WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Master Thesis

Detecting bias in data-driven artificial intelligence systems

Nicole Gogol

Subject Area: Digital Economy

Studienkennzahl: h12125133

Supervisor: Dr. Sabrina Kirrane

Date of Submission: 10. October 2023

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK
UND OPERATIONS MANAGEMENT DEPARTMENT
OF INFORMATION SYSTEMS AND OPERATIONS
MANAGEMENT

Contents

List of Figures

List of Tables

List of Acronyms

AIF360 - AI Fairness 360

AI - Artificial Intelligence

DIR - Disparate Impact Remover

FAO - Food and Agriculture Organization

GFSI - Global Food Security Index

LIME - Local Interpretable Model-agnostic Explanations

ML - Machine Learning

RSME - Root Mean Squared Error

R-squared (R^2) - The coefficient of determination in regression analysis

SHAP - SHapley Additive exPlanations

UN - United Nations

WFP - World Food Programme

xAI - Explainable Artificial Intelligence

Acknowledgement

Für meine geliebte Cousine Karina, weil du immer noch ein essenzieller Teil von mir bist und von allem, was ich tue. Das wird so bleiben - für immer.

Von Herzen Danke, Mama, Arthur, Papa und Finn, für alles. Eure bedingungslose Liebe und Unterstützung ist und bleibt die wichtigste Stütze in meinem Leben.

Abstract

The study at hand explores the complex relationship between machine learning models and their potential biases, specifically in the context of agriculture and food security. With the increasing adoption of predictive models in the agricultural sector, the study highlights the need to understand and mitigate biases that may inadvertently affect model outcomes. Leveraging the AIF360 library, the study unraveled the fairness of models after introducing biases in two different use cases: regional differences and crop types. For both use cases, bias detection tools were effective in identifying discrepancies across sub-regions as well as different crop types. While some mitigation techniques were promising, others at times exacerbated the biases. The results emphasize the complexity of balancing model accuracy with fairness. As the link between food security and Machine Learning grows stronger, this study underscores the need for proactive bias detection and mitigation, and lays the groundwork for future research in this area.

1 Introduction

According to the United Nations (UN), the world's population is expected to exceed 9 billion by 2050 [26]. Hereby, one of the most critical sustainability challenges is to find ways to achieve food security in a world with a population this high and while reducing further environmental degradation [78]. The same organization defines food security as:

> "*... all people, at all times, have physical, social, and economic access to sufficient, safe, and nutritious food that meets their food preferences and dietary needs for an active and healthy life*" [66].

Navigating the path to achieving food security in a world that is grappling with the ramifications of climate change is particularly difficult. Climate change holds pronounced risks to food security and its effects are evident in scientific studies on food security [47]: It affects crop production [82, 47], inter alia through evolving weather conditions and the increasing frequency and severity of extreme weather incidents [60]. Expected increases in the frequency and intensity of extreme weather events, especially floods and droughts, will further affect food production [42, 73] and thereby also affect food prices. Alongside climate-related concerns, food security is tightly interwoven with other dimensions, such as the limited availability of natural resources, the unpredictability of agricultural markets, and overall technological and socio-cultural changes all play a role in the food security equation [34]. This marks a big societal and economic problem carrying profound implications, affecting health, well-being, and economic stability. Disruptions in food supply can lead to price spikes, social unrest, and even geopolitical conflicts [36]. Countries with food insecurities face malnutrition, stunted economic growth, and increased vulnerability to other societal challenges. Recent data from the Food and Agriculture Organization of the UN (FAO) underscores the magnitude of this challenge by stating that almost a billion people lack adequate calories and more than two billion lack adequate nutrients [1].

Simultaneously, there is an unprecedented surge in technological advancement, with the development of Artificial Intelligence (AI) emerging as transformative force across multiple industries. By now, AI is not only applied in the finance, manufacturing, and healthcare fields, but also in the food security and agriculture sector [15]. By the same token, the world's rising food demands are commencing to be met with technological developments and AI [34]. Therefore, recent developments in agriculture and food security have brought a boost in Machine Learning (ML) applications. These devel-

opments range from soil and water management to crop management, with a particular prominence in crop yield prediction [79].

While the significance of AI continues to expand, it is equally important to identify and address its inherent concerns. One of the most pressing concerns across the entire AI landscape is bias. This bias can be systemic [31] or stem from historical data which potentially encodes historical, human-induced bias [71]. The difficulty lies in identifying this bias, with fairness metrics emerging as the primary tool, complemented by visualization and model interpretability tools [57, 61, 64, 20, 6]. Especially in domains as critical as food security, it is of utmost importance to ensure that such biases are maximally reduced or even eliminated and do not compromise their validity and integrity [5].

1.1 Research Questions

As ML applications in food security spread, the intersection of these domains becomes central. While both, ML in food security on the one hand and bias detection in AI/ML on the other hand are witnessing a significant increase in research, there is a clear research gap at their intersection. Moreover, in the context of an area as vital as food security, the challenges of bias detection become paramount [7, 21]. This thesis is intended to contribute to this important topic by answering the following overarching research question:
How can bias be detected in AI based yield predictions for food security? This question is answered through the following three sub-questions.

RQ1.1 What does the chosen bias detection method reveal about its impact on AI-based yield predictions for food security?

RQ1.2 How effective are the selected bias mitigation techniques in reducing bias in AI-based food security yield prediction models?

RQ1.3 How do potential biases in AI-based yield prediction models affect food security forecasts?

1.2 Structure of the Thesis

Beginning with Chapter 1, the thesis emphasizes the relevance of dealing with bias in ML models for food security. Chapter 2 delves into distinct streams of literature and carefully weaves them together to lay a robust groundwork for the thesis. This thorough examination of the state of the art not only provides context, but also identifies a gap that this thesis aims to fill. Next, Chapter 3 presents the chosen research approach, explaining why action research is the chosen methodology and describing the iterations undertaken during the study. It serves as a bridge from the theoretical foundation to the practical application outlined in the following chapters. Chapter 4 starts the technical journey. It first presents a detailed description of the data retrieval process, emphasizing the merging of different datasets and the associated cleaning needed to ensure their usefulness. From this foundation, the chapter dives into a comprehensive data exploration that yields insights and addresses two distinct use cases. Leveraging these discoveries, the chapter proceeds to build baseline ML models for both use cases, providing a reference point for subsequent analyses. The focus then moves to understanding the importance of different features in the trained model. In particular, this understanding serves as a basis for deliberately introducing bias into the models. With biased models at hand, Chapter 5 dives into the application of IBM's AI Fairness 360 (AIF360) toolkit. The chapter systematically approaches bias detection using a set of fairness metrics, before moving on to implement algorithms intended to mitigate the introduced bias. Chapter 6 acts as a reflective space, examining the journey undertaken in the preceding chapters. It allows a deep dive into the discussion of the chosen bias detection and mitigation practices, their implications, and how they affect food security. Within this discussion, the research questions are answered. In addition, the chapter recognizes the inherent limitations of the study and suggests avenues for future research to maintain the continuity of the work. Finally, Chapter 7 concludes the entire thesis project by reviewing the milestones achieved, summarizing the key findings, and reaffirming the contribution of this research to the wider academic and practical landscape of ML based food security models.

2 State of the Art

This chapter is concerned with a comprehensive review of the state of the art in food security as well as in ML applications in crop yield prediction. Furthermore, this chapter reviews recent advances in bias detection and bias mitigation. Thereby, it provides a basis for understanding the link between food security and predictive fairness tools.

2.1 Food Security

Food security models are multi-layered and complex [25]. The Global Food Security Index[1] (GFSI) is a dynamic benchmarking model measuring drivers of food security across 113 countries [36]. It is based on three key pillars. The first one, affordability, looks at consumers' purchasing power, their vulnerability to price fluctuations, and the existence of policies that provide support in times of food crisis. This is measured by metrics such as gross domestic product, poverty rates, and the existence of safety net programs. The second pillar, availability, examines the reliability of national food supplies, taking into account domestic production volumes, possible threats to supply chains, and the nation's commitment to agricultural research. The third pillar, Quality and Safety, assesses the nutritional adequacy and safety standards of available foods, also considering dietary diversity. By assessing and weighting each indicator of these three pillars, the GFSI provides an annual, comprehensive view of food security for each country.

In 2022, the GFSI depicts a deteriorating global food environment for the third year in a row. Since 2019, the GFSI has been declining due to economic and socio-political shocks like the Covid-19 pandemic and the war in Ukraine. Among others, these shocks lead to an immense cost increase for food and a vulnerable global food environment. The top ten performers in the GFSI in 2022 are comprised of eight high-income countries in Northern and Central Europe, led by Finland, Ireland, and Norway as well as by high-income Japan and Canada. On the other end of the spectrum, six of the bottom ten scoring nations in 2022 come from Sub-Saharan Africa, including the democratic republic of Congo, Nigeria, and Sudan. The aforementioned imbalance is also clearly evident in the GFSI statistics and this gap has been steadily widening in recent years. The disparity between the best performing country, Norway (83.7/100) and the worst performer Syria (36.3/100) amounts to 47.4 [36]. In practice, this translates into very divergent living conditions and challenges.

[0]Note: All links mentioned in this thesis have been last accessed on 29.09.23.

[1]http://foodsecurityindex.eiu.com/

For Norwegians, the high score reflects the secure access to a wide range of quality foods and a stable food market. In contrast, Syria's low score points to severe food shortages, where large proportions of the population could face hunger, malnutrition, and related health concerns. These inequalities can have profound long-term implications, impacting on everything from children's cognitive development to societal stability. Overall, the report finds that the problem of hunger is most acute in Africa, more specifically in Sub-Saharan Africa. In addition, the increase of the world's population to more than 9 billion people by 2050, as projected by the UN, will primarily occur in developing countries [26]. Such levels of growth will demand an increase in food supply and agricultural production of up to 75% by 2050 [26]. At the same time, the needs of developing countries as a whole will double, including Sub-Saharan Africa by 250% [69]. In order to combat this, various programs and efforts are in place.

As a key agency of the UN, the World Food Programme (WFP) is at the leading forefront of the fight against global malnutrition and food insecurity. One of its most prominent efforts is its nutrition program[2], which focuses on vulnerable populations such as women, children, and infants. As part of this initiative, WFP inter alia provides specialized nutritious foods, while stressing dietary diversification and fortification of staple foods [83]. In addition, WFP's Purchase for Progress[3] initiative is an important tool in empowering smallholder farmers. By linking them to markets, ensuring fair prices and training them in post-harvest handling, the initiative aims at minimizing food losses and increasing overall agricultural productivity [51].

In parallel, the FAO is promoting sustainable agricultural practices that combine economic profitability on the one hand and environmental responsibility on the other hand. This is reflected in efforts towards conservation agriculture, agroecology, and integrated pest management [53]. For instance, the Global Information and Early Warning System[4] of the FAO is designed to timely avert food crises by closely monitoring food supply and demand and accordingly issue warnings of potential food shortages [55]. Another key initiative of the FAO is the Codex Alimentarius Commission[5], which is a joint venture with the World Health Organization. This joint initiative is central to establishing international food standards, promoting food safety and quality, and fostering fair trade practices [17].

[2]https://www.wfp.org/nutrition/
[3]https://www.wfp.org/purchase-for-progress/
[4]https://www.fao.org/giews/en/
[5]https://www.fao.org/fao-who-codexalimentarius/en/

More broadly, the UN has set clear guidelines through Sustainable Development Goal 2 [6], which targets eradicating hunger, strengthening food security, and promoting sustainable agriculture by 2030 [80]. Similarly, the Zero Hunger Challenge[7], another UN initiative, aims for global access to food, zero cases of malnutrition among children under two, sustainable food systems, increased productivity of smallholder farmers, and a substantial decrease in food waste [11]. The Committee on World Food Security complements these endeavors by providing an integrative platform for stakeholders and promoting policy alignment [50]. However, and despite various efforts and UN programs and policies, food security in the global sense has not been achieved yet. Worldwide, there is an extreme and complex imbalance in food security seeing as certain regions experience a high surplus of food, while at the same time, it involves increasing malnutrition for others [69].

Despite the deteriorating global food environment and its vulnerability towards shocks, one positive development can be seen in the GFSI 2022: Access to agricultural technology, education and resources has risen by 10.1% [36]. Modern technologies can increase global food production by increasing soil fertility, advancing genetics, harnessing solar energy, and using AI to improve crops [69]. Among these modern technologies, AI is currently emerging as a dominant force driving innovations across multiple industries, including food security and agriculture [15].

2.2 ML in Crop Yield Predictions

In recent years, there has been a growing body of literature on various applications of ML in agriculture and food security. On the one hand, fields such as soil management and water management are extensively investigated. Among others, Motia and Reddy [59, 41] emphasize the use of ML techniques for predicting and assessing soil properties, leading to improved soil health management. Huang et al. [35] on the one hand demonstrate the various benefits achieved when utilizing ML to predict numerous water indicators in both natural and engineered water systems, including water quality prediction and contamination mapping. On the other hand, also with respect to crop management, there are various theoretical ML approaches. As argued by Waldamichael et al. [81], Amitab et al. [76], and others, these include early disease detection in crops, as well as weed detection as examined by Islam et al. [37] and Osorio et al. [62] among others.

[6] https://www.un.org/sustainabledevelopment/hunger/

[7] https://sdgs.un.org/partnerships/zero-hunger-challenge-zhc

Table 1 below lists ML approaches for crop yield prediction in the literature of the last seven years. The table is structured chronologically and contains information on the year of publication, the examined crop, the features used to predict the crop, the ML algorithms applied as well as the model results. It is compiled using the keywords *machine learning in crop prediction*, *machine learning in yield prediction*, *machine learning in crop yield prediction*, *artificial intelligence in crop prediction*, *artificial intelligence in yield prediction*, and *artificial intelligence in crop yield prediction*. Using these keywords, a variety of approaches emerges from the literature. Upon closer inspection, these publications can be systematically categorized according to their algorithmic direction.

Crop yield prediction using ML Crop yield prediction is essential for achieving food security and improving agricultural productivity. With the emergence of ML, a multitude of studies use this technology to improve the accuracy and robustness of predictions. Shahhosseini et al. [75] propose an ML ensemble framework for maize yield forecasting, focusing on the integration of weather data. Their optimized model shows a solid RSME of 9,5%.

