healthcare for tasks such as disease prediction and patient risk assessment. For example, it can predict whether a patient has a particular disease (yes/no) based on various medical test results and patient characteristics.
- Credit Scoring: Banks and financial institutions use logistic regression to assess credit risk. It helps in determining whether a loan applicant is likely to default on a loan or has a good creditworthiness based on their financial history, income, and other relevant factors.
- Marketing and Customer Churn: Logistic regression is employed in marketing to predict customer behavior. For instance, it can determine whether a customer is likely to purchase a product or cancel a subscription based on historical data, demographics, and interactions with the company.
- Sentiment Analysis! In natural language processing, logistic regression can be used for sentiment analysis. It classifies text as positive, negative, or neutral based on the sentiment expressed in the text. For example, it can analyze social media comments to gauge public sentiment about a product or brand.
- Spam Detection: Logistic regression is a common tool in email spam filters. It classifies incoming emails as spam or not spam based on features like keywords, sender information, and email content.

- Quality Control: In manufacturing and quality control, logistic regression can be used to predict whether a product is likely to be defective or meet certain quality standards based on various measurements and parameters.
- Employee Attrition: Human resources departments use logistic regression to predict employee attrition or turnover. It helps in identifying factors that contribute to employees leaving a company, such as job satisfaction, salary, and work environment.
- Default Prediction: Logistic regression is used in the finance industry to predict loan defaults. It assesses whether a borrower is likely to default on a loan based on their financial history, credit score, and other relevant factors.
- Real Estate: In real estate, logistic regression can be used to predict whether a house will be sold within a certain period based on features like location, price, and property characteristics.
- Social Sciences: Logistic regression is widely used in social sciences for various research purposes, including predicting voting behavior, analyzing survey data, and studying the impact of variables on a specific outcome. [61]

6.3. Applications of logistic regression

Logistic regression has several real-world applications in many different industries.

- Manufacturing: Manufacturing companies use logistic regression analysis to estimate the probability of part failure in machinery. They then plan maintenance schedules based on this estimate to minimize future failures.
- Healthcare: Medical researchers plan preventive care and treatment by predicting the likelihood of disease in patients. They use logistic regression models to compare the impact of family history or genes on diseases.
- Finance**:** Financial companies have to analyze financial transactions for fraud and assess loan applications and insurance applications for risk. These problems are suitable for a logistic regression model because they have discrete outcomes, like high risk or low risk and fraudulent or not fraudulent.
- Marketing: Online advertising tools use the logistic regression model to predict if users will click on an advertisement. As a result, marketers can analyze user responses to different words and images and create high-performing advertisements with which customers will engage. [62]

6.4. Types of Logistic Regression

On the basis of the categories, Logistic Regression can be classified into three types:

- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”. [63]

6.5. Algorithm

6.5.1. Logistic Function (Sigmoid Function)

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold, value tends to 1, and a value below the threshold values tends to 0.

6.5.2. Assumptions for Logistic Regression

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

6.5.3. Logistic Regression Equation

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \quad (8)$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}; 0 \; for \; y = 0, and \; infinity \; for \; y = 1 \quad (9)$$

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n + \quad (10)$$

The above equation is the final equation for Logistic Regression. [52]

6.6. Advantages of the Logistic Regression Algorithm

- Logistic regression performs better when the data is linearly separable
- It does not require too many computational resources as it's highly interpretable
- There is no problem scaling the input features—It does not require tuning
- It is easy to implement and train a model using logistic regression
- It gives a measure of how relevant a predictor (coefficient size) is, and its direction of association (positive or negative) [64]

6.7. Disadvantages of the Logistic Regression Algorithm

- Overfitting on high dimensional data

- Nonlinear problems can't be solved with logistic regression since it has a linear decision surface
- Assumes linearity between dependent and independent variables.
- Fails to capture complex relationships.
- Only important and relevant features should be used otherwise model's predictive value will degrade. [65]

7. Conclusion

In conclusion, the second chapter of this study represents a continuation of our journey into the world of analytical and predictive modeling. In this chapter, we focused on analyzing several different models and delving into their details comprehensively. We provided a thorough definition of each model, along with reviewing the various types that each model may follow. Emphasis was placed on explaining the different dimensions of each model and their role in the analysis and prediction process. By exploring the algorithms and techniques employed in each model, this chapter aimed to provide the reader with a deep understanding of the different characteristics and applications of these models, and how to select the most appropriate model for specific research objectives.

In the next chapter, we will continue our journey in the world of modeling by applying a specific model to some products.

CHAPITER 3

PROPOSED MODELS FOR FOOD SAFETY

1. Introduction

In recent years, there has been an increasing trend towards using machine-learning techniques in analyzing and predicting food safety risks, as previous studies have demonstrated the ability of these techniques to accurately predict and classify food safety risks. Additionally, efforts have been directed towards building analytical models to forecast the prices of certain products and classify them based on food safety risks, with the aim of mitigating market disruptions and providing a framework for effective decision-making in pricing. In this chapter, we will review previous research related to food safety risks and apply analytical and predictive models to several products, with the goal of illustrating how modern techniques can be used to estimate and classify prices according to food risk levels

2. State of the art

Food security and crisis management pose significant threats to human well-being and social stability. Traditional approaches to addressing these challenges often rely on manual data analysis and decision-making processes, which can be time-consuming and error-prone. In recent years, Machine Learning (ML) models have demonstrated their ability to extract valuable insights from large and complex datasets. By leveraging ML algorithms, food-related data can be analyzed more efficiently, leading to improved decision-making and resource allocation.

ML techniques can significantly enhance the performance of crisis prediction and early warning systems by analyzing large, diverse datasets and identifying complex patterns and relationships that may be difficult for humans to discern. For example, ML models can be trained on historical climate data, market fluctuations, conflict patterns, and other relevant variables to predict the likelihood and severity of impending crises. These models can then be integrated into early warning systems to provide decision-makers with timely and actionable insights.

Several studies have applied ML models to food insecurity and crisis management. For instance, [66] attempted to find the best way to predict food prices for the average Canadian consumer, while [67] proposed Food Security Prediction framework and applied machine and deep learning models on heterogeneous data.

In the context of food insecurity, [68] proposed ML approaches to infer predictors of food insecurity in southern Malawi, while [69] explored how the COVID-19 pandemic influenced urban food insecurity patterns, and [70] highlighted the role of AI and big data analytics for forecasting disruptions in global food value chains to tackle food insecurity. Using remote sensing and ML-Based prediction, [71] tackled food insecurity of wheat grain crops and used three regression techniques including Random Forest, Xtreme Gradient Boosting (XGB)

regression, and Least Absolute Shrinkage&Selection Operator (LASSO) regression.