Deep Learning in crop yield prediction Deep learning, a subset of ML, demonstrates a notable effectiveness in crop yield prediction. Schwalbert et al. [73] use Long-Short Term Memory networks for predicting soybean yields using satellite images and weather data. Their models outperform other algorithms, such as a Linear Regression, in accuracy for most prediction data, highlighting the applicability of deep learning in this area. Luo et al. [54] also successfully use Long-Short Term Memory networks to obtain spatial distributions of wheat harvesting areas to predict yield. Pantazi et al. [65] combine supervised self-organizing maps and artificial neural networks to predict wheat. By including soil data and satellite images, their model shows high accuracy, especially for the low yield class. Wang et al. [82] further illustrate the effectiveness of deep learning algorithms in agriculture by applying such a model to predict wheat yield in China during winter, obtaining a strong R^2 of 0.77.

Random Forest in crop yield prediction Random Forest proves to be a robust method for crop yield prediction. Josephine et al. [42] apply the Random Forest algorithm for predicting millet yield and achieve a very high accuracy of 99.74%. In much the same vein, Kumar et al. [27] emphasize the algorithm's ability to predict yields using historical data, including variables such as temperature, humidity, and rainfall. Their method exploits the power of Random Forest to build a robust model from historical data, thus effectively predicting yields in the agricultural sector by identifying the optimal crops for certain weather conditions in the field. Charoen-Ung et al. [16] further extend the use of the Random Forest algorithm to predict sugarcane yield quality, surpassing human experts in their assessments.
Summarizing, recent crop prediction research frequently explores neural networks and employs the Random Forest algorithm. These studies often incorporate a mix of data, including climate, soil, fertilizer, vegetation, and precipitation data. Interestingly, a significant portion of the studies rely on satellite imagery for data collection.

In essence, there is a growing interest in ML applications for crop prediction. All of the papers reviewed, regardless of the algorithm used, produced satisfactory results, emphasizing the applicability and power of these methods. However, it is worth noting that the current state of the art tends toward theoretical exploration rather than practical implementation.

Reference	Year	Crop	Features	Algorithms	Results
[54]	2022	Wheat	Satellite data, soil data, climate data, vegetation data	Random Forest, Light Gradient Boosting Machine	R2: 6.7% RMSE: 6.3%
[82]	2020	Wheat	Climate data, vegetation indices	Deep Neural Network	R2: 0.77, RMSE: 721 kg/ha
[75]	2020	Corn	Soil data, crop data, historical yield data, climate data	Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine	RMSE: 9.56%
[42]	2020	Millet	Climate data, rain data	Random Forest	Accuracy: 99.7%
[73]	2020	Soybean	Satellite data, weather data	Regression, Random Forest, Long Short-Term Memory	MAE: 0.24
[16]	2019	Sugarcane	Crop data, rain data, fertilizer data	Random Forest	Accuracy: 71.88%
[12]	2019	Wheat	Crop data, yield data, satellite data, climate data	Random Forest, Support Vector Machine, Neural Network	R2: 0.75
[65]	2016	Wheat	Satellite data	Artificial Neural Network	Accuracy 81.65%
[27]	2016	Rice	Weather data, precipitation data.	Support Vector Machine	RSME: 0.39

Table 1: Summary of ML Approaches for Crop Yield Prediction

2.3 Bias Detection and Mitigation

While ML demonstrates its potential in crop prediction through several successful applications, these algorithms are not free from certain pitfalls such as bias. In general, algorithmic biases can occur due to several factors, making the topic of bias in itself very intersecting and therefore complex [10]. For instance, a hiring algorithm can on the one hand use biased training data and thereby predominantly favour the resumes of male applicants (algorithmic bias) as they were historically preferred over female applicants (selection bias) [30]. On the other hand, the interpretation of the algorithm's results can at the same time be biased, e.g., when assuming that certain roles are better suited for a particular gender, this assumption further compounds the bias in the hiring process (interpretation bias) [61]. Depending on the means by which the bias is generated, literature provides different definitions of bias. Ntousi et al. [61] define bias as: "*... the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair*". Mehrabi et al. subdivide bias further into three categories [57]:

Data bias addresses bias that is inherent in the data, meaning that the algorithm absorbs pre-existing inequalities including historical and social bias.

Algorithm bias refers to bias that is introduced exclusively by the algorithm, for instance through optimization functions or regularization.

User interaction bias is induced by the interaction with the user, as the interface lets the user impose his behaviour for a self-selected interaction that can inter alia include behavioural or presentation bias [57].

In much the same vein, Ntoutsi et al. [61] classify bias detection and mitigation into three categories.

Pre-processing approaches focus on addressing bias in the input data. The underlying idea is that the fairer and more balanced the training data is, the less biased the resulting AI model will be [61]. In order to achieve this, several methods manipulate the original data distribution by modifying the class labels of selected instances that are close to the decision boundary [43], by assigning varying weights to instances based on their group membership boundary or by sampling from each group [45]. These methods aim at balancing the protected and unprotected groups in the training set [61]. By making the data more balanced, the subsequent ML model is less likely to generate biased results.

In-processing approaches intervene during the training process of the ML model to minimize bias. In contrast to pre-processing, in-processing approaches focus on altering the training algorithm itself. Furthermore, it aims at restating the classification problem by including a model's discrimination behaviour into the objective functions. This can be done through regularization, constraints, or training on latent target labels. For instance, a reguliser can be incorporated to minimize indirect bias [61]. Kamiran et al. [45] change the splitting criterion of decision tree algorithms to account for the effect of splitting on the protected attributes, whereas Dwork et al. [22] redefine the classification problem by treating similar individuals similarly, i.e. reducing an arbitrary loss function subject to the individual fairness constraint. As opposed to this, Krasanakis et al. [48] consider the presence of latent fair classes. For such classes, they suggest changing the in-training weights of the instances iteratively. Although not many, there are also in-processing approaches which do not relate to classification. In an unsupervised model, Samadi et al. [72] impose equal reconstruction errors for both, protected and unprotected groups [61].

Post-processing approaches are implemented after an ML model is trained, tweaking the model's predictions to reduce bias. Essentially, these approaches are comprised of two strategies. On the one hand white-box approaches to post-processing involve altering the model's internals or probabilities, such as correcting probabilities in Naive Bayes models [13] or changing the class label at the leaves of decision trees [45]. On the other hand, black-box approaches to post-processing involve altering the model's predictions by promoting the proportionality of decisions between protected and unprotected groups [46] or by overlaying a classifier on top of a base classifier [3]. The advantage of this approach is that it allows for correcting bias without retraining the model.

Beyond these approaches to bias in ML literature, researchers use the terms bias and unfairness interchangeably [57, 61, 64, 20]. In much the same vein, fairness in this context is referred to as "*... the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics*" [57]. For this reason, bias detection methods are commonly equated to quantitative fairness metrics, evaluating how a ML model's prediction or decision aligns with fairness criteria [4]. Table 2 provides an overview of the fairness metrics which have emerged as common state of the art metrics in recent years, along with reference papers which examine these metrics and a brief explanation of the metrics' objective.

Reference	Metric	Explanation
[64, 6, 39]	*Average Odds Difference*	Measures fairness by comparing false positive and true positive rates across groups.
[64, 6, 38]	*Demographic Parity*	Ensures predictions are independent of protected attributes, targeting equal acceptance rates across groups.
[64, 6, 23, 10]	*Disparate Impact*	Evaluates whether seemingly neutral policies affect one group more than another in model predictions.
[64, 6, 23, 32, 67, 30, 70]	*Equalized Odds*	Aims for equal true positive and false positive rates for protected and non-protected groups.
[64, 6, 23, 67, 39, 70])	*Equal Opportunity Difference*	Assesses disparities in true positive rates between protected and non-protected groups.
[64, 6, 23, 10]	*Statistical Parity Difference*	Compares the probability of positive results for protected versus non-protected groups.

Table 2: Fairness Metrics

Navigating the field of fairness metrics, it is important to note that each metric, while broadly concerned with bias and fairness, focuses on different aspects of the wider challenge. *Demographic Parity* and *Statistical Parity Difference* have a common basis in achieving equality of outcomes across groups. Both aim to provide equal likelihood of positive outcomes regardless of protected characteristic [6]. *Demographic Parity* addresses this goal directly, while *Statistical Parity Difference* quantifies disparity by calculating the gap in positive prediction rates [70]. *Disparate Impact*, based on U.S. employment law principles, examines whether a model's decisions could have a disproportionately negative impact on a protected class, thus taking a broader view of fairness [24]. On the error based fairness side, *Average Odds Difference* is a comprehensive metric. It assesses differences in both false positives and true positives, providing a richer perspective than *Equalized Odds Difference*, which concentrates solely on differences in true positive rates between groups. [39]. Finally, *Equalized Odds* is notable for its commitment to preventing models from discriminating in their errors. This metric stresses

that both types of errors, false positives and false negatives, are consistent across groups, thus guarding against models whose errors might unfairly bias one group over another [33].

While all of these metrics share the objective of fairness, their individual nuances contribute to a multi-layered picture of how a model behaves across multiple dimensions of fairness.

2.4 Fairness Toolkits

Given the importance of bias detection and mitigation, several open source tools have been developed and gained popularity in recent years. While theoretical explorations of bias provide foundational understanding, practical tools translate these concepts into actionable implementations. The following section examines tools which demonstrate this transition from theory to practice, emphasizing contributions to fairness.

Tensorflow Responsible AI[8] provides tools to promote fair and transparent ML practices within Google's Tensorflow framework [2]. It aims at addressing model fairness, interpretability, and robustness. Tensorflow's Fairness Indicators facilitate a systemic way to evaluate models by means of fairness metrics, such as *Demographic Parity*, *Disparate Impact*, and *Equal Opportunity Difference*. This allows researchers and developers to examine disparities in model results. Besides, Tensorflow Responsible AI contains tools that explore the interpretability of ML models. The What-If Tool offers interactive exploration and visualization of model decisions, thereby enabling a deeper understanding. For instance, if a model trained to recognize items in pictures constantly misidentifies a particular category, the What-If Tool can be used to uncover these discrepancies and diagnose potential sources of bias [2]. In addition, Tensorflow Responsible AI contains a privacy library focusing on training models with differential privacy. Thereby, the tool ensures that a model does not overfit to datapoints and expose private information. For instance, if an ML model is trained on medical records, this tool can prevent patients' data to be reverse engineered from a model's output [2]. Tensorflow Responsible AI also stresses clear documentation and community collaboration on its platform [2]. Nevertheless, its broad adoption in both research and industry seems to be behind more established methods in this field. While Tensorflow Responsible AI offers tools to tackle fairness concerns, the wider community's exploration of these tools, in-depth studies, and large-scale practical applications might still be in its infancy. As AI fairness continues to

gain momentum, it's reasonable to expect more research and practical applications to emerge.

Fairlearn[9] is created by Microsoft and offers a Python package that enables developers to assess and improve fairness in their AI systems [8]. This can be done through Fairlearn's set of tools that assess potential ML biases as well as provide methods to mitigate them. At the core of Fairlearn's functionality lie two elements, an interactive visualization dashboard as well as a set of bias mitigation algorithms [8]. The toolkit's visualization dashboard on the one hand allows users to detect potential classes that may be negatively impacted by a model. On the other hand, it presents a mechanism to compare multiple models on fairness and performance metrics. Thereby, it includes a total of five fairness metrics suitable for both, classification and regression tasks. These fairness metrics inter alia include *Demographic Parity* and *Equalized Odds*, providing a detailed insight into the fairness of different groups defined by sensitive attributes such as gender or disability status. The dashboard's customization settings allow users to choose the sensitive attribute and performance metric of interest. Based on this selection, visualizations are produced that illustrate the impact of the model on different groups [8]. With respect to mitigation algorithms, Fairlearn provides post-processing and reduction algorithms. The post-processing algorithms modify the predictions of an already trained model to better match a certain fairness metric while the model's performance is maintained or even improved, but not reduced. The reduction algorithms modify data weights and iteratively retrain the models until the model matches the specified fairness metric more accurately yet still achieves a good performance [8]. For instance, Caputo [14] investigates the application of Fairlearn in addressing minority bias in healthcare data sets, specifically in diabetes diagnosis and changes in therapy for diabetic patients. Using Fairlearn, the study validates the methodology, creates a comprehensive list of metrics that detect this bias, and successfully explores mitigation strategies for this bias in diagnostic contexts [14].

AI Fairness 360[10] is a Python toolkit, developed by IBM and designed to address algorithmic bias in ML models [6]. The toolkit provides a set of fairness metrics, including *Disparate Impact*, *Statistical Parity Difference*, and *Equal Opportunity Difference*. These metrics allow for an in-depth exploration of how a model treats diverse demographic groups [57]. It is designed to identify any immanent bias within models that can be pinpointed to and assessed. With regard to bias mitigation, AIF360 includes algorithms that act at different stages of the ML process. Pre-processing techniques, such as *Reweighing* methods, alter data prior to the model training stage. In-processing methods modify the model training process, whereas post-processing methods adjust a model's predictions to comply with fairness criteria. The toolkit also provides a holistic framework for tackling fairness in order to ensure that biases are addressed across the entire lifecycle of a model. Furthermore, it serves as a collaborative platform that enables researchers to disseminate their algorithms [6]. Through its integration with popular ML libraries and provision of educational resources, AIF360 also acts as a resource and a guide to AI fairness [57]. Zhang and Zhou [87] examine the application of the AIF360 toolkit in the context of ML-based loan approval system. More specifically, they focus on potential biases against persons of diverse ethnicities. The assessment reveals that these minority groups experience higher loan denial rates. Using AIF360, the researchers detect these biases and implement techniques to mitigate them, resulting in a fairer decision-making model [87].

It can be deduced that fairness metrics serve as the primary measure of model bias by providing clear, numerical insight into how a model's predictions or decisions compare to fairness criteria. In the context of bias detection, fairness metrics can also be complemented by model interpretability tools [63]. Tools such as Local Interpretable Model-agnostic Explanations[11] (LIME) and SHapley Additive exPlanations[12] (SHAP) dive deeper into the model's decision-making process. By revealing the importance of different features in a given prediction, these tools make the model's predictions more explainable. Although they play a secondary role to fairness metrics, they provide valuable information about the importance of features and potential sources of bias [28].

[8] https://www.tensorflow.org/responsible_ai
[9] https://fairlearn.org/
[10] https://github.com/Trusted-AI/AIF360
[11] https://interpret.ml/docs/lime.html
[12] https://shap.readthedocs.io/en/latest/

LIME targets individual predictions by perturbing the input data of a given instance, monitoring the resulting changes in the model's predictions, and then fitting a simpler, more interpretable model to match the model's logic for that specific instance [49]. Such localized explanations are especially useful for revealing latent biases [85]. To exemplify, if an ML model in recruiting tends to unfairly weight age as a key factor for tech roles, a LIME analysis can expose this bias, even if the overall behavior of the model appears balanced. Although LIME's power lies in its granular, localized insights, it's noteworthy that these explanations are specific to particular instances. Therefore, they may not reflect the overall behavior of the model [49]. Yet, in the effort to make ML models more interpretable, LIME provides a valuable tool allowing stakeholders to dive deep into model decisions and understand their rationale.

SHAP is emerging as a key technique designed to shed light on the decision-making process of ML models. Inspired by the Shapley value in game theory, SHAP quantifies the influence of each feature in a dataset, thereby signifying its contribution in the given prediction [77]. The core of SHAP lies in its dual ability to reveal insights into the overall behavior of the model, i.e. global interpretability, while simultaneously revealing individual predictions, i.e. local interpretability [58]. SHAP is designed around three core principles: consistency (i.e. features maintain their contribution regardless of model changes), linearity (i.e. SHAP values correspond to feature weights in linear models), and local accuracy (i.e. SHAP values for a prediction sum to the difference between that prediction and the model's average output) [77]. Thus, by using SHAP, researchers can identify which features predominantly influence ML predictions.

Both tools are model agnostic, making them applicable across different ML models. In addition, biases in the training data itself may emerge through these explanations. If either LIME or SHAP suggests that certain features are overly emphasized, this could reflect an existing bias in the dataset on which the model was trained.

Overall, the tools studied above allow for standardized evaluations and have facilitated the widespread detection of bias in existing AI systems. Fairness metrics are the bedrock of bias detection in ML. They enable the rigorous and quantitative analysis required to assess bias. At the same time, interpretation tools can augment these tools, ensuring a comprehensive view of the models' biases. However, it is important to note that even by using these tools, researchers face challenges in bias detection. One such challenge is the definition of fairness. What is considered fair is context-specific and therefore difficult to define universally. Binns [7] argues that fairness in ML should be informed by political and moral philosophy. Another challenge faced by researchers are trade-offs in terms of performance. Corbett-Davies et al. [21] state that achieving fairness can come at the expense of model accuracy, in which the the right balance is often application-specific. Furthermore, Pagano et al. [63] address the evaluation dilemmas, which arise as the use of a single metric might not capture the holistic bias within a system. However, multiple metrics might conflict and give different perspectives on the present bias, leading to evaluation challenges.

3 Methodology

The following chapter discusses the research method, highlighting the applied Action Research approach. In addition, the different iterations undertaken for bias detection and bias mitigation in the context of this thesis project are outlined.

3.1 Action Research

In order to address the aforementioned research gap, an Action Research approach is applied [19]. This research approach consists of cycles of Planning, Acting, Observing, and Reflecting as depicted in Figure 1 below. Action Research is often used in fields such as education, healthcare, and community development, where the goal is to improve a specific situation or address a particular problem [29]. Yet, with its reflective principles, Action Research also provides a well-suited framework for algorithm development. Its iterative cycles coherently align with the typical process of algorithmic development as it frequently deviates from a linear path and requires multiple iterations to optimize the algorithm's accuracy. The process can be repeated several times until the required performance metrics are met. The following paragraph describes how the Action Research cycle is set up in the context of this thesis, more precisely in the context of detecting and mitigating bias in food security.

Planning is the first stage in Action Research, aiming at identifying the problem to be addressed. Therefore, in the context of this thesis, Planning entails a thorough review of literature and the state of the art, examining food security fundamentals, delving into ML applications for crop yield prediction and bias detection methodologies. Once the problem is identified, the plan is put into **Action**. As part of this step, a possible solution, i.e. a suitable algorithm is implemented and the corresponding output data is collected. This leads to the **Analysis** phase, where the collected data is thoroughly examined, the model performance is evaluated, patterns are identified, and findings are derived. Then, in the **Conclusion** stage, the research is reflected. Based on the outcomes of the previous steps, the solution is evaluated in detail. If the results of the conclusion are not yet satisfactory, i.e. in case the algorithms do not perform well or no bias can be detected, the cycle is repeated until an adequate solution is found. Figure 1 outlines the steps taken in this thesis at each stage of the action research cycle more precisely.

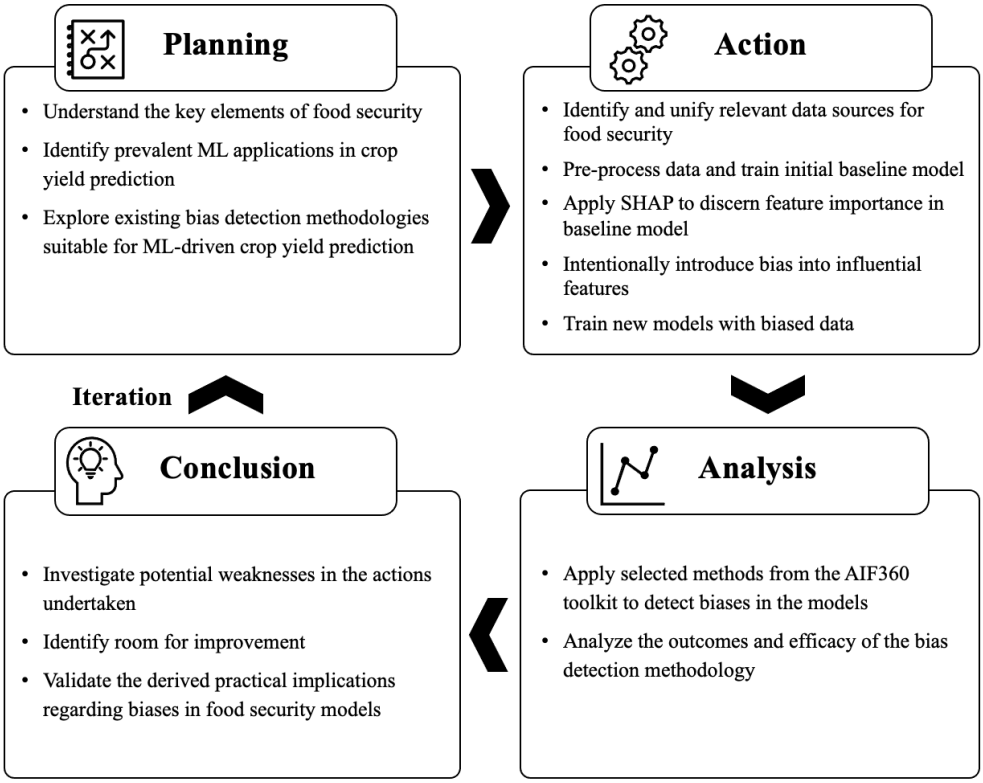

Figure 1: Action Research Cycle

3.2 Iterations

The thesis project comprises four distinct iterations, each of which builds on the learnings and insights of the previous iteration. Throughout these iterations, continuous reflection and adaptation are emphasized, ensuring that the research is both responsive and dynamic.

Iteration 1 - Initial Model Training In the first iteration, a distinct Random Forest Regressor model is trained on the dataset for each specified use case. This initial iteration is vital as it establishes unbiased baseline models and demonstrates their underlying performance capabilities. Subsequent modifications in the model can then be compared to these baseline models, allowing to measure and quantify the impact of modifications and introduced biases in later iterations.

Iteration 2 - Bias Introduction With the baseline models in place, the second iteration introduces specific biases into the dataset. This is done by significantly lowering the input values of features which prove to have a strong influence on the model's predictions. After intentionally distorting these feature values, the models are retrained to comprehend and capture the impact of the biases on their predictive abilities.

Iteration 3 - Bias Detection The third iteration involves applying bias detection methods. This is critical for measuring the extent of bias and understanding its impact on model results. It is paramount to not rely solely on a single metric, but instead, try and use multiple metrics as part of this iteration to ensure a comprehensive understanding of the biases. This phase helps quantify the differences and identify areas of concern.

Iteration 4 - Bias Mitigation Having identified and quantified the biases, the final iteration centres around applying bias mitigation techniques. Prior to this step, it is essential that a bias in the models has been identified. This allows for a comparative analysis of metrics before and after applying bias mitigation methods. The goal is to correct the identified biases and thereby improve the fairness of the ML models. This is facilitated by applying suitable bias mitigation algorithms of the AIF360 toolkit and analysing their results. Hence, this phase is essential to determine the effectiveness of different mitigation strategies in dealing with the identified biases.

4 Technical Implementation

This chapter covers the practical aspects of data retrieval, data pre-processing, and data exploration. It includes detailed steps that illustrate the transformation of raw weather and agriculture data into actionable insights within the food security domain. Subsequently, the foundation for the technical implementation is laid by training the baseline ML models to predict yield. This is followed by the application of SHAP, which is used to improve the understanding of the ML models' decisions and gain deeper insights into the corresponding food security predictions. The source code for the steps performed in this thesis project is available in the GitHub repository described in Appendix A.

4.1 Data Retrieval and Pre-processing

For the purpose of this research, the dataset in use is not downloaded as-is but curated from various sources, including the FAO[13], the UNstats[14], and the World Bank[15]. In light of this research, the primary tool used for data analysis is Python. More specifically, the Pandas library[16] used for data manipulation. The different datasets are therefore imported by means of a Pandas DataFrame as it eases an efficient and accurate data analysis [56]. The objective is to investigate the relationships of different features and their impact on crop production. Thus, a total of five datasets are merged not only containing weather information, such as temperature and rain, but also information on the region, the usage of pesticides and the crop yield.

Each of the imported datasets includes multiple columns upon initial inspection. In order to create a lean data structure, duplicate columns are systematically removed. Moreover, redundant columns such as area codes, are removed as they are considered irrelevant for the objectives of this thesis. This step allows to reduce data noise and focus on strategically important columns [84]. Renaming columns, such as converting 'value' into the more meaningful `Pesticides in tonnes` further increases clarity. In addition, data gaps, outliers, and inconsistencies are thoroughly addressed in a way that does not compromise the accuracy of the thesis project and ensures data integrity. The commonality across all datasets is the presence of the columns `Year` and `Country`. Individually, these columns do not contain unique entries. However, combining their values allows the datasets to be accurately merged

[13]https://data.apps.fao.org/catalog/dataset
[14]https://unstats.un.org/sdgs/dataportal
[15]https://databank.worldbank.org/
[16]https://pandas.pydata.org/is

on this composite key [84]. The aggregated dataset now includes the following variables: `Crop`, `Year`, `Country`, `Pesticides in tonnes`, `Yield in hg/ha`, `Avg rain in mm`, `Avg temperature`, `Region`, and `Sub-region`. All in all, the performed data pre-processing steps lead to a robust and multi-dimensional dataset streamlined for gaining insights into the complex area of food security.

4.2 Dataset Description

Overall, the dataset consists of a total of 28,242 observations containing the following environmentally and agriculturally pertinent variables.

`Region`, `Sub-region`, and `Country` are categorical variables denoting in which location the respective data was collected. This regional data adds geographic and demographic nuances to the agricultural data, thereby enriching the context and understanding of food security across different regions. Moreover, `Region` and `Sub-region` range from broader regions like North America to sub-regions like Sub-Saharan Africa. They serve as geographical classification which is of importance in the further research process of this thesis project.

`Crop` specifies the particular crop species for which the data is compiled, allowing analyses across agricultural commodities, including the ten most produced crops worldwide as of 2023 [26]. These crops are quantified in the integer variable `Yield in hg/ha`. Indicating the crop's yield in hectograms per hectare, this variable serves as a metric for measuring agricultural productivity. `Year`, captured as numeric integer variable, denotes the year in which the data was collected, i.e. from 1993 to 2013 in this dataset, providing a full temporal scope over a time period of 20 years. The floating-point variable `Average rainfall in mm` represents the average annual precipitation in millimeters. Similarly, the floating-point variable `Average temperature` represents the average annual temperature in degrees Celcius. Thereby, these two variables offer insights into the prevailing climate conditions. `Pesticides in tonnes`, another floating-point variable, measures the overall quantity of pesticides used for a particular location in a particular year.

Variable name	#Unique	Min	Mean	Max
`Year`	20	1993	2003	2013
`Yield in hg/ha`	10,218	80.00	78,559.65	501,412.00
`Avg rain in mm`	873	51.00	1,146.99	3,240.00
`Pesticides in tonnes`	1,534	0.25	38,373.61	367,778.00
`Avg temperature`	1,729	1.30	20.54	30.65

Table 3: Descriptive Statistics of Numeric Variables

Table 3 presents a high-level statistical snapshot of the variables, including their unique occurrences, minimum, mean, and maximum values. Thereby, the table highlights the wide range and central tendencies of these key agricultural factors. This summary underscores the variability of the dataset variables. For example, the wide range between the minimum and maximum values for variables such as `Yield in hg/ha` and `Pesticides in tonnes` suggests a diverse agricultural landscape. This diversity could be due to a variety of factors, such as differences in agricultural practices, climatic patterns, or economic status among the countries in the dataset.

In much the same vein, Table 4 lists the dataset's string variables, showing both the number of unique instances and the different values that each variable is associated with. This presentation shows the diversity of the data and facilitates an understanding of the distribution and uniqueness of the values within the dataset.