Handling more specific products, authors in [72] proposed ML methods to predict the Turkish mercantile exchange wheat index. They showed that tree-based methods revealed better overall performance, while authors in [73]explored neural network model for forecasting problems in datasets of daily prices over periods of greater than 50 years for coffee, corn, cotton, oats, soybeans, soybean oil, sugar, and wheat.

Analytics of social media data investigated ML efficacy in food security. For instance, [74] introduced "HungerGist", a multi-task deep learning model utilizing news texts and NLP techniques. Authors exploited for the study, a corpus of over 53,000 news articles from nine African countries. They demonstrated that their proposed model can predict three critical objectives (food insecurity, food price, and social insecurity). With our proposed models, we aim to predict food price and safety risk of several products using datasets from various sources.

3. Dataset Description

3.1. Collect of the Dataset

The data was collected from multiple sources, including the Al-Kadhi market for price data and the Mateo website for weather data. This diversity in sources enables a deeper understanding of the factors influencing the food market and product quality, as well as how weather conditions impact these processes. By utilizing this diverse data, advanced analytical models can be developed to monitor and assess food safety risks, and build effective strategies to address these risks and ensure food safety for consumers.

3.2. **List of Products:** We have selected a set of 34 products presented in the table 3.1.

Green Legumes		**Dry Legumes**	**Fruits**
Artichoke	Green Olive	Dry Beans	Apple
Black olive	Lettuce	Corn	Bananas
Beans	Onion	Chickpeas	Lemon
Beetroot	Onion Bundle	Fouila	Orange
Cabbage	Pea	Lentils	Strawberry
Carrot	Pepper		
Cauliflower	Potato		
Cucumber	Pumpkin		
Eggplant	Prickly Artichoke		
Fennel	Tomato		
Garlic	Turnip		
Green Beans	Zucchini		

Table 3. 1: List of Products of the study

3.3. Detailed description of the dataset

The Dataset spans a few months from January to April, providing us with daily prices of all the products listed in Table 3.1. We collected real data making for us a great opportunity to create an advanced model for predicting food prices and monitoring food safety risks. This model is

comprehensive and accurate, analyzing a wide range of variables to understand the conditions that may affect food quality and safety.

- Days: The number of days reflects the time period for data collection, allowing for the analysis of changing conditions throughout the year, contributing to an understanding of seasonal variations in food safety.
- Temperature، Humidity, Wind, and Precipitation**:** These variables provide insights into the surrounding weather conditions, which influence bacterial growth and food spoilage, and are thus used to estimate food safety risks and take appropriate preventive measures.
- Prices: Changes in prices reflect economic challenges that may impact food quality and safety, allowing for an estimation of the effects of economic conditions on food safety risks.
- Food safety risk: To determine whether the price fluctuations of a product indicate a potential crisis or not, we relied on the collective knowledge and extensive experience of numerous experts specializing in the fields of economics and trading.

3.4. A Sample of collected data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Days							**Price (DA)**							**Weather**		
	Garlic	onionBundle	onion	pepper	Tomato	Potato	Cucumber	Zucchini	Eggplant	green Bean	Cauliflower	Carrot	Temperature (C°)	Humidity (%)	Wind (km/h)	Precipitation (mm)
1	300	30	50	125	55	50	45	140	65	60	50	40	12	41	4	0
2	330	35	55	120	50	40	60	120	70	50	50	45	15	48	7	0
3	330	35	55	120	50	40	60	120	70	50	50	45	14	44	2	0
4	350	25	45	140	55	55	50	170	80	50	50	35	14	44	4	0
5	350	25	45	140	55	55	50	170	80	50	50	35	9	71	4	0
6	350	25	45	160	55	55	50	170	80	60	50	35	12	47	46	6,1
7	350	30	50	160	60	50	45	180	90	60	45	55	11	47	37	0
8	350	30	50	155	60	50	45	190	80	60	40	45	11	35	41	0
9	350	25	50	150	50	40	45	185	85	60	35	40	8	66	6	0
10	350	30	55	150	60	40	45	190	90	60	30	40	8	76	7	1
11	350	30	55	150	60	40	45	190	90	60	30	40	12	72	4	2
12	250	30	50	150	55	55	45	190	100	50	35	50	12	55	22	0
13	250	30	50	150	55	55	45	190	100	50	35	50	11	58	6	0
14	250	30	50	150	55	55	45	190	100	50	35	50	11	62	7	0
15	300	30	50	155	55	55	45	190	100	50	35	50	18	21	20	0

Figure 3. 1: A sample of collected data

4. Development Environment

- PyCharm (Version 2023.1.2): PyCharm serves as a dedicated integrated development environment (IDE) crafted for Python programming. It provides essential features like code analysis, a graphical debugger, integrated unit testing, seamless integration with version control systems, and support for web development.
- Scikit-learn (Version 1.2.1): Scikit-learn, or sklearn, emerges as a comprehensive machine-learning library tailored for Python. It encompasses a plethora of algorithms and utilities addressing diverse machine learning tasks, spanning from classification and regression to

clustering and dimensionality reduction.

- NumPy (Version 1.24.1): NumPy emerges as a fundamental Python library facilitating numerical computing. It enables efficient and high-performance operations on multidimensional arrays, crucial for endeavors such as scientific computing, data analysis, and machine learning.
- Pandas (Version 1.5.3): Pandas enjoys widespread adoption as an open-source library specializing in data manipulation and analysis within Python. Its core lies in offering a user-friendly and efficient approach to handling structured data through its DataFrame data structure. With Pandas, tasks like data cleaning, filtering, transformation and aggregation are streamlined. Moreover, it provides robust tools for data visualization and seamless integration with other libraries. Pandas stands as a favored choice among data scientists and analysts for managing tabular data in Python.

These tools, along with matplotlib. pyplot, which facilitates data visualization, collectively empower the implementation and evaluation of our approach. [75]

5. Proposed Regression models

Regression models are a class of statistical models used in data analysis and machine learning to explore the relationship between one or more independent variables (often called predictors or features) and a dependent variable (often called the target or outcome). The primary goal of regression analysis is to understand how changes in the independent variables are associated with changes in the dependent variable. This understanding can help in making predictions and explaining the observed data. There are various types of regression models, each suited to different types of data and research questions. Some common types of regression models include: [75]

We conducted an analytical study of vegetable prices, applying two different regression models to predict these prices. These models included linear regression and polynomial regression.