Variable name	#Unique	Unique values
`Country`	101	Albania, Algeria, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, and 93 others
`Crop`	10	Cassava, Maize, Plantains, Potatoes, Rice paddy, Sorghum, Soybeans, Sweet potatoes, Wheat, Yams
`Region`	5	Africa, Americas, Asia, Europe, Oceania
`Sub-region`	13	Australia and New Zealand and Melanesia, Central and Eastern Asia, Eastern Europe, Latin America and the Caribbean, Northern Africa, Northern America, Northern Europe, South-eastern Asia, Southern Asia, Southern Europe, Sub-Saharan Africa, Western Asia, Western Europe

Table 4: Descriptive Statistics of String Variables

4.3 Data Exploration

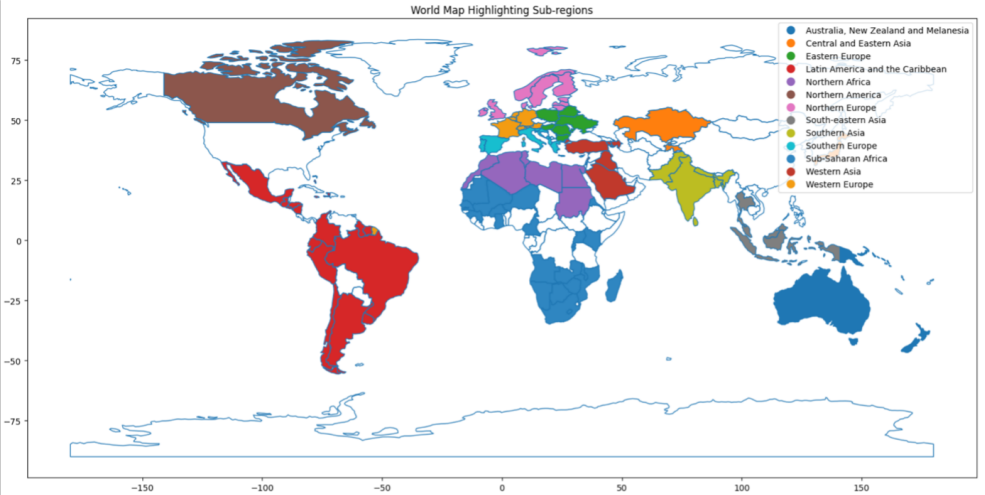

Figure 2: World Map Highlighting Sub-regions

This section takes a closer look at the individual variables. For a clearer understanding of the variables, visualizations are created using the Python Seaborn library[17].

The dataset contains 101 different countries from five regions (Europe, Africa, Americas, Asia, Oceania) clustered into 13 different sub-regions. The sub-regions and the countries they contain are highlighted in Figure 2, each sub-region identified by a different color. Having set the geographic context, the exploration continues to understand the distribution of yield across different sub-regions.

[17]https://seaborn.pydata.org/

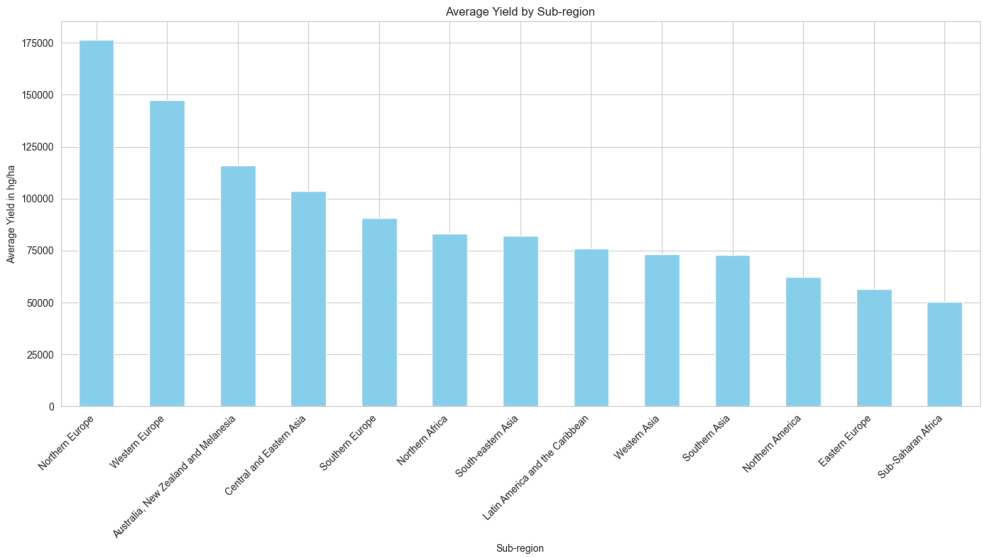

Figure 3: Average Yield by Sub-region

Figure 3 illustrates the average yield by sub-region for all 13 sub-regions, measured in hectograms per hectare. Moreover, the figure demonstrates that the yield varies significantly across these areas, underlining regional differences in agricultural productivity. While Northern Europe exhibits the largest and Sub-Saharan Africa the smallest yield, South-eastern Asia lies exactly in the middle. Having explored the overall yield averages across the sub-regions, it is now essential to deconstruct these averages further in order to understand the contribution of individual crops to these averages. In order to see which crops dominate in specific sub-regions, the stacked bar chart below is compiled.

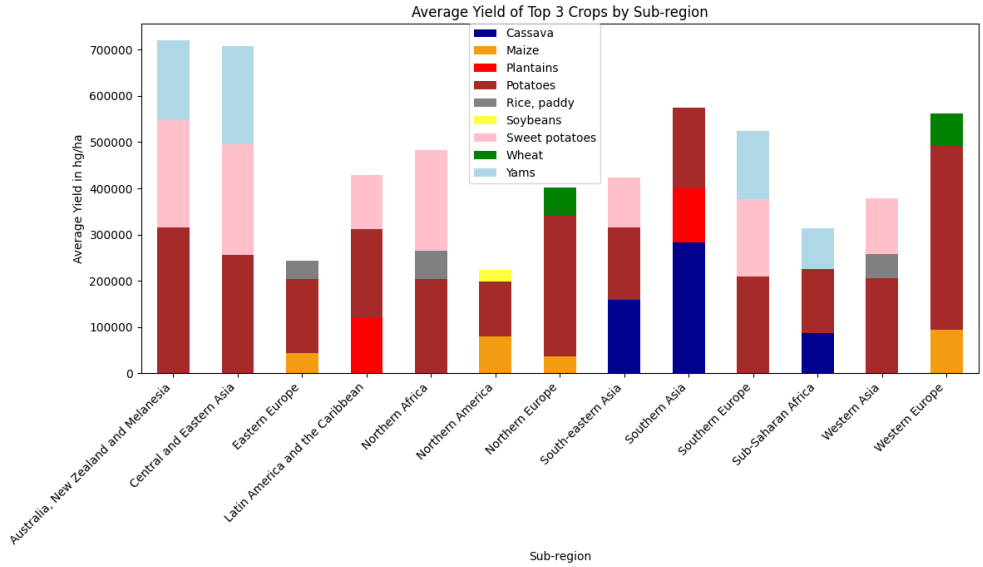

Figure 4: Average Yield of Top 3 Crops by Sub-regions

Figure 4 shows the average yield of the top three crops across all 13 sub-regions. Each of the three elements of a bar corresponds to a specific crop and is color coded for easy identification. This way, the plot provides insights into the distribution of yield among the top three crops in each sub-region and the different agricultural landscape of the sub-regions. Several sub-regions, such as Central and Eastern Asia, Southern Europe, and Sub-Saharan Africa show a balanced proportion of different crops. This is indicative of a diversified agricultural approach in which single crops do not dominate. This contrasts with e.g. Eastern Europe, Northern Africa, and Western Asia, where one or two crops appear to dominate. Such patterns highlight the crucial role of the respective top crop(s) in the agricultural output in these sub-regions.

Overall, it can be deduced that certain Sub-regions emphasize a diversified approach while others focus on specific crops, reflecting different agricultural strategies and potentially mirroring differences in regional market demands or conditions. Looking at the specific crops, it can be seen that potatoes, sweet potatoes, and yams are among the most important crops in many of the Sub-regions. This is also seen in the bar chart below.

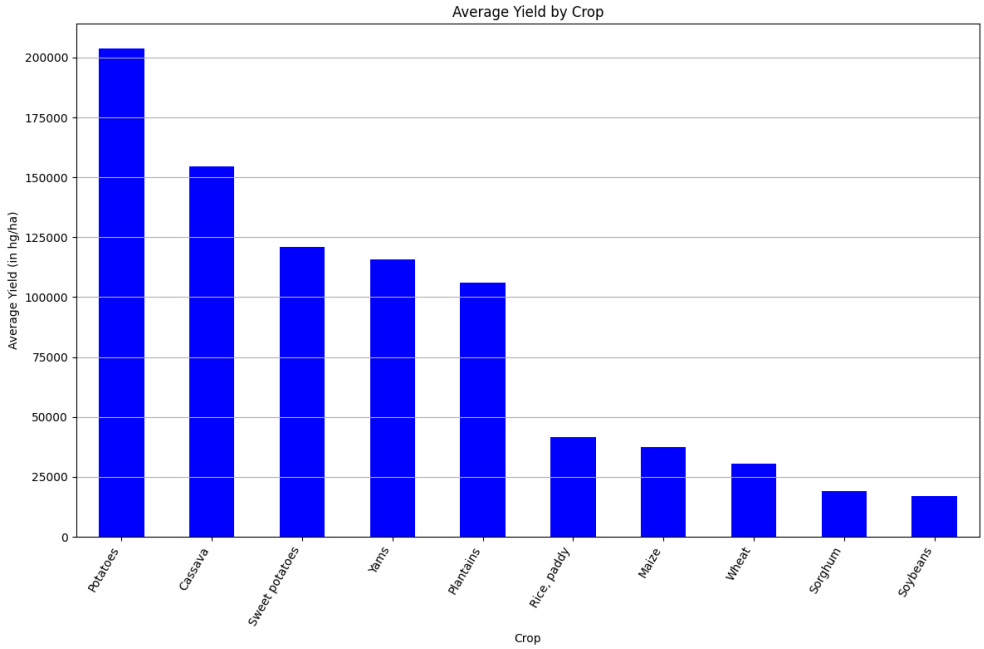

Figure 5: Average Yield by Crop

Figure 5 illustrates the average yield of each crop type. The crops are arranged in a descending order of average yield from left to right. While potatoes rank as the highest-yielding crop, soybeans rank as the lowest-yielding. Plantains as well as rice, paddy hold the intermediate positions, but plantains significantly exceed rice, paddy in yield. This is also reflected in the heatmap below in Figure 6.

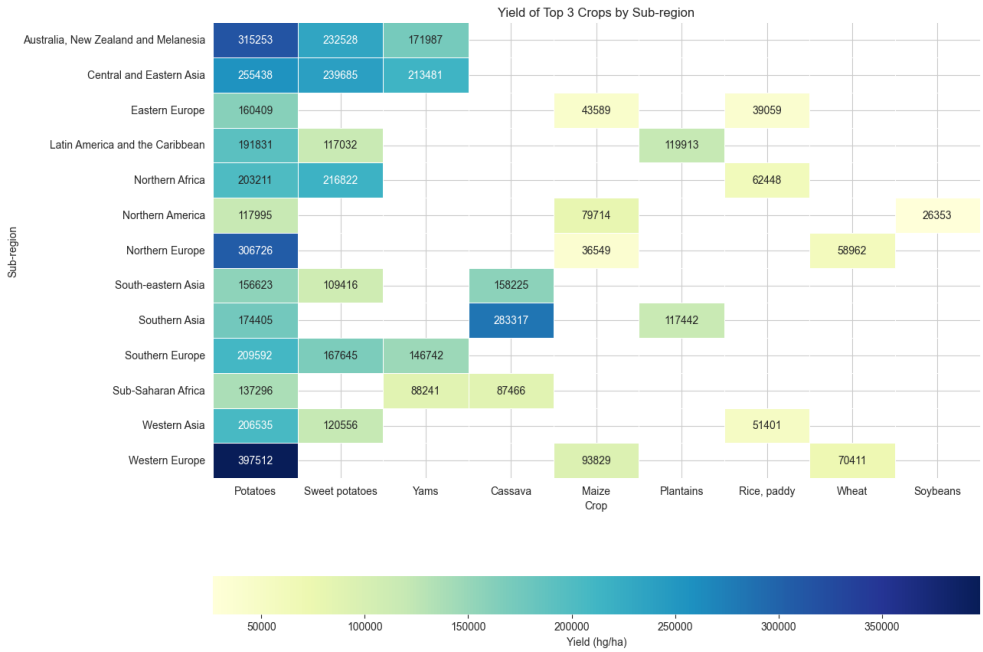

Figure 6: Yield of Top 3 Crops by Sub-region

This heatmap provides a condensed representation of the top three crops per sub-region. Based on their frequency as top three yielder, the crops are ordered from left to right in descending order. By means of the layout, the dominant crops are illustrated as well as the specific regions in which these crops prosper. In addition, the color gradients display the variations in crop yield across the different geographic areas. The heatmap on the one hand reveals that potatoes are consistently ranked in the top three crops of every sub-region in the dataset. On the other hand, soybeans only appear in the top three crops for Northern America. Sorghum does not rank among the top three crops in any of the sub-regions and is therefore not represented in Figure 6.

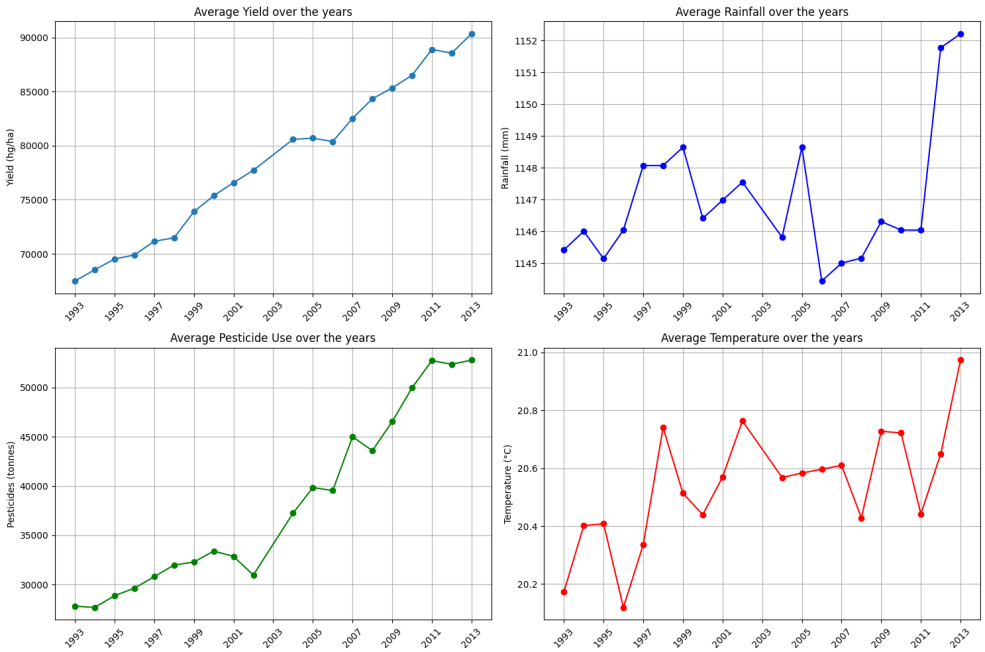

Figure 7: Average Values of Numeric Features over the Years

The four plots in Figure 7 present a composite view of agricultural metrics from 1993-2013. More specifically, they display the courses of the numeric variables, namely the average `Yield in hg/ha`, the average `Pesticides in tonnes`, the `Average rain in mm`, and the `Average temperature`.