5.1. Simple Linear regression

As an illustrative example, a simple model was created to analyze and predict the price of **bananas** using linear regression, where the correlation coefficient (R-squared) was used to evaluate the model's ability to explain the variance in the data. The data was split into two parts, with half of the data used to train the model and the other half-used to test its performance based on the following code excerpt:

```
# Splitting the data
X_train, X_test, Y_train, Y_test = train_test_split( *arrays: X, Y, test_size=0.5, random_state=0)

# Training the model
regressor = LinearRegression()
regressor.fit(X_train, Y_train)

# Predicting values
Y_pred = regressor.predict(X_test)

# Calculating the coefficient of determination (R-squared)
r_squared = r2_score(Y_test, Y_pred)
print(f'R square: {abs(r_squared)}')

# Displaying results in a plot
plt.scatter(X_test[:, 0], Y_test, color='red', label='Actual')
plt.plot( *args: X_test[:, 0], Y_pred, color='blue', label='Predicted')
plt.xlabel('Days')
plt.ylabel('Bananas Price')
plt.title('Actual vs Predicted Bananas Price')
plt.legend()
plt.grid(True)
plt.show()
```

Exemple obtained result :

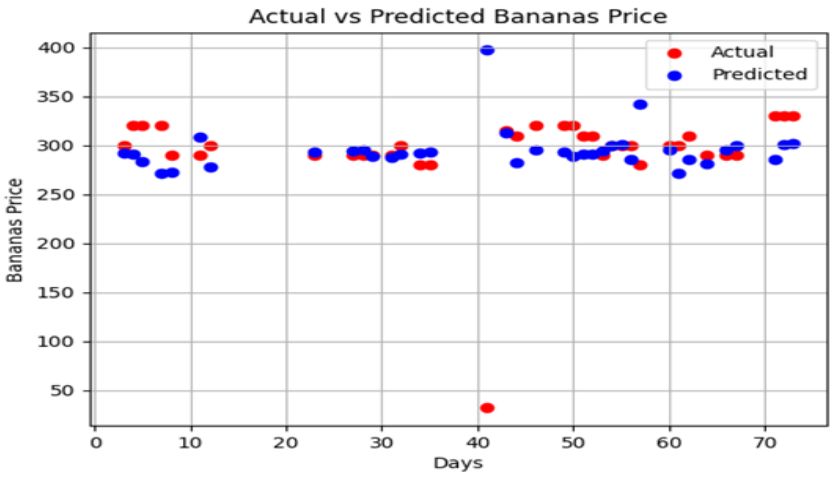

Figure 3. 2: Simple Linear regression for Bananas

5.2. Polynomial Regression

An illustrative example using polynomial regression model was created to analyze the price of **garlic**, where the correlation coefficient (R-squared) was used to assess the model's ability to explain the variance in the data. The model was trained using all available data, and the predicted results are presented accordingly.

```
# Try increasing the degree of polynomial
degrees = [4, 6, 8, 10]  # Try different degrees
best_r_squared = -1
best_degree = None

for degree in degrees:
    poly_reg = PolynomialFeatures(degree=degree)
    X_poly = poly_reg.fit_transform(X)
    pol_reg = make_pipeline( *steps: StandardScaler(), LinearRegression())
    pol_reg.fit(X_poly, y)


    # Calculating R-squared
    y_pred = pol_reg.predict(X_poly)
    r_squared = r2_score(y, y_pred)

    if r_squared > best_r_squared:
        best_r_squared = r_squared
        best_degree = degree

# Fit the best model with the chosen degree
poly_reg = PolynomialFeatures(degree=best_degree)
X_poly = poly_reg.fit_transform(X)
pol_reg = make_pipeline( *steps: StandardScaler(), LinearRegression())
pol_reg.fit(X_poly, y)

# Visualizing the Polynomial Regression results
plt.scatter(X, y, color='red')
plt.plot( *args: X, pol_reg.predict(X_poly), color='blue')
plt.title('Garlic Price Prediction (Polynomial Regression)')
plt.xlabel('Days')
plt.ylabel('Garlic Price')
plt.show()

# Print the best R-squared value and the corresponding degree
print(f'Best R-squared: {best_r_squared}')
print(f'Best degree: {best_degree}')
```

Figure 1.3:

Example obtained result for Garlic product:

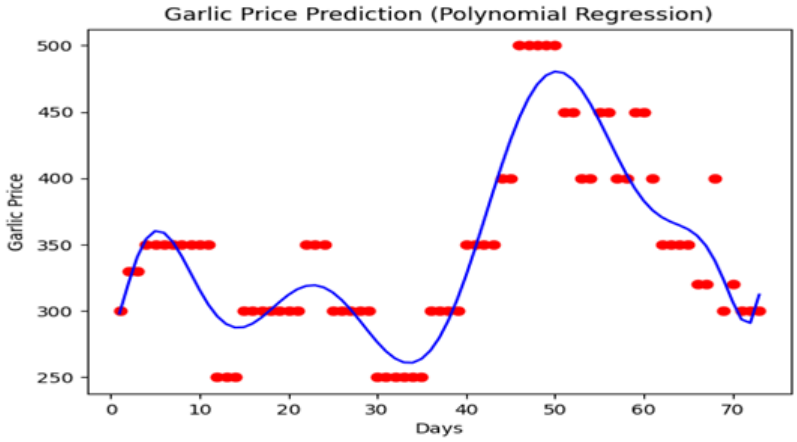

Figure 3. 3: Polynomial Regression for Garlic product

6. Proposed classification models

After conducting the regression process using the different models we have studied, we tackled the classification modeling in order to predict for each product whether its price variation may conduct to a food safety risk or not. We identified several important variables for predicting vegetable prices, such as the time (expressed by days of the year), temperature, humidity, wind speed, rainfall, and prices. The class label indicating the potential safety risks and expressed as a binary "yes" or "no" is determined, thanks to expertise of food market traders and economists depending on the aforementioned factors. Utilizing these variables, we trained four models: Random Forests, Decision Trees, Support Vector Machine (SVM), and Logistic regression.

6.1. Random forest

A random forest model was used where the data was divided into two sets: The training set on which the model was trained with 75%, and the test set with 25%. The confusion matrix and measures of precision, accuracy, positive precision, and recall were used to evaluate the model's performance and effectiveness in classification. Based on the following code excerpt:

```
# Splitting the dataset into the Training set and Test set
X_Train, X_Test, Y_Train, Y_Test = train_test_split( *arrays: X, Y, test_size=0.25, random_state=0)

# Feature Scaling
sc_X = StandardScaler()
X_Train = sc_X.fit_transform(X_Train)
X_Test = sc_X.transform(X_Test)

# Fitting the classifier into the Training set
classifier = RandomForestClassifier(n_estimators=200, criterion='entropy', random_state=0)
classifier.fit(X_Train, Y_Train)

# Predicting the test set results
Y_Pred = classifier.predict(X_Test)

# Calculate confusion matrix
cm = confusion_matrix(Y_Test, Y_Pred)


# Calculate accuracy, precision, and recall
accuracy = accuracy_score(Y_Test, Y_Pred)
precision = precision_score(Y_Test, Y_Pred, average='weighted', zero_division=0)
recall = recall_score(Y_Test, Y_Pred, average='weighted', zero_division=0)

print("Confusion Matrix:")
print(cm)
print("Accuracy: {:.2f}%".format(accuracy * 100))
print("Precision: {:.2f}%".format(precision * 100))
print("Recall: {:.2f}%".format(recall * 100))
```

Results obtained from Dry Beans Dataset :

```
Confusion Matrix:
[[5 2 2]
 [0 2 0]
 [0 0 8]]
Accuracy: 78.95%
Precision: 86.32%
Recall: 78.95%
```

6.2. Decision Trees

Decision trees are a machine learning technique used to solve classification and prediction problems. As an illustrative example, we used decision tree analysis to investigate vegetable price data, specifically to predict safety risk of the **pepper**. The process included the following steps:

We started by reading data from a **CSV** file, leveraging the **Pandas** library.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, plot_tree
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder

# Read data from CSV file
df = pd.read_csv( filepath_or_buffer: 'pepper1.csv', delimiter=';')

```

Next, we convert the columns to their appropriate data types, ensuring data validity by applying the **pd.to_numeric** function.

```
## Convert columns to the correct data types
df[['Days', 'Temperature', 'Humidity', 'Wind', 'Precipitation', 'pepper']] =\
    df[['Days', 'Temperature', 'Humidity', 'Wind', 'Precipitation', 'pepper']].apply(pd.to_numeric, errors='coerce')
```

The text values in the “Risk” column was then converted to classifiable numeric values using the **LabelEncoder** library.

```
# Convert text values in the "Label" column to numbers
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(df['Risk'])

```

Next, we split the data into two subsets: a training set and a test set, adhering to an 80:20 split accomplished through the **train_test_split** function.

```
# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)
```

We built the model using the decision tree algorithm **DecisionTreeClassifier**.

```
# Build the model using the decision tree
model = DecisionTreeClassifier()
model.fit(X_train, y_train)
```

We then determined the number of **nodes** in the decision tree and evaluated its **accuracy** on the test data.

```
# Number of nodes in the decision tree
node_count = model.tree_.node_count
print("Number of nodes in the decision tree:", node_count)

# Determine accuracy on test data
accuracy = model.score(X_test, y_test)
print("Accuracy:", accuracy)
```

Finally, we plotted a decision tree to visualize the decisions the model makes based on different input variables, using the **plot_tree** function.

```
# Plot the decision tree
plt.figure(figsize=(20,8))
plot_tree(model, feature_names=X.columns, class_names=label_encoder.classes_, filled=True, rounded=True)
plt.show()
```

The obtained decision tree of pepper is presented in Fig. 3.4.

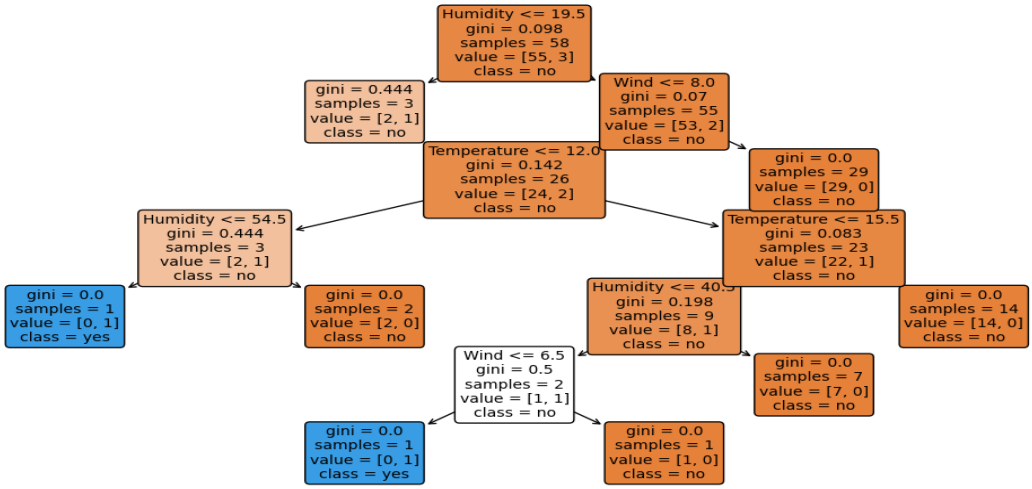

Figure 3. 4: Decision tree classification for Pepper product

The tree starts from the root node at the top, then branches out into child nodes based on the values of the variables. At each node, a decision is made based on the value of a particular variable. If the conditions are met, the tree moves to the next node; otherwise, it moves to a different node. Ultimately, the tree reaches a leaf node (end node) where the final decision is made.

For example, if the wind speed is less than 3.45 m/s, the humidity is less than 44.5%, and the temperature is 0.49 degrees Celsius, the tree will reach the final decision of "yes" (the phenomenon occurs). Other paths can be read in the same way to make different decisions.

```
"C:\Users\MXP TAIBECHE\project\pythonProject\.venv\Scripts\python.exe" "C:\Users\MXP TAIBECHE\project\pythonProject\classification tree.py"
Number of nodes in the decision tree: 15
Accuracy: 0.8666666666666667
```

Overall, this provides an overview of the number of nodes in the decision tree generated and the resulting classification accuracy when applied to the dataset.

6.3. Support Vector Machine (SVM)

We applied a Support Vector Machine (SVM) algorithm to predict food safety risk based on the aforementioned variables. In this context, we divided the data into a training set and a test set. We then trained the model and used it to predict safety. Next, we calculated the accuracy to evaluate the model's performance. In addition, we printed a classification report to get a deeper understanding of the model's performance. Finally, we plot actual and forecast safety risk of the selected product. The following code illustrates all these steps for the orange product as an example.

```
y = df['Orange']

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)

# Train the model
svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)

# Make predictions
y_pred = svm_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Print classification report
print("Classification Report:")
print(classification_report(y_test, y_pred, zero_division=0))

# Plot data and predictions
plt.scatter(X_test, y_test, color='blue', label='Actual')
plt.scatter(X_test, y_pred, color='red', label='Predicted')
plt.xlabel('Days')
plt.ylabel('Orange Price')
plt.title('Predicted Orange Prices')
```

The obtained result for Orang product:

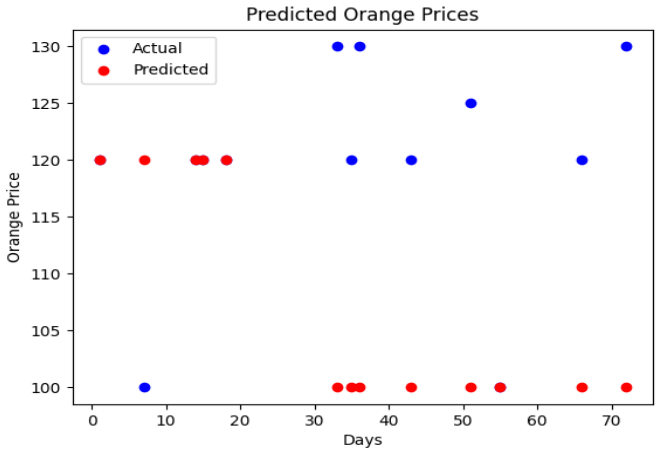

Figure 3. 5: Support Vector Machine (SVM)