The average `Yield in hg/ha` is almost consistently rising, only experiencing small dips in 2006 and 2012. Overall, during the span of 20 years, the yield of the ten crops in focus grows from 67.480 hg/ha in 1993 to 90.357 hg/ha in 2013, marking an increase by approx. 34%. The average `Pesticides in tonnes` use over the years displays a similar trend. It also increases consistently with only slight fluctuations, especially in 2002. While the average `Pesticides in tonnes` is at 27.289 tonnes in 1993, it is almost doubling in the course of 20 years, reaching 52.790 tonnes in 2013, thereby marking an increase of 90%. The `Average rain in mm` appears to be relatively consistent over the years with only a few fluctuations evident in the data. While some fluctuations, such as the spike in 2012, may appear to be substantial at first glance, a deeper inspection reveals that even this apparent change only accounts for an increase of around 5 mm. This variance corresponds to a fluctuation of 0.04%. Nevertheless, it is important to note that in agriculture, even a rather modest fluctuation in rain can result in a substantial impact on yield. A similar progression is seen in the plot for the `Average temperature`

over the years. Over the course of two decades, the curve presents some fluctuations, but shows an overall upward tendency in temperature. From 1993 to 2013, the `Average temperature` across all sub-regions rises by 0.8 Celsius (from 20.17 to 20.97). In the broader context of global warming and the climate crisis, even such seemingly subtle temperature increases are critical since rising temperatures have major effects on food security [25].

The four graphs in Figure 7 provide a high-level view of agricultural metrics over the two decades in focus. Remarkably, the pesticide curve displays a similar progression as the yield curve. This resemblance suggests a possible relationship between these two variables. In order to quantify this connection, a brief correlation analysis with the numeric variables in the dataset is conducted. Correlation is a widespread statistical measure describing the degree to which two variables change together. When a variable tends to increase whenever the respective other variable increases, there is a positive correlation between these two. In much the same vein, if a variable tends to decrease when another decreases, there is a negative correlation [52]. A measure that quantifies the linear relationship between two numeric variables is the Pearson correlation coefficient [18]. Ranging between -1 and 1, coefficients close to -1 indicate a strong negative linear relationship, and coefficients close to 1 indicate a strong positive linear relationship, while coefficients close to 0 imply little to no linear association. The Pearson correlation coefficient is widespread in the scientific sphere. However, it is important to mention that the Pearsons coefficient assumes a constant relationship of the variables across the whole data range and is rather sensitive to outliers [52]. This correlation analysis is briefly applied here, seeking to understand potential interactions between environmental factors and agricultural practices.

As seen in Table 5, the strongest relationship of variables is detected between `Pesticides in tonnes` and `Yield in hg/ha` with a value of r=0.965. The effect is positive linear and notably strong, suggesting that as the use of pesticides increased over the years, the yield also substantially increased. This potentially reflects the role of pesticides in crop protection. Both `Average rain in mm` and `Average temperature` also correlate positively with yield, although, with values of r=0.324 and r=0.640, significantly less then `Pesticides in tonnes`. This behaviour might reflect the inherent role which these two factors play in the growing conditions for certain crops. Yet, despite the positive correlations, it is important to note that correlation does not imply causation [52]. There could be other, uninvestigated factors that play a significant role or the relationships could possibly be coincidental.

No.	Variables	Correlation Coefficient r	Effect
1	`Pesticides in tonnes` and `Yield in hg/ha`	0.965	Strong positive relationship
2	`Average rain in mm` and `Yield in hg/ha`	0.324	Moderate positive relationship
3	`Average temperature` and `Yield in hg/ha`	0.640	Moderate positive relationship

Table 5: Pearson Correlation Coefficient of Numeric Variables

4.4 Setup of Baseline ML Models

Prior to diving into the complexities of bias detection and mitigation, the establishment of a baseline is vital. Training a baseline algorithm on the data establishes a reference for subsequent analyses and comparisons [9].

The ML algorithm A Random Forest regressor algorithm is used for this analysis. It is an ensemble learning method that constructs a large number of decision trees at training time and produces the mean prediction of the individual trees for regression problems [56]. In the literature, Random Forest algorithms are referred to as powerful and versatile method for crop yield prediction due to its accuracy and precision on the one hand and ease of use and utility on the other hand. Therefore, it is found highly capable of predicting crop yields and outperforms other models such as multiple linear regression benchmarks [40, 68]. The algorithm is set up as follows.

Libraries and Packages Several Python libraries and packages are employed to facilitate data processing, modeling, and evaluation. The Pandas library provides data structures and operations for efficient manipulation of numeric tables and time series. Here, it is extensively used for the purpose of data manipulation and analysis. Furthermore, the Scikit-learn library[18] is used as it offers a wide range of tools, making it a popular ML library in Python [56]. Here, it is employed for various tasks. For data pre-processing, its `train_test_split` [19] package is used to split the dataset into training and test subsets. For modeling, the sklearn library's `RandomForestRegressor` [20]

[18]https://scikit-learn.org/stable/

[19]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[20]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

is employed as the baseline ML algorithm for this analysis. For evaluation, metrics like `root_mean_squared_error` [21] and `r2_score` [22] are used to evaluate the performance of the model. These metrics provide a comprehensive assessment of the models and are explained in detail in Chapter 5. In addition, the Numpy library[23] is used for the numerical operations it provides to Python. This includes support for large multi-dimensional arrays and matrices, as well as a collection of mathematical functions to operate on these arrays [56]. Here, it is used to compute the square root of the MSE.

Model Training To initialize the Random Forest regressor, the default parameters are used, which encompass the number of trees in the forest (n_estimators=100), the function used to measure the quality of a split (criterion='mse' for regression), and the strategy to choose the split at each node (splitter='best'). Based on empirical evidence, these parameters help in generating well-performing general purpose models with low chances of overfitting [56].

Use Cases For the purpose of this research, two different baseline models are developed in order to address two use cases concerning bias in crop yield predictions. Thereby, the potential bias in predictions can be viewed at from two different perspectives, as seen in Figure 8 below. Understanding how crop yield predictions vary across different sub-regions is the basis of the first use case in this thesis, which is referred to as sub-region use case. By isolating the information on sub-regions and assessing its impact on the predictions, it can be investigated if certain sub-regions are consistently receiving higher or lower predictions, which may be indicative of a regional bias in the model. The objective of the second use case, the crop type use case, is to examine whether the model demonstrates preferential treatment to certain crops. This could highly influence agricultural strategies and indicate a bias in the model. By evaluating these two use cases individually, a comprehensive picture of potential biases in the predictive algorithms can be obtained.

[21] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

[22] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[23] https://numpy.org/

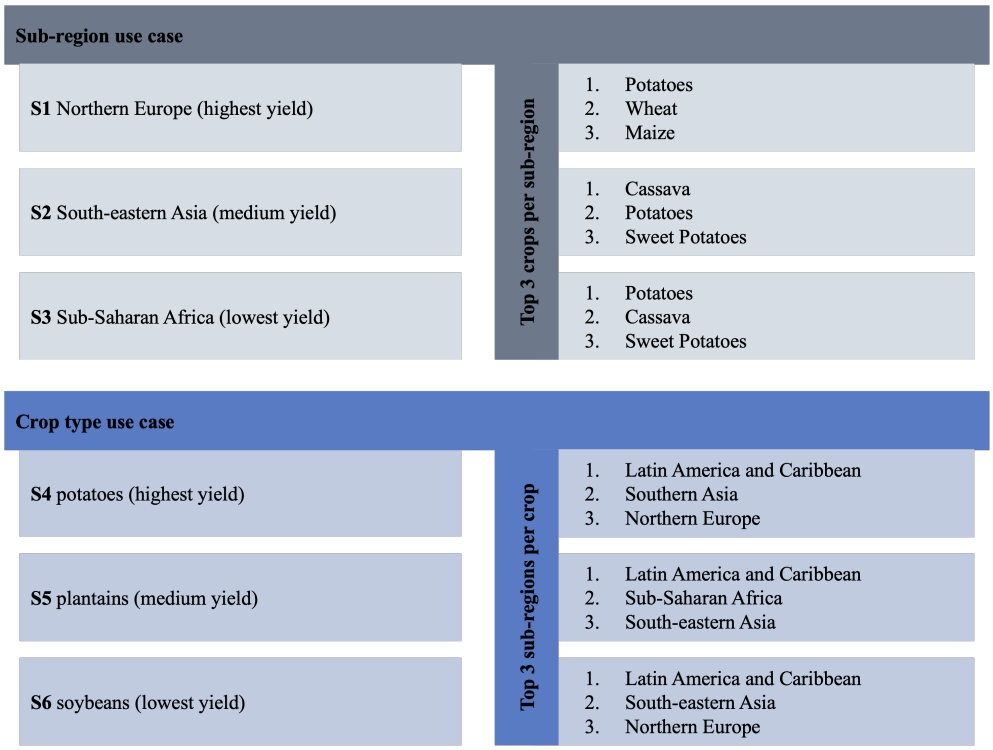

Figure 8: Overview of Use Cases and Scenarios

As the dataset is assembled, it catalogues crop yield delineated by the geographic variable `Sub-region`. While the dataset initially also included the variables `Country` and `Region`, these columns are streamlined. The rationale for this is that the `Sub-region` variable already captures the geographic distinction, thus allowing the analysis to be both informative and efficient without compromising the variety of geographic context.

Moreover, in order to conduct a more focused analysis and examine the differential impact of bias on different yield patterns, the data is strategically narrowed down to three distinctive Sub-regions: Northern Europe, South-eastern Asia, and Sub-Saharan Africa. Northern Europe contains the areas with the highest yields, positioning the Sub-regions as a benchmark for peak agricultural activity. In contrast, Sub-Saharan Africa features the lowest yields, thereby providing a perspective from the other end of the spectrum. South-eastern Asia represents the middle ground with its medium yields, bridging the two extremes.

Additionally, for a more granular exploration of these Sub-regions, their top three crops are determined by means of the yield as selection criterion. This approach is effective in isolating the key crops which are dominating the agricultural production in these areas.

The next phase of data preparation involves feature engineering. Preserving key variables such as `Yield in hg/ha` and `Year` in their original format is vital to maintaining the integrity of the dataset. Variables of categorical nature are one-hot encoded, meaning that a binary column is created for each category. Thereby, these variables are transformed into a format that is suitable for ML. Following feature engineering, the prepared dataset is split to ease model training and testing. A common split ratio is applied, with 80% of the data used for training and the remaining 20% reserved for testing purposes. This split not only facilitates model training, but also allows for reliable performance metrics during the evaluation.

In order to thoroughly explore the dataset, a dual-model approach is used. First, a Random Forest regressor is constructed and used for the SHAP summary plot in the next section of this thesis project, aiming at unraveling the nuances of each features influence on the prediction. In parallel to the regression analysis, a Random Forest classification model is also built in order to ensure compatibility with a wide range of tools available in the AIF360 toolkit. Since many of these tools are optimized for binary outcomes, constructing a classifier allows for an effective exploration of the AIF360 fairness metrics.

The initial phase of the analysis focuses on the baseline Random Forest regressor models which are used for the SHAP analysis, while the additional models are explored in Chapter 5. This allows for a layered exploration of the data and its inherent dynamics. These future scenarios will inject biases into the dataset, which will then be analysed by means of the AIF360 library to detect and potentially mitigate these injected biases.

4.5 Application of SHAP

Following the training of the models, SHAP is employed to interpret the baseline models predictions. Along with the growth of ML applications, the field of explainable Artificial Intelligence (xAI) has gained momentum in recent years. As ML models grow in complexity, the need to understand and interpret decisions and predictions made by these models grows as well [77]. The SHAP summary plots in Figure 9 and Figure 10 provide insights into the relevance of different features in the baseline model. Each point on the plot corresponds to a prediction for an observation. The y-axis represents the features, which are ordered from top to bottom based on their importance. The x-axis represents the SHAP values, reflecting the impact of that feature value on the prediction. A positive SHAP value implies that the feature pushes the model's prediction higher, whereas a negative SHAP value implies that the feature pushes the prediction lower [58]. The color indicates the feature value, red representing high values and blue representing low values.

Prior to delving into the interpretation of the SHAP plots, it is important to note that certain features in the plot incorporate values within their variable names, such as `Crop_Potatoes` or `Sub-region_Northern Europe`. This naming convention results from one-hot encoding that was carried out earlier to prepare the data for model training. One-hot encoding converts categorical variables into binary vectors, allowing ML models to handle categorical data without falsely interpreting those as ordinal [74].

In the sub-region use case, the SHAP summary plot in Figure 9 reveals that the following variables seem to have a high influence on the prediction.