```
"C:\Users\MXP TAIBECHE\project\pythonProject\.venv\Scripts\python.exe"
Accuracy: 0.38461538461538464
Classification Report:
              precision    recall  f1-score   support

       100.0       0.12      0.50      0.20         2
       120.0       0.80      0.57      0.67         7
       125.0       0.00      0.00      0.00         1
       130.0       0.00      0.00      0.00         3

    accuracy                           0.38        13
   macro avg       0.23      0.27      0.22        13
weighted avg       0.45      0.38      0.39        13
```

6.4. Logistic regression

Logistic Regression model is used to predict food safety risk. The data is split into training and testing sets with an 80:20 ratios, and feature scaling (standardization) is applied to the independent variables. After training the model, its accuracy is calculated. The following code is predicting safety risk for the lettuce product as an example.

```
# Prepare data
X = df[['Days']]
y = df['Lettuce']

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)

# Scale the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Build the model using Logistic Regression with increased max_iter
logistic_model = LogisticRegression(max_iter=1000)
logistic_model.fit(X_train_scaled, y_train)

# Make predictions on scaled data
y_pred = logistic_model.predict(X_test_scaled)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Plot data and predictions
plt.figure(figsize=(10, 6))
plt.scatter(X_test, y_test, color='blue', label='Actual')
plt.scatter(X_test, y_pred, color='red', label='Predicted')
plt.title('Actual vs Predicted Lettuce Prices')
```

The obtained result is as follows:

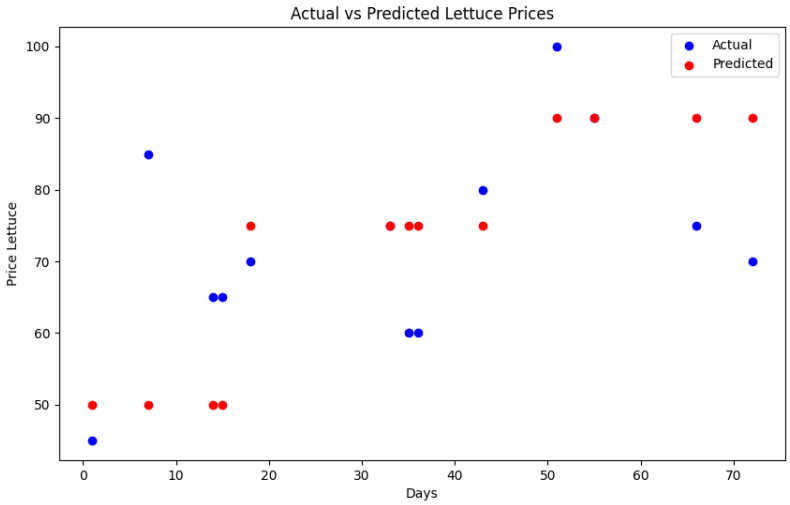

Figure 3. 6: Logistic regression

```
"C:\Users\MXP TAIBECHE\project\pythonProject\.venv\Scripts\python.exe"
Accuracy: 0.15384615384615385
```

7. Conclusion

This chapter reviewed the process of developing ML models to predict the prices of a set of vegetables and forecasting their safety risks using market data and weather information. Regression and classification models such as Linear and polynomial regressions, random forests, decision trees, support vector machines, and logistic regression were applied. The performance of these models was evaluated through measures such as the correlation coefficient, confusion matrix, accuracy, positive predictive value, as well as plotting the data and predictions to visually assess the models.

In the next chapter, detailed results of applying these models to all the vegetable products in the dataset will be presented, providing an opportunity to evaluate the performance of the models in more detail and allowing for a comparison of results across different products.

CHAPITER 4

RESULTS AND DISCUSSION

1. Introduction

To evaluate the performance of the proposed models for predicting and classifying food safety risks, we conducted comprehensive and extensive experiments, and valuable tangible results were collected. In this section, we will present the results of the corresponding experiments to illustrate the effectiveness of each model.

2. Obtained Results and Discussion

2.1. Regression models

After applying regression models, which are used as a means of prediction, and calculating the correlation coefficient for 34 products, the correlation coefficient results were obtained for each model. These results illustrate the relationship between the independent variables and the dependent variable for each product, helping to provide accurate forecasts and make informed decisions in the future. Table (4.1) shows the summary of all studied products (34) product.

Product	Linear regression	Polynomial Regression	Product	Linear regression	Polynomial Regression
Apple	0.47	**0.87**	Green olive	0.24	**0.81**
Artichoke	0.36	**0.89**	Prickly Artichoke	0.08	**0.84**
Bananas	**0.93**	0.18	Lemon	0.12	**0.76**
Green Beans	0.16	**0.79**	Lentils	0.03	**0.47**
Dry Beans	0.71	**0.78**	Lettuce	0.04	**0.49**
Beetroot	0.44	**0.77**	Onion	0.09	**0.75**
Black olive	0.11	**0.69**	Onion Bundle	0.11	**0.74**
Cabbage	0.12	**0.86**	Orange	0.22	**0.70**
Carrot	0.03	**0.50**	Pea	0.14	**0.47**
Cauliflower	0.42	**0.85**	Pepper	0.28	**0.31**
chickpeas	0.17	**0.72**	Potato	0.03	**0.57**
Corn	0.42	**0.83**	Pumpkin	0.27	**0.85**
Zucchini	0.43	**0.91**	Strawberry	0.21	**0.50**
beans	0.04	**0.88**	Tomato	0.09	**0.56**
Cucumber	0.52	**0.84**	Turnip	0.37	**0.74**
Eggplant	0.08	**0.46**	Fouila	0.10	**0.80**

Table 4.1: R-Squared of Regression Models

In Table 4.1, we present the correlation coefficients (R_squard) calculated for all the products of the dataset. It can be seen that most products are better predicted using the Polynomial Regression model. This interpretation is logical because Linear Regression models may not adequately capture the complexity of the relationship between variables of the dataset, while Polynomial Regression can handle more complex patterns and capture non-linear relationships between variables by fitting a non-linear regression line. Furthermore, the prediction ratio (R^2) for certain products, such as **Onion Bundle, Carrot, Turnip, Pumkin and Lemons**, was suboptimal when using regression models, likely due to the significant price volatility exhibited by these commodities during the study period.

2.2. Classification models

When applying classification models to 34 products and analyzing the classification accuracy for each model, results were obtained that reflect the effectiveness of each model in accurately classifying the products. These results highlight the predictive ability of each model in correctly classifying the products, contributing to providing accurate recommendations and making informed decisions in the future. Table 4.2 shows the accuracy of the used classification models for the 34 studied products.

Products	Random Forest	Decision Tree	SVM	Logistic regression
Apple	57.89	60	**76.92**	69.23
Artichoke	36.84	**60**	38.46	46.15
Bananas	42.11	**86.66**	23.07	30.76
Green Beans	57.89	**60**	38.46	30.76
Beans	**78.95**	53.33	61.53	61.53
Beetroot	21.05	**86.66**	69.23	46.15
Black olive	**78.95**	73.33	53.84	46.15
Cabbage	52.63	**53.33**	30.76	23.07
Carrot	26.32	**86.66**	30.76	7.69
Cauliflower	31.58	**66.66**	46.15	30.76
chickpeas	84.21	**86.66**	30.76	23.07
Corn	36.84	**93.33**	7.69	20
Zucchini	21.05	**73.33**	38.46	23.07
Crispy beans	84.21	66.66	**92.30**	69.23
Cucumber	21.05	**73.33**	7.96	15.38
Eggplant	47.37	**86.67**	38.46	46.15
Fennel	57.89	**66.67**	61.53	61.53
Garlic	52.63	**80**	15.38	23.07
Green olive	63.16	**66.67**	38.46	53.84
Prickly Artichoke	47.37	66.67	**69.23**	**69.23**
Lemon	47.37	**86.67**	69.23	69.23
Lentils	78.95	**93.33**	48.16	76.92

Lettuce	26.32	**66.67**	15.38	15.38
Onion	36.84	**86.67**	25	16
Onion bundle	57.89	**99**	46.15	46.15
Orange	26.32	**93.34**	38.33	38.46
Pea	57.89	73.33	**76.92**	23.07
Pepper	42.11	**86.67**	23.07	15.38
Potato	47.37	**93.33**	15.38	15.38
Pumpkin	36.84	**80**	53.48	46.15
Strawberry	63.16	**93.33**	46.15	61.53
Tomato	15.79	**86.66**	7.96	15.38
Turnip	31.58	**86.66**	53.84	46.15
Fouila	84.21	80	**92.30**	76.92

Table 4. 2: Accuracy of Classification Models

Table 4.2 shows a summary of accuracy results using classification models. It can be seen that Decision Tree model exhibited strong predictive performance for the majority of products (**27 products**), which aligns with the well-established effectiveness of this model across various domains, particularly in anticipating potential crises. Nonetheless, SVM model demonstrated superior accuracy for certain commodities, while only a handful of products achieved optimal prediction through Random Forest and surprisingly, Logistic Regression model failed to accurately predict the safey risk for any product.

3. Conclusion

After conducting comprehensive experiments on various models for predicting food prices and forecasting food safety risks, the results showed that **Polynomial Regression** and **Decision Tree** models delivered the best performance. These models demonstrated high accuracy in predicting and correctly classifying the products, confirming their effectiveness and efficiency in food safety applications. These results contribute to providing accurate recommendations and making informed decisions to enhance food safety in the future.

GENERAL CONCLUSION

Machine learning models have the potential to revolutionize food insecurity and crisis management by providing predictive analytics, optimizing resource allocation, and enhancing decision-making.

We would like to emphasize the significant importance of food security and its direct impact on the health and well-being of individuals and communities. This study addressed the analysis and prediction of food safety risks through the development of predictive models based on data and its analysis. Various models were reviewed and applied to real market data, focusing on predicting product prices and assessing food safety risks.

This study highlighted two primary objectives: 1) identifying influential predictors to enable price prediction and crisis forecasting using regression models, and 2) leveraging well-established classification models for food crisis early warning systems.