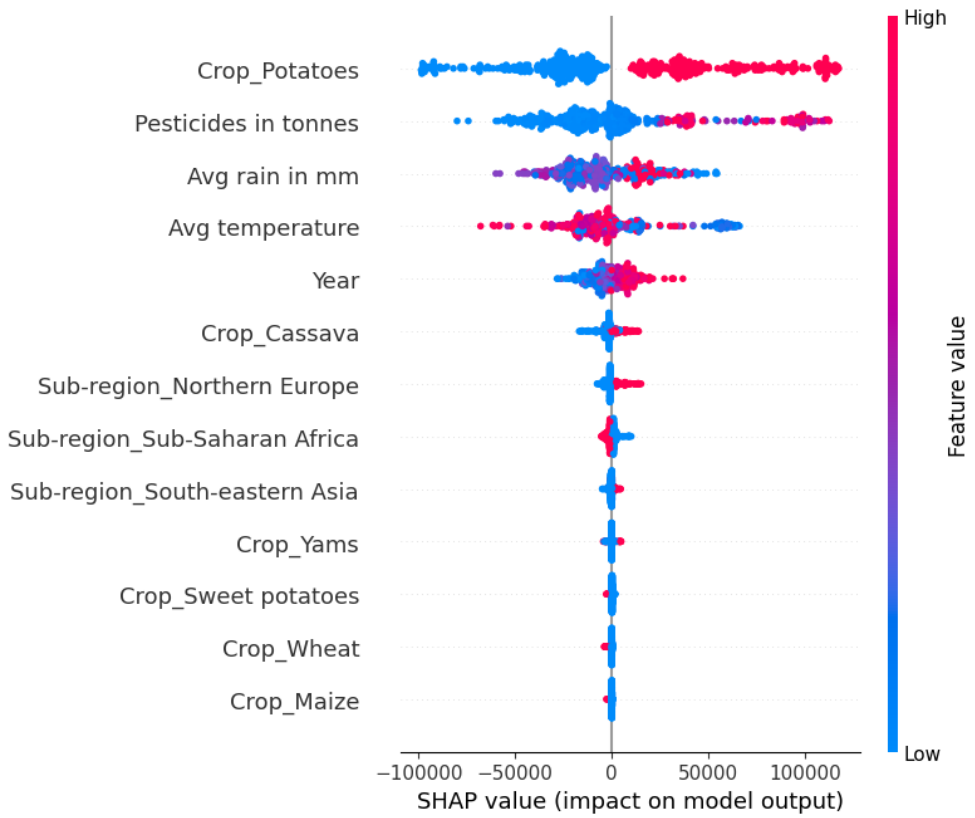

Figure 9: SHAP Summary Plot for Sub-region Use Case

`Crop_Potatoes` has a significant influence on the model's predictions. There is a trend towards higher yield predictions when the dataset includes potatoes as crop (highlighted in red). Similarly, the absence of potatoes as crop often leans towards lower yield predictions (highlighted in blue).

`Pesticides in tonnes` also displays an important influence. The red points leaning to the right indicate that increased pesticide use is associated with higher yield predictions, while the blue points, especially those leaning towards the left, imply a lower yield in case of reduced pesticide use.

`Avg rain in mm` indicates that higher rainfall tends to lead to higher yield (red points) while lower rainfall tends to lead to lower yield (blue points), highlighting the importance of water in crop growth.

The remaining features such as `Avg_temperature`, `Year`, the various crops and sub-regions suggest a mixed influence on the models predictions. These different effects on yield predictions indicate that several factors may interact with their effects.

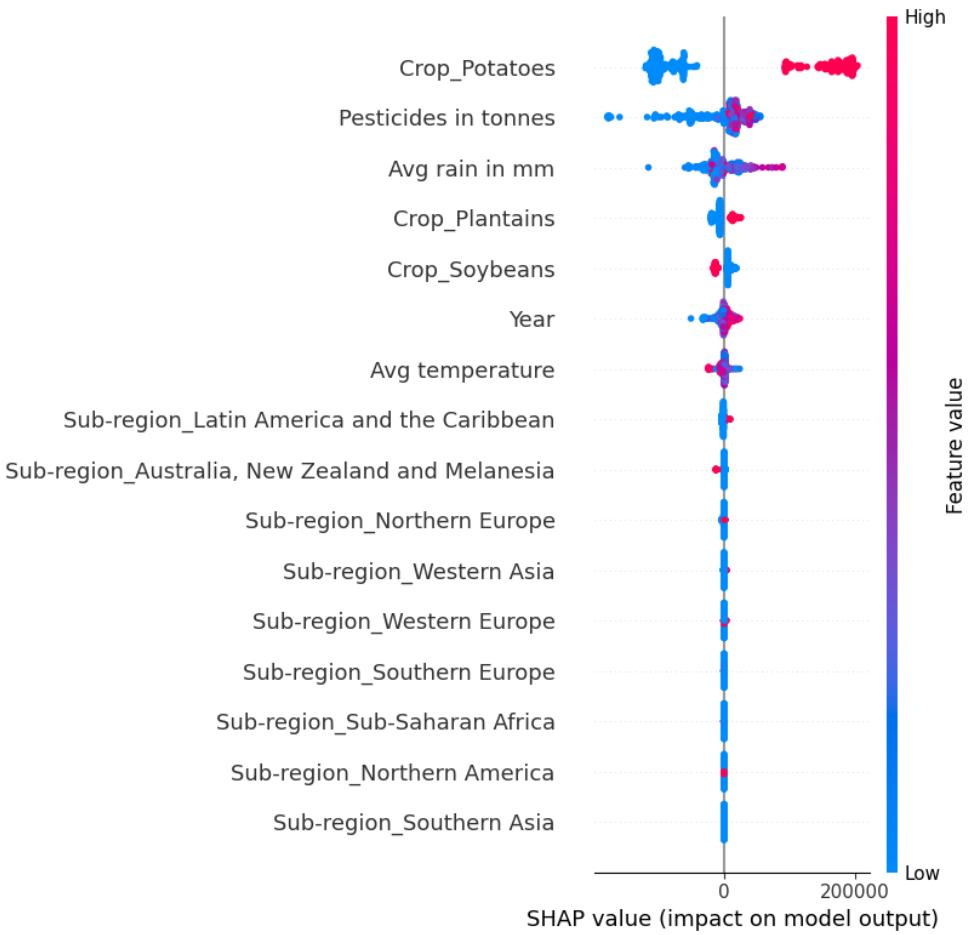

Figure 10: SHAP Summary Plot for Crop Type Use Case

In the crop type use case, the SHAP summary plot in Figure 10 reveals that the same variables seem to have a high influence on the prediction as in the sub-region use case, even though at another magnitude.

`Crop_Potatoes` has a significant influence on the model's predictions here as well. The presence of potatoes as crop type predicts higher yields, just as the absence of potatoes as crop type tends to predict lower yields.

`Pesticides in tonnes` displays a similar influence in the crop type use case as in the sub-region use case. Higher pesticide usage is positively associated with higher yields, while lower pesticide usage can lead to lower yields.

`Avg rain in mm` reflects the trend observed in the sub-region use case as well. Here, increased rainfall is also linked to higher yield predictions and vice versa.

The remaining features including `Crop_Plantains`, `Crop_Soybeans`, `Year`, `Avg temperature` and various sub-regions display a mixed influence on the models predictions, underscoring the interplay of several factors determining yield outcomes.

All in all, the SHAP plot reveals that in both use cases `Crop_Potatoes`, `Pesticides in tonnes`, and `Avg rain in mm` are the most influential features in predicting crop yield. Potatoes generally lead to higher yield predictions, and pesticide usage and rainfall positively influence yield.

5 Application of ML Models

This chapter focuses on firstly training algorithms on data that is intentionally biased [86]. Secondly, the nuances of the introduced biases are explored in depth across all six scenarios. Thirdly, by means of the AIF360 toolkit, this chapter then detects the manifested biases, and examines possible AIF360 mitigation techniques, shedding light on the intricacies of bias in ML models.

5.1 Set up of Biased ML Models

As the basis for this thesis project and the application of the AIF 360 toolkit, a notebook is used which is published by an AIF360 contributor, Oren Zeev-Ben-Mordehai, on his Github account[24] [6]. In Zeev-Ben-Mordehai's project, gender bias is artificially induced into a cardiology dataset. This is done by deliberately removing 40% of women's cardiac events, which leads to an (artificial) gender imbalance in the model. Subsequently, a model is trained on the biased data in which women are predicted to have a lower likelihood of being diagnosed with cardiovascular diseases. By means of different AIF360 tools, the researchers aim at detecting the artificially introduced bias.

In accordance with that, this thesis project follows a similar endeavor of manually introducing bias into the model. After establishing a non-biased baseline model and identifying the influence of features by means of a SHAP analysis in Chapter 4, the following section is concerned with establishing biased ML models. Based on the results of the SHAP analysis, bias is deliberately introduced into the model by modifying its most influential features. More specifically, the values of `Pesticides in tonnes` and `Avg rain in mm` are both reduced by 90%. This manual intervention facilitates the examination of ramifications of bias within the model, which paves the way for a profound analysis.

In the following analysis, in which bias is introduced across all scenarios, the ML model assessment relies on the following metrics to evaluate the models' performances:

[24] https://github.com/zbenmo/detecting-and-mitigating-bias-in-machine-learning-models-using-shap-and-aif360/blob/main/Copy_of_milestone_3_instructions.ipynb

RSME (Root Mean Squared Error) This metric measures the standard deviation of the residuals and provides insight into the accuracy of the model. A lower value indicates greater accuracy.

R-squared ($\mathbf{R^2}$) This statistical measure represents the proportion of variance in the dependent variable which can be explained by the independent variables in the model. The closer the R^2 value is to 1, the better the model fits the data.

Accuracy This metric indicates the ratio of correctly predicted observations to the total number of observations.

Precision This metric represents the number of correct positive results divided by the number of all positive results.

Recall (sensitivity) This metric expresses the number of correct positive results divided by the number of positive results which should have been produced.

F1 Score This measure balances precision and recall. A higher F1 score is desirable.

Sub-region use case

Scenario 1 Northern Europe Northern Europe has the highest yield across all sub-regions in the dataset. Scenario 1 Northern Europe (S1 Northern Europe) deals with introducing bias into this particular sub-region by manipulating the values of pesticide use and rain. The following results depicted in Figure 11 are obtained after introducing the bias and retraining both ML models. When bias is introduced to Northern Europe, both the Random Forest regressor and the Random Forest classifier still show high levels of performance. In the model, the *RSME* stands at 17,621.79, indicating a deviation between the model's predictions and the actual values. Looking at this value in relative terms, the prediction error of the model in the presence of bias amounts to 13,30%. Even with the manipulated features, the model achieves an R^2 of 0.970, implying that it is able to explain 97% of the variance in crop yield. The Random Forest classifier achieves an *Accuracy* of 96,61%, while both, the "Low-Med. Yield" and "Med.-High Yield" classes exhibit a high *Precision*, *Recall* and *F1 Score*. This underscores the model's ability in differentiating among yield categories, even under the introduced biases. The rather subtle shifts in the metrics compared to the baseline model

Sub-region use case			
S1 Northern Europe			
RF Regressor	RSME absolute		17621.79
	RSME relative		13.30%
	R2		0.970
RF Classifier		Accuracy	96.61%
	Low - Med. Yield	Precision	0.98
		Recall	0.95
		F1 Score	0.97
	Med. - High Yield	Precision	0.96
		Recall	0.98
		F1 Score	0.97

Figure 11: Model Measures S1 Northern Europe

suggest that in terms of predictive quality, the models display a considerable degree of resilience to the introduced bias.

Scenario 2 South-eastern Asia Shifting the bias to Scenario 2 South-eastern Asia (S2 South-eastern Asia), a median yield region, the resilience of both the Random Forest regressor and the Random Forest classifier can still be observed. Looking at the metrics of the Random Forest regressor, the absolute *RSME* is at 17,677.15, suggesting a slight increase in the spread of residuals compared to Northern Europe. Similarly, the relative *RSME* slightly rises to 13.34%. The R^2 value for this scenario is at 0.969, thereby displaying a comparable ability to account for the variance as the model in the previous scenario. Regarding the Random Forest classifier, an *Accuracy* of 96.19% is achieved, which is again slightly lower compared to Northern Europe. *Precision*, *Recall* and *F1 Score* metrics for both classes remain fairly stable compared to the previous scenario, ranging between 0.95 and 0.97. When comparing S2 South-eastern Asia in Figure 12 below to S1 Northern Europe in Figure 11, there is a marginal decrease in each metric. Nevertheless, the models for both sub-regions exhibit a notable stability with only minor variations in performance.

Sub-region use case			
S2 South-eastern Asia			
RF Regressor	RSME absolute		17677.15
	RSME relative		13.34%
	R2		0.969
RF Classifier		Accuracy	96.19%
	Low - Med. Yield	Precision	0.97
		Recall	0.95
		F1 Score	0.96
	Med. - High Yield	Precision	0.95
		Recall	0.97
		F1 Score	0.96

Figure 12: Model Measures S2 South-eastern Asia

Scenario 3 Sub-Saharan Africa Introducing bias in Scenario 3 Sub-Saharan Africa (S3 Sub-Saharan Africa), the sub-region with the lowest yield across all sub-regions in the dataset, the model showcases robustness through the following metrics described in Figure 13 below. The Random Forest regressor demonstrates an absolute *RSME* of 18,214.44, showing a modest increase compared to scenarios S1 Northern Europe and S2 South-eastern Asia, implying slightly more error in the prediction. Naturally, this is also observed in the relative *RSME*, which marginally increases to 13.74%. In spite of this, the model still maintains a high level of predictability with an R^2 value of 0.968, illustrating that it still captures approximately 96.8% of the variance in the dependent variable. The Random Forest classifier displays similar robust values. With an *Accuracy* of 96,61%, it is just as accurate as the previous scenarios. With values between 0.96 and 0.97, both yield classes demonstrate strong *Precision*, *Recall*, and *F1 Scores*. This robust classification across both yield classes, despite introducing bias towards Sub-Saharan Africa, underscores the model's robustness and its capability to efficiently distinguish between different levels of yield. Summarized, across all three scenarios, bias in Northern Europe, South-eastern Asia, and Sub-Saharan Africa, the models exhibit a marginal decrease in performance with each successive scenario. Nevertheless, their overall resilience remains strong, highlighting their ability to navigate sub-regional biases while sustaining robust forecasting results.