By applying several ML models on a dataset spanning 34 products (legumes and fruits) over an entire year, the results demonstrated promising potential for predicting food insecurity risks for the majority of the commodities. The results showed that **Polynomial Regression** and **Decision Tree** models were the most efficient in accurately predicting and classifying the products, confirming their effectiveness in food safety applications. These findings contribute to providing accurate recommendations and making informed decisions to enhance food safety in the future.

We conclude by highlighting that achieving sustainable food security requires international cooperation, technological advancements in agriculture, and promoting sustainability to face increasing challenges. By utilizing predictive models and analytical tools, we can improve our ability to address future food crises and ensure the availability of safe food for everyone.

We hope that this study has contributed to enhancing our understanding of the challenges and potential solutions in the field of food security and serves as an impetus for further research and development in this vital area.

However, several challenges and opportunities remain to be addressed, including improving data quality and increasing the availability of relevant data sources. However, in the future, we plan to explore deep learning techniques like CNN, LSTM, etc., to analyze datasets for food crisis forecasting. In addition, we plan to integrate more predictors like supply chains for enhancing the accuracy of insecurity risks.

REFERENCES

[1] S. d. Pee, Encyclopedia of Human Nutrition (Third Edition), 2013.

[2] f. S. Yasmine, biotechnologie et sécurité alimentaire en Algérie, université Constantine 1, 2013.

[3] “World Demand for Food,” [Online]. Available: Savemyexams.com. [Accessed 10/02/2024].

[4] “) الامن الغذائي : مفهومه أهميته انواعه معوقاته”.

[5] “ hunger in our region,” [Online]. Available: https://www.capitalareafoodbank.org. [Accessed 10/02/2024].

[6] “concern, OCR Food security a global,” [Online]. Available: https://www.bbc.co.uk. [Accessed 11/02/2024].

[7] "HLPE_Price volatility and food security," [Online]. Available: https://www.fao.org. [Accessed 11/02/2024].

[8] “IMF,” 30/09/2022. [Online]. Available: https://www.imf.org. [Accessed 11/02/2024].

[9] “ Department of Economic and Social AffairsEconomicv Analysis,” [Online]. Available: https://www.un.org. [Accessed 11/02/2024].

[10] “worldbank,” [Online]. Available: https://www.worldbank.org. [Accessed 14/02/2024].

[11] “Famine in Somalia, 2011–2012,” [Online]. Available: https://academic.oup.com. [Accessed 15/02/2024].

[12] “the Pakistan Flood 2010,” [Online]. Available: https://earthobservatory.nasa.gov. [Accessed 14/02/2024].

[13] “ Pakistan Floods of 2010,” 2010. [Online]. Available: https://www.britannica.com.

[14] “ Locust Plague in EastAfrica, Locust Plague in East Africa,” 06/2020. [Online]. Available: https://news.harvard.edu. [Accessed 15/02/2024].

[15] “Syrian_civil_war,” [Online]. Available: https://en.wikipedia.org. [Accessed 17/02/2024].

[16] “ COVID-19 Brief: Impact on Food Security,” [Online]. Available: https://www.usglc.org. [Accessed 20/02/2024].

[17] “questions for research and preventive action Famine in Gaza,” [Online]. Available: https://www.nature.com. [Accessed 17/02/2024].

[18] “, Environmental Impacts of Food Production,” [Online]. Available: https://ourworldindata.org. [Accessed 20/02/2024].

[19] “https://www.ncbi.nlm.nih.gov/,” [Online]. [Accessed 22/02/2024].

[20] “Pollution IndustrialAgricultural, Pollution Industrial Agricultural,,” [Online]. Available: https://www.nrdc.org. [Accessed 22/02/2024].

[21] “Food Systems and the Environment,” [Online]. Available: https://www.genevaenvironmentnetwork.org. [Accessed 22/02/2024].

[22] “Feeding_America,” [Online]. Available: https://en.wikipedia.org. [Accessed 22/02/2024].

[23] “food-insecurity Healthcare,” [Online]. Available: https://www.healthcarevaluehub.org. [Accessed 22/02/2024].

[24] “natfoodatfood,” [Online]. Available: https://www.nature.com. [Accessed 22/02/2024].

[25] “food-chemistry,” [Online]. Available: https://www.sciencedirect.com. [Accessed 22/02/2024].

[26] “Food Security and Political Stability in the Asia-Pacific Region,” Asia-Pacific, 11/09/1998.

[27] “geriatric-medicine,” [Online]. Available: https://academic.oup.com. [Accessed 15/02/2022].

[28] “climate-change-global-food-security,” [Online]. Available: https://www.preventionweb.net.

[Accessed 22/02/2024].

[29] “international-organizations-for-worldwide-food-security,” [Online]. Available: https://www.biolabmag.com. [Accessed 22/02/2024].

[30] “30 Organizations Working to End Hunger,” [Online]. Available: https://www.humanrightscareers.com. [Accessed 22/02/2024].

[31] “vwikipedia,” [Online]. Available: www.wikipedia.com. [Accessed 01/03/2024].

[32] “ What Is Linear Regression?,” [Online]. Available: https://www.spiceworks.com. [Accessed 02/03/2024].

[33] A. Chakure, “Types of Linear Regression,” Jun 29, 2019. [Online]. Available: https://medium.datadriveninvestor.com.

[34] “Polynomial Regression,” [Online]. Available: https://www.studysmarter.co.uk. [Accessed 10/03/2024].

[35] “Linear Regression,” [Online]. Available: https://www.vedantu.com/. [Accessed 10/03/2024].

[36] T. Das, “Steps for Linear Regression Algorithm (Simplified),” 5 Apr 2021. [Online]. Available: https://medium.datadriveninvestor.com. [Accessed 11/03/2024].

[37] S. Hogarty, “Decision trees: Definition, analysis, and examples,” 9 Nouvember 2022. [Online]. Available: https://www.wework.com. [Accessed 05/03/2024].

[38] “What is random forest?,” [Online]. Available: https://www.ibm.com. [Accessed 05/03/2024].

[39] “What is a decision tree (parts, types & algorithm examples),” [Online]. Available: https://slickplan.com/. [Accessed 06/03/2024].

[40] “Decision Tree Classification Algorithm,” [Online]. Available: https://www.javatpoint.com. [Accessed 05/03/2024].

[41] “Decision Tree Algorithm in Machine Learning: Advantages, Disadvantages, and Limitations,” 15 October 2022 . [Online]. Available: https://www.analytixlabs.co.in. [Accessed 06/03/2024].

[42] “ What is a Random Forest?,” [Online]. Available: https://deepai.org. [Accessed 05/03/2024].

[43] “ hyperparameters-random of forests,” [Online]. Available: https://www.linkedin.com. [Accessed 05/03/2024].

[44] “Random Forest AlgorithmRandom Forest Algorithm,” 2023. [Online]. Available: https://www.simplilearn.com/. [Accessed 07/03/2024].

[45] “Algorithme N°2 – Comprendre comment fonctionne un random forest en 5 min,” [Online]. Available: https://france.devoteam.com. [Accessed 07/03/2024].

[46] “What is random forest?,” [Online]. Available: https://www.ibm.com. [Accessed 07/03/2024].

[47] “Difference between Random Forests and Decision tree,” [Online]. Available: https://stats.stackexchange.com/. [Accessed 07/03/2024].

[48] L. AI, “Random Forest Algorithm,” 2023. [Online]. Available: https://medium.com. [Accessed 07/03/2024].

[49] N. Donges, “Random Forest: A Complete Guide for Machine Learning,” [Online]. Available: https://builtin.com. [Accessed 07/03/2024].

[50] A. Sharma, “Random Forest vs Decision Tree | Which Is Right for You?,” 16 Feb 2024. [Online]. Available: https://www.analyticsvidhya.com. [Accessed 07/03/2024].

[51] F. Tabsharani, “support vector machine (SVM),” [Online]. Available: https://www.techtarget.com. [Accessed 08/03/2024].

[52] V. Kanade, “What Is a Support Vector Machine? Working, Types, and Examples,” [Online]. Available: https://www.spiceworks.com. [Accessed 08/03/2024].

[53] “Support Vector Machine (SVM) Algorithm,” 10 Jun 2023. [Online]. Available:

https://www.geeksforgeeks.org. [Accessed 08/03/2024].

[54] “Support Vector Machine Algorithm,” [Online]. Available: https://www.javatpoint.com. [Accessed 08/03/2024].

[55] S. Ray, “Learn How to Use Support Vector Machines (SVM) for Data Science,” [Online]. Available: https://www.analyticsvidhya.com. [Accessed 08/03/2024].

[56] V. Kanade, “What Is a Support Vector Machine? Working, Types, and Examples,” [Online]. Available: https://www.spiceworks.com. [Accessed 08/03/2024].

[57] F. Tabsharani, “Articles by Fred Tabsharani”.

[58] a. Mar, “Advantages and disadvantages of SVM,” 2012. [Online]. Available: https://stats.stackexchange.com. [Accessed 09/03/2024].

[59] “Support Vector Machine (SVM) Algorithm,” 10 Jun 2023. [Online]. Available: https://www.geeksforgeeks.org. [Accessed 08/03/2024].

[60] “What is logistic regression?,” [Online]. Available: https://www.ibm.com/. [Accessed 20/03/2024].

[61] Sayed-Qasim, “Logistic Regression in Predictive Analytics,” [Online]. Available: https://www.linkedin.com. [Accessed 20/03/2024].

[62] “What is Logistic Regression?,” [Online]. Available: https://aws.amazon.com/. [Accessed 21/03/2024].

[63] “Logistic Regression in Machine Learning,” 2024. [Online]. Available: https://www.geeksforgeeks.org. [Accessed 22/03/2024].

[64] M. Banoula, “An Introduction to Logistic Regression in Python,” [Online]. Available: https://www.simplilearn.com. [Accessed 22/03/2024].

[65] M. Pant, Advantages and Disadvantages of Logistic Regression..

[66] J. Harris, *“A machine learning approach to forecasting consumer food prices”,* Canada, Dalhousie University,Halifax, Nova Scotia, Aug. 2017.

[67] H. Deléglise, *“Food security prediction from heterogeneous data combining machine and deep learning methods.”,* Vol. 190, 116189. 2022..

[68] S.Gholami, *“Food security analysis and forecasting: A machine learning case study in southern Malawi”,* Data & Policy, Vol. 4, e33. 2022.

[69] J. Drake, *“Food insecurity and disasters: predicting disparities in total and first-time food pantry visits during the COVID-19 pandemic ” Food Security,* Vols. Vol. 15(2),, pp. p. 493-504.

[70] P. Tamasiga, *Forecasting disruptions in global food value chains to tackle food insecurity: The role of AI and big data analytics–A bibliometric and scientometric analysis”.,* vol. Vol. 14, Journal of Agriculture and Food Research, 2023, p. 100819.

[71] U. Shafi, *“Tackling food insecurity using remote sensing and machine learning based crop yield prediction”,* 2023.

[72] H. A. Burhan, *“Comparison of prediction performances of regression models in machine learning: an application on the turkish mercantile exchange wheat index”,* Nişantaşı Üniversitesi Sosyal Bilimler Dergisi , 2023 , pp. p. 602-623.

[73] X. Xu, *“Commodity price forecasting via neural networks for coffee, corn, cotton, oats, soybeans, soybean oil, sugar, and wheat”,* Intelligent Systems in Accounting, Finance and Management, 2022, pp. Vol. 29(3), p. 169-181.

[74] Y. Ahn, *“HungerGist: An Interpretable Predictive Model for Food Insecurity”,* In 2023 IEEE International Conference on Big Data, 2023 , pp. . p. 1591-1600.

[75] Z. Bechere M'hamed Ayoub, *Prediction Model for Forests Fire Spread in M’sila,* 2023.

[76] H. Report, “Price Volatility and Food Security”.

[77] “American Heart Association”.

[78] “https://www.foodunfolded.com/,” [Online].

[79] “Prevention Web - How to mitigate the effects of climate change on global food security”.

[80] “we work: Decision trees: Definition, analysis, and examples by Steve Hogarty,” November 9, 2022.

[81] “java point: Decision Tree algorithm in machine learning”.

[82] “Rebellion Research - What Are the Disadvantages of Random Forest”.

[83] “javatpoint- machine-learning-support-vector-machine-algorithm”.

[84] “Towards Data Science. (Source: "Everything About SVM Classification Above and Beyond")”.

[85] “stats.oarc.ucla.edu_logistic-regression”.

[87] “report says Nearly one-third of the DMV has had food insecurity in the past year,” [Online]. Available: https://www.nbcwashington.com. [Accessed 10/02/2024].

[88] A. Mohan, “Decision Tree Algorithm With Hands-On Example,” 23 Jan 2019. [Online]. Available: https://medium.datadriveninvestor.com. [Accessed 06/03/2024].

[89] “Random Forest Algorithm,” [Online]. Available: https://www.javatpoint.com. [Accessed 07/03/2024].

[90] V. Kanade, “What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices,” 2024. [Online]. Available: https://www.spiceworks.com. [Accessed 22/03/2024].

الملخص

في هذا البحث، تم استعمال ست خوارزميات لتعلم الآلة (الانحدار الخطي البسيط، الانحدار المتعدد الحدود، الانحدار اللوجستي، الغابات العشوائية، شجرة القرار، وآلة متجه الدعم) للتنبؤ بمخاطر الأمن الغذائي وتصنيفها. تم جمع مجموعة بيانات شاملة تتضمن بيانات الأسعار وبيانات الطقس التي تؤثر على مخاطر الأمن الغذائي، واستخدمت لتدريب واختبار النماذج. تم استخدام معايير الأداء مثل الدقة والنقاوة وإعادة الاستدعاء لتقييم النماذج. أظهرت النتائج أن جميع الخوارزميات قدمت أداءً جيدًا، ولكن نموذج الانحدار المتعدد الحدود حقق أفضل معامل ارتباط بالنسبة لنماذج التنبؤ، بينما حققت شجرة القرار أعلى دقة في نماذج التصنيف. يسلط هذا البحث الضوء على فعالية تعلم الآلة في التنبؤ بمخاطر الأمن الغذائي وتصنيفها، ويبرز أهمية استغلال التقنيات المتقدمة للحد من هذه المخاطر وضمان سلامة الغذاء للمستهلكين.

كلمات مفتاحية: الأمن الغذائي، التعلم الآلي، السلامة الغذائية، توقع المخاطر، الانحدار، التصنيف

.

Summary

In this thesis, six machine learning algorithms (simple linear regression, polynomial regression, logistic regression, random forests, decision tree, and support vector machine) were explored to predict and classify food safety risks. A comprehensive dataset including price data and weather data affecting food safety risks was collected and used to train and test the models. Performance metrics such as accuracy, precision, and recall were used to evaluate the models. The results showed that all algorithms performed well, but the polynomial regression model achieved the best correlation coefficient for prediction models, while the decision tree achieved the highest accuracy for classification models. This research highlights the effectiveness of machine learning in predicting and classifying food safety risks and underscores the importance of leveraging advanced techniques to mitigate these risks and ensure food safety for consumers.

Keywords: Food security, machine learning, food safety, risk prediction, regression, classification.

Résumé

Dans ce mémoire, six algorithmes d'apprentissage automatique (régression linéaire simple, régression polynomiale, régression logistique, forêts aléatoires, arbre de décision et machine à vecteurs de support) ont été explorés pour prédire et classer les risques de sécurité alimentaire. Un ensemble de données complet comprenant les données de prix et les données météorologiques affectant les risques de sécurité alimentaire a été collecté et utilisé pour entraîner et tester les modèles. Des mesures de performance telles que la précision, la pureté et le rappel ont été utilisées pour évaluer les modèles. Les résultats ont montré que tous les algorithmes ont offert de bonnes performances, mais le modèle de régression polynomiale a obtenu le meilleur coefficient de corrélation pour les modèles de prédiction, tandis que l'arbre de décision a atteint la plus haute précision pour les modèles de classification. Cette recherche met en évidence l'efficacité de l'apprentissage automatique pour prédire et classer les risques de sécurité alimentaire, et souligne l'importance d'exploiter les techniques avancées pour atténuer ces risques et assurer la sécurité alimentaire des consommateurs.

Mots clés: Sécurité alimentaire, apprentissage automatique, sureté alimentaire, prévision du risque, régression, classification