Sub-region use case			
S3 Sub-Saharan Africa			
RF Regressor	RSME absolute		18214.44
	RSME relative		13.74%
	R2		0.968
RF Classifier		Accuracy	96.61%
	Low - Med. Yield	Precision	0.97
		Recall	0.96
		F1 Score	0.97
	Med. - High Yield	Precision	0.96
		Recall	0.97
		F1 Score	0.97

Figure 13: Model Measures S3 Sub-Saharan Africa

Crop type use case

Scenario 4 potatoes The next analysis shifts the focus from regional biases to crop-specific biases. This next use case examines how the introduction of biases based on the crop type affects the models' performances. Starting with Scenario 4 potatoes (S4 potatoes), which provides the highest yields compared to the other crops in the dataset, the metrics depicted in Figure 14 below are observed. Interestingly, the Random Forest regressor demonstrates a slight improvement compared to the baseline model. The absolute *RSME* value decreases from 18,585.88 (baseline) to 17.858,39, which is also reflected in the relative *RSME* which decreases from 10.81% (baseline) to 10.39%. Similarly, although minimal, the R^2 value increases from 0.986 to 0.987, implying that despite the introduced bias, the model is able to explain the variance in the dependent variable marginally better than the baseline model. In contrast to this, the Random Forest classifier displays a slight decrease in *Accuracy*, reducing from 96.65% (baseline) to 96.28%. For the "Low-Med. Yield" class, the *Precision* remains constant at 0.97, while the *Recall* value drops slightly from 0.97 to 0.95, resulting in an *F1 Score* of 0.96. The other class, "Med.-High Yield", shows *Precision* and *Recall* values of 0.95 and 0.98 respectively, resulting in an *F1 Score* of 0.96. This indicates a minor shift in the classifier's ability to correctly predict this particular class. Overall, the performances of the models remain reasonably stable, even after introducing bias. This underscores the resilience of the models in dealing with such biases.

Crop type use case			
S4 potatoes			
RF Regressor	RSME absolute		17858.39
	RSME relative		10.39%
	R2		0.987
RF Classifier		Accuracy	96.28%
	Low - Med. Yield	Precision	0.97
		Recall	0.95
		F1 Score	0.96
	Med. - High Yield	Precision	0.95
		Recall	0.98
		F1 Score	0.96

Figure 14: Model Measures S4 Potatoes

Scenario 5 plantains In the case of Scenario 5 plantains (S5 plantains), introducing bias in the crop which exhibits a medium yield in the dataset, the models present the following results. As seen in Figure 15 below, the Random Forest regressor's absolute *RSME* value is 18,591.447, which is slightly higher than the 17,858.39 observed for potatoes. Thereby, also the relative *RSME* increases from 10.39% for potatoes to 10.81% for plantains, indicating that the predictions for the latter are slightly further from the actual values than they are for potatoes. With 0.986, the R^2 value for plantains is slightly lower than for potatoes, yet very high as the model can effectively explain 98.6% of the variance in the data. The values for the classifier metrics exhibit a similar pattern. The accuracy drops from 96.28% for potatoes to 95.66% for plantains. For the "Low-Med. Yield" class, plantains demonstrate a *Precision* of 0.94, which is lower than the 0.97 for potatoes, although the *Recall* for plantains increase to 0.97 here, compared to the 0.95 observed for potatoes. For the other class, "Med.-High Yield", the *Precision* for plantains mirrors that of potatoes at 0.97. However, with 0.94, the *Recall* for plantains is lower compared to the high *Recall* of 0.98 for potatoes. The *F1 Score* remains consistent for each class in both scenarios. It can be deduced that, while the introduction of bias for plantains reveals a slight dip compared to potatoes in certain Random Forest classifier metrics, the Random Forest regressor metrics remain consistent. Such variations highlight the model's ability to respond to different crops with different yields.

Crop type use case			
S5 plantains			
RF Regressor	RSME absolute		18591.47
	RSME relative		10.81%
	R2		0.986
RF Classifier		Accuracy	95.66%
	Low - Med. Yield	Precision	0.94
		Recall	0.97
		F1 Score	0.96
	Med. - High Yield	Precision	0.97
		Recall	0.94
		F1 Score	0.96

Figure 15: Model Measures S5 Plantains

Scenario 6 soybeans Turning to Scenario 6 soybeans (S6 soybeans), which constitute the lowest yield in the dataset, the model continues to be stable as seen in Figure 16 below. The absolute *RSME* value of the Random Forest regressor is at 18,314.93, which is between the values observed for potatoes and plantains. In much the same vein, the relative *RSME* value is 10.65%, again indicating a slight improvement for plantains and a slight decrease for potatoes. Interestingly, the R^2 value remains stable at 0.987, reflecting the same performance of the model as when biased towards potatoes. In the analysis of the classifier metrics, the *Accuracy* of soybeans at 95.66% is slightly below that of potatoes at 96.28%, but equivalent to that of plantains. In the "Low-Med. Yield" class, soybean's *Accuracy* of 0.94 is slightly lower than potato's 0.97. Similarly, in the "Med.-High Yield" class, soybean is on the same level as plantain, but slightly behind potato's *Recall* of 0.98. The *F1 Score* of 0.96 remains consistent across all three scenarios, demonstrating the stable performance of the models despite the different crop biases.

Comparing all three scenarios, bias in potatoes, plantains, and soybeans, reveals that as the bias is introduced from the highest yielding crop to the lowest yielding crop, the model's regression metrics remain largely consistent, while the classifier shows nuanced variations. Soybeans, despite having the lowest yield, show metrics comparable to plantains, pointing to the model's adaptability and strength in sustaining prediction quality across different crops with distinct yields.

Crop type use case			
S6 soybeans			
RF Regressor	RSME absolute		18314.93
	RSME relative		10.65%
	R2		0.987
RF Classifier		Accuracy	95.66%
	Low - Med. Yield	Precision	0.94
		Recall	0.97
		F1 Score	0.96
	Med. - High Yield	Precision	0.97
		Recall	0.94
		F1 Score	0.96

Figure 16: Model Measures S6 Soybeans

5.2 Application of AIF360

As part of the previous section, bias is successfully introduced into the ML models. The next step is to assess the extent to which this introduced bias can be detected. This is done by means of the AIF360 library. Out of the three fairness toolkits investigated in Chapter 2, AIF360 is chosen as the primary tool due to the following reasons. Compared to Tensorflow Responsible AI, AIF360 appears to be more suitable in the context of this thesis project as it not only offers bias detection, but also bias mitigation approaches. Tensorflow Responsible AI primarily offers tools and libraries for fairness metrics rather than explicit bias mitigation tools [2]. Likewise, AIF360 is chosen over Fairlearn because, while Fairlearn offers both bias detection and mitigation capabilities, it primarily focuses on post-processing and reduction algorithms for bias mitigation and is therefore not equipped with the pre-processing strategies required for this thesis project.

In particular, the following AIF360 fairness metrics are leveraged in the context of this thesis project: *Disparate Impact*, *Equal Opportunity Difference*, *Average Odds Difference*, and *Statistical Parity Difference*. According to the literature studied in Chapter 2, these metrics collectively provide insight into how the models behave across different groups, highlighting potential disparities in false positives, true positives, and overall favorable outcomes, thus allowing for a holistic assessment of the models' fairness [6].

In terms of interpreting these metrics, *Disparate Impact* values near 1 point to fairness, while deviations point to bias. For *Equal Opportunity Difference*, a value of 0 indicates equal true positive rates across groups, with deviations suggesting imbalances. For *Average Odds Difference*, values near zero signal fairness, while non-zero values reveal bias in favor of either the unprotected or protected group. *Statistical Parity Difference* values should approximate 0, while values away from 0 suggest disparities in favorable outcomes [63]. Beyond its capabilities in bias detection, the AIF360 toolkit includes a set of nine bias mitigation algorithms targeted at correcting biases in ML models. These algorithms are used based on the location where they can intervene within the ML pipeline. Pre-processing algorithms can modify the training data [6]. As the biases in the context of this thesis were introduced by strongly modifying data, two of the pre-processing algorithms from AIF360 are used for bias mitigation purposes:

Reweighing rebalances the weights of the training examples to mitigate bias in the decision boundaries, fostering a fairer classifier [44].

Disparate Impact Remover adjusts the features of the data set to reduce the correlation between sensitive attributes and the decision process, thereby ensuring that decisions remain unbiased by these features [6].

Starting with the sub-region use case, this section delves into detecting and mitigating biases which exist within the sub-regions. Afterwards, the same tools are applied to the crop type use case.

Sub-region use case

S1 Northern Europe Looking at the fairness metrics for in Figure 17 below, where bias is introduced for Northern Europe, the following values are produced.

Bias Detection The *Disparate Impact* metric is captured at 0.9636. Since this metric is close to the ideal value of 1, it suggests that there is a modest, but not profound difference between the favorable outcomes for the protected and non-protected groups. The *Equal Opportunity Difference* shows a positive value of 0.0287. This suggests that the protected group has a slightly higher true positive rate than the non-protected group. Furthermore, the *Average Odds Difference* holds a minimal value of 0.0016. Such a low value indicates that the results between the two groups are almost equal. In contrast to this, the *Statistical Parity Difference* comes to -0.0192. This negative value implies a small disadvantage for the protected group in terms of positive classifications.

Sub-region use case		
S1 Northern Europe		
DETECTING BIAS	Accuracy	96.61%
	Disparate Impact	0.9636
	Equal Opportunity Difference	0.0287
	Average Odds Difference	0.0016
	Statistical Parity Difference	-0.0192
MITIGATING BIAS	***Reweighing***	
	Accuracy (after Reweighing)	96.61%
	Disparate Impact (after Reweighing)	0.9746
	Equal Opportunity Difference (after Reweighing)	0.0345
	Average Odds Difference (after Reweighing)	0.0076
	Disparate Impact Remover	
	Accuracy (after DIR)	67.58%
	Disparate Impact (after DIR)	0.5551
	Equal Opportunity Difference (after DIR)	0.0460
	Average Odds Difference (after DIR)	-0.3983

Figure 17: Bias Metrics S1 Northern Europe

Bias Mitigation Applying the *Reweighing* mitigation algorithm, the model maintains its *Accuracy* of 96.61%. The *Disparate Impact* shows slight increase to 0.9746, suggesting that the model gradually moves towards more equitable results after *Reweighing*. Both, *Equal Opportunity Difference* and *Average Odds Difference* also show modest increases, indicating further refinement of the classification results. However, the application of the *Disparate Impact Remover* results in substantial changes. The accuracy drops to 67.58% and the *Disparate Impact* scores 0.5551, marking a significant deviation from the fair treatment mark. The new *Equal Opportunity Difference* and the *Average Odds Difference* underscore the profound negative impact of this mitigation strategy.

Sub-region use case		
S2 South-eastern Asia		
DETECTING BIAS	Accuracy	96.19%
	Disparate Impact	1.6400
	Equal Opportunity Difference	0.0230
	Average Odds Difference	-0.0149
	Statistical Parity Difference	0.2917
MITIGATING BIAS	***Reweighing***	
	Accuracy (after Reweighing)	96.61%
	Disparate Impact (after Reweighing)	1.6720
	Equal Opportunity Difference (after Reweighing)	0.0364
	Average Odds Difference (after Reweighing)	-0.0059
	Disparate Impact Remover	
	Accuracy (after DIR)	85.59%
	Disparate Impact (after DIR)	1.7704
	Equal Opportunity Difference (after DIR)	0.1810
	Average Odds Difference (after DIR)	0.0566

Figure 18: Bias Metrics S2 South-eastern Asia

S2 South-eastern Asia Determining the fairness metrics for S2 Southeast Asia reveals the values seen in Figure 18.

Bias Detection Turning to the fairness metrics, the *Disparate Impact* is particularly striking at 1.6400. This is substantially higher than the ideal benchmark of the value 1, and contrasts sharply with the outcome of S1 Northern Europe. It implies that the protected group in South-eastern Asia receives significantly more favorable outcomes than the unprotected group. The *Equal Opportunity Difference* is recorded at 0.0230. This is a small positive value and seems to indicate that the protected group has a slightly higher true positive rate. The value is close to that observed in S1 Northern Europe, potentially suggesting a common trend in the way bias affects true positive rates in different regions. The *Average Odds Difference* is at -0.0149, pointing to a minimal negative bias when both, false and true positive rate disparities, are combined. The *Statistical Parity Difference* shows a value of 0.2917. This value, which is notably positive, marks a bias towards positive

classifications for the protected group, a trend that is more accentuated than the one observed for S1 Northern Europe.

Bias Mitigation In mitigation, the use of the *Reweighing* algorithm results in a minor increase in *Accuracy* to 96.61%. The *Disparate Impact*, already higher than ideal, increases slightly more to 1.6720. The new *Equal Opportunity Difference* and *Average Odds Difference* also undergo adjustments that bring the model results closer to fairness. However, applying the *Disparate Impact Remover* has a significant impact, just as seen in S1 Northern Europe. While the drop in *Accuracy* is less pronounced, settling at 85.59%, the *Disparate Impact* jumps to 1.7704, indicating a further increase in the biased results. The remaining metrics, the new *Equal Opportunity Difference* and the *Average Odds Difference*, show significant shifts, reflecting the transformative potential of this particular mitigation strategy.

S3 Sub-Saharan Africa Continuing the exploration of bias within a sub-regional context, S3 Sub-Saharan Africa leads to the results depicted in Figure 19 below.

Bias Detection Taking into account the fairness metrics, the *Disparate Impact* stands out at 0.7003. This is below the optimal metric of 1, signaling that the protected group in Sub-Saharan Africa receives less favorable outcomes relative to the unprotected group. This contrasts with the elevated *Disparate Impact* observed for S2 South-eastern Asia and the near-neutral value for S1 Northern Europe, indicating region-specific disparities. The *Equal Opportunity Difference* scores -0.0434. This negative value means that the protected group faces a marginally reduced true positive rate. This pattern, although in the same negative direction, is more pronounced than the slight positive bias observed for S1 Northern Europe and S2 South-eastern Asia. The *Average Odds Difference* is barely negative at -0.0040. This indicates a minimal negative bias, a result that lies between the results of the other two sub-regions. The *Statistical Parity Difference* at -0.1827 marks a tendency against positive classifications for the protected group in Sub-Saharan Africa, reinforcing the prevailing trend of reduced positive outcomes for this group.

Sub-region use case		
S3 Sub-Saharan Africa		
DETECTING BIAS	Accuracy	96.61%
	Disparate Impact	0.7003
	Equal Opportunity Difference	-0.0434
	Average Odds Difference	-0.0040
	Statistical Parity Difference	-0.1827
MITIGATING BIAS	***Reweighing***	
	Accuracy (after Reweighing)	96.40%
	Disparate Impact (after Reweighing)	0.6934
	Equal Opportunity Difference (after Reweighing)	-0.0535
	Average Odds Difference (after Reweighing)	-0.0091
	Disparate Impact Remover	
	Accuracy (after DIR)	77.97%
	Disparate Impact (after DIR)	0.5295
	Equal Opportunity Difference (after DIR)	-0.1437
	Average Odds Difference (after DIR)	-0.3048

Figure 19: Bias Metrics S3 Sub-Saharan Africa

Bias Mitigation Venturing into the bias mitigation strategies, applying the *Reweighing* algorithm leads to a minimal dip in accuracy, which is now 96.40%. The *Disparate Impact*, already lower than the ideal benchmark, drops further to 0.6934. The reweighed *Equal Opportunity Difference* and *Average Odds Difference* both move in directions that amplify the detected bias, which indicates that *Reweighing* may not be the optimal solution for bias in this particular context. When applying the *Disparate Impact Remover*, there is a noticeable drop in accuracy to 77.97%. The *Disparate Impact* decreases substantially to 0.5295, exacerbating the disparity in favorable outcomes. Likewise, the *Equal Opportunity Difference* and the *Average Odds Difference* indicate exacerbated bias, underscoring the significant transformative and, in this case, detrimental impact of this mitigation technique for Sub-Saharan Africa.

All in all, for bias detection and mitigation in the sub-region use case, in which bias was introduced at the sub-regional level, there are distinct differences in fairness metrics across S1 Northern Europe, S2 South-eastern Asia and S3 Sub-Saharan Africa. While all three sub-regions maintain high accuracy, the effectiveness of mitigation strategies varies, with Northern Europe demonstrating rather minor biases, Southeast Asia showing favorable results, and Sub-Saharan Africa experiencing consistent challenges in overcoming its introduced biases.

Crop type use case

S4 potatoes This crop type use case illustrates the impact of introduced bias towards the crop type potatoes in Figure 20 below.

Bias Detection The *Disparate Impact* for potatoes is at 1.0734, indicating a minor bias toward potatoes compared to other crops. This signifies that potatoes have slightly more favorable outcomes than the baseline, but not by any significant amount. The positive *Equal Opportunity Difference* of 0.0294 implies that potatoes have a minor better true positive rate compared to the other crops. Similarly, the *Average Odds Difference* of 0.0124 supports this notion by underlining a favorable rate of correct predictions for potatoes.

Bias Mitigation After applying the *Reweighing* algorithm, the *Disparate Impact* rises slightly to 1.0848, indicating that the mitigation approach surprisingly makes the model slightly more biased in favor of potatoes. The *Equal Opportunity Difference* also increases to 0.0348, indicating an increase in the true positive rate difference between potatoes and other crops. Applying the *Disparate Impact Remover*, there's a noticeable drop in the *Disparate Impact*, which decreases to 0.7548, signifying a less favorable outcome for potatoes. The *Equal Opportunity Difference* rises to 0.2000 and the *Average Odds Difference* becomes negative to -0.1996. This demonstrates that potatoes experienced a significant decrease in true positives and an increase in false positives after the application of the *Disparate Impact Remover*.

While the bias towards potatoes produces subtle differences in prediction outcomes, the mitigation efforts, particularly with the *Disparate Impact Remover*, produce more pronounced effects, moving the balance away from the initially introduced bias.

Crop type use case		
S4 potatoes		
DETECTING BIAS	Accuracy	96.28%
	Disparate Impact	1.0734
	Equal Opportunity Difference	0.0294
	Average Odds Difference	0.0124
	Statistical Parity Difference	0.0374
MITIGATING BIAS	***Reweighing***	
	Accuracy (after Reweighing)	96.28%
	Disparate Impact (after Reweighing)	1.0848
	Equal Opportunity Difference (after Reweighing)	0.0348
	Average Odds Difference (after Reweighing)	0.0178
	Disparate Impact Remover	
	Accuracy (after DIR)	63.21%
	Disparate Impact (after DIR)	0.7548
	Equal Opportunity Difference (after DIR)	0.2000
	Average Odds Difference (after DIR)	-0.1996

Figure 20: Bias Metrics S4 Potatoes

S5 plantains Obtaining the fairness metrics for S5 plantains, where bias is introduced for a medium yield crop, reveals the values shown in Figure 21 below.

Bias Detection The *Disparate Impact*, measured at 0.6016, points to a lower favorability for plantains compared to other crops. This value implies that the results for plantains are about 60% as favorable as those for other crops, indicating the presence of bias against the crop plantains. The negative *Equal Opportunity Difference* of -0.0374 indicates that plantains are slightly disadvantaged when it comes to true positive rates compared to other crops. This finding is further reflected in the *Average Odds Difference* of -0.0178, which is indicative of a slightly unfavorable rate of both true and false positive predictions for plantains. Also, the *Statistical Parity Difference* of -0.2274 adds further evidence to this bias, signaling that plantains are less prone to favorable outcomes.

Crop type use case		
S5 plantains		
DETECTING BIAS	Accuracy	95.66%
	Disparate Impact	0.6016
	Equal Opportunity Difference	-0.0374
	Average Odds Difference	-0.0178
	Statistical Parity Difference	-0.2274
MITIGATING BIAS	***Reweighing***	
	Accuracy (after Reweighing)	95.95%
	Disparate Impact (after Reweighing)	0.6059
	Equal Opportunity Difference (after Reweighing)	-0.0374
	Average Odds Difference (after Reweighing)	-0.0129
	Disparate Impact Remover	
	Accuracy (after DIR)	74.57%
	Disparate Impact (after DIR)	0.5676
	Equal Opportunity Difference (after DIR)	-0.0532
	Average Odds Difference (after DIR)	-0.2045

Figure 21: Bias Metrics S5 Plantains

Bias Mitigation After the application of the *Reweighing* algorithm, the *Accuracy* of the model slightly improves to 95.95%. The new *Disparate Impact* also increases, albeit minimally, to 0.6059, indicating a slight bias reduction for plantains. Yet, the *Equal Opportunity Difference* remains unchanged, revealing that the true positive rate difference between plantains and other crops persists. The *Average Odds Difference* improves marginally to -0.0129, reflecting a minor correction in the balance between true and false positive rates. After applying the *Disparate Impact Remover*, the *Disparate Impact* value further decreases to 0.5676, exacerbating the existing bias against plantains. The *Equal Opportunity Difference* further deteriorates to -0.0532, and the *Average Odds Difference* turns more negative to -0.2045. These post *Disparate Impact Remover* metrics hint at the fact that the mitigation strategy may have accentuated the bias against plantains inadvertently.

Overall, in S5 plantains, a bias against plantains is noticeable. The implemented strategies, particularly the *Disparate Impact Remover*, appear to further exacerbate the imbalance, highlighting the complexity of addressing biases for specific crop types.

S6 soybeans Continuing the bias exploration in the crop type context, S6 soybeans examines the implementation of bias towards the lowest yielding crop in the dataset, generating the results presented in Figure 22.

Bias Detection The *Disparate Impact* metric amounts to 1.3793, signifying a bias in favor of soybeans over other crops. This implies that the results for soybeans are approximately 38% more favorable compared to other crops, revealing the presence of a positive bias in favor of soybeans. The positive *Equal Opportunity Difference* of 0.0147 is indicative of a minor advantage for soybeans in determining true positives over other crops. This is further demonstrated by the *Average Odds Difference* of 0.0228, which also gives a slightly favorable rate for both true and false positive predictions for soybeans. Additionally, the *Statistical Parity Difference* of 0.1725 supports a positive indication of this positive bias, suggesting that soybeans are more prone to favorable outcomes.

Crop type use case		
S6 soybeans		
DETECTING BIAS	Accuracy	95.66%
	Disparate Impact	1.3793
	Equal Opportunity Difference	0.0147
	Average Odds Difference	0.0228
	Statistical Parity Difference	0.1725
MITIGATING BIAS	***Reweighing***	
	Accuracy (after Reweighing)	95.95%
	Disparate Impact (after Reweighing)	1.3337
	Equal Opportunity Difference (after Reweighing)	0.0060
	Average Odds Difference (after Reweighing)	0.0010
	Disparate Impact Remover	
	Accuracy (after DIR)	83.24%
	Disparate Impact (after DIR)	1.0673
	Equal Opportunity Difference (after DIR)	0.0247
	Average Odds Difference (after DIR)	-0.0981

Figure 22: Bias Metrics S6 Soybeans

Bias Mitigation In regard to mitigation, when employing the *Reweighing* algorithm, the model's *Accuracy* increases slightly to 95.95%. However, the *Disparate Impact* decreases to 1.3337, representing a slight reduction in favorability of soybeans. The *Equal Opportunity Difference* drops to 0.0060, which is a reduction in the true positive rate advantage which soybeans initially held. The *Average Odds Difference* also shrinks dramatically to 0.0010, meaning that the favorable predictions for soybeans, both true and false positives, become nearly equal. The application of the *Disparate Impact Remover* remarkably reduces the disparity to 1.0673, pointing to a much more balanced result between soybeans and other crops. Interestingly, the new *Equal Opportunity Difference* climbs to 0.0247, inferring that soybeans obtained a small advantage in true positive rates after mitigation. Conversely, the *Average Odds Difference* produces a negative value of -0.0981, thus illustrating an unfavorable prediction balance for soybeans, particularly in terms of false positives.

Overall, S6 soybeans demonstrates a clear positive bias favoring soybeans in the detection phase. Although mitigation strategies, especially the *Disparate Impact Remover*, progress in leveling the field, they also create complexity in terms of prediction scores, underlining the challenges of managing bias in crop types.

In the crop type use case, the model identifies distinct biases for potatoes, plantains, and soybeans. Potatoes are slightly overpredicted, plantains are consistently underpredicted, while soybeans show an overprediction that can almost be offset through mitigation. Despite varying biases, the model accuracy remains similarly high across all crops, although the effectiveness of mitigation varies for each crop.

All in all, it can be deduced that in the sub-region use case, Northern Europe, South-eastern Asia, and Sub-Saharan Africa each demonstrate unique biases when manipulated, with the models indicating different levels of resilience across the regions. In the crop type use case, introducing bias for potatoes, plantains, and soybeans leads to different predictive outcomes, with mitigation strategies leading to different outcomes. In both use cases, while the model accuracy remains constant, the variation in results emphasizes the complexity of managing bias in different contexts. A comprehensive overview of all calculated metrics, which allow for comparison of the different values, is shown below in Figure 23 for the sub-region use case and in Figure 24 for the crop type use case.

Sub-region use case				
		S1 Northern Europe	**S2 South-eastern Asia**	**S3 Sub-Saharan Africa**
Baseline Random Forest Regressor Algorithm				
RSME absolute		**17621.79**	17677.155	18214.443
R2		**0.9697**	0.9695	0.9676
RSME relative		**13.30%**	13.34%	13.74%
Baseline Random Forest Classifier Algorithm				
Low - Medium Yield Class	Precision	**0.98**	0.97	0.97
	Recall	0.95	0.95	**0.96**
	F1 Score	**0.97**	0.96	**0.97**
Medium - High Yield Class	Precision	**0.96**	0.95	**0.96**
	Recall	**0.98**	0.97	0.97
	F1 Score	**0.97**	0.96	0.97
Bias Detection Metrics				
Accuracy		**96.61%**	96.19%	**96.61%**
Disparate Impact		0.9636	**1.6400**	0.7003
Statistical Parity Difference		-0.0192	**0.2917**	-0.1827
Equal Opportunity Difference		0.0287	0.0230	**-0.0434**
Average Odds Difference		0.0016	**-0.0149**	-0.0040
Bias Mitigation Strategy 1: Reweighing				
Accuracy after Reweighing		96.61%	96.61%	96.40%
Disparate Impact after Reweighing		**0.9746**	1.6720	0.6934
Equal Opportunity Difference after Reweighing		**0.0345**	0.0364	-0.0535
Average Odds Difference after Reweighing		0.0076	**-0.0059**	-0.0091
Bias Mitigation Strategy 2: Disparate Impact Remover				
Accuracy after DIR		67.58%	**85.59%**	77.97%
Disparate Impact after DIR		0.5551	**1.7704**	0.5295
Equal Opportunity Difference after DIR		**0.0460**	0.1810	-0.1437
Average Odds Difference after DIR		-0.3983	**0.0566**	-0.3048

Figure 23: Overview Results for Sub-region Use Case

Crop type use case				
		S4 potatoes	**S5 plantains**	**S6 soybeans**
Baseline Random Forest Regressor Algorithm				
RSME absolute		**17858.39**	18591.466	18314.928
R2		**0.9873**	0.9862	0.9866
RSME relative		**10.39%**	10.81%	10.65%
Baseline Random Forest Classifier Algorithm				
Low - Medium Yield Class	Precision	**0.97**	0.94	0.94
	Recall	0.95	**0.97**	**0.97**
	F1 Score	**0.96**	**0.96**	**0.96**
Medium - High Yield Class	Precision	0.95	**0.97**	**0.97**
	Recall	**0.98**	0.94	0.94
	F1 Score	**0.96**	**0.96**	**0.96**
Bias Detection Metrics				
Accuracy		**96.28%**	95.66%	95.66%
Disparate Impact		1.0734	**0.6016**	1.3793
Statistical Parity Difference		0.0374	**-0.2274**	0.1725
Equal Opportunity Difference		0.0294	**-0.0374**	0.0147
Average Odds Difference		0.0124	-0.0178	**0.0228**
Bias Mitigation Strategy 1: Reweighing				
Accuracy after Reweighing		**96.28%**	95.95%	95.95%
Disparate Impact after Reweighing		1.0848	0.6059	**1.3337**
Equal Opportunity Difference after Reweighing		0.0348	-0.0374	**0.0060**
Average Odds Difference after Reweighing		0.0178	-0.0129	**0.0010**
Bias Mitigation Strategy 2: Disparate Impact Remover				
Accuracy after DIR		63.21%	74.57%	**83.24%**
Disparate Impact after DIR		0.7548	0.5676	**1.0673**
Equal Opportunity Difference after DIR		**0.2000**	-0.0532	0.0247
Average Odds Difference after DIR		**-0.1996**	-0.2045	-0.0981

Figure 24: Overview Results for Crop Type Use Case

6 Discussion and Limitations

This chapter on the one hand aims at summarizing the results and thereby answering the research questions and on the other hand at stating potential limitations of the study. This thesis is concerned with the overarching Research Question: How can bias be detected in AI based yield predictions for food security? In the following, this question is answered by answering its three sub-questions.

6.1 The Applied Bias Detection Method

RQ1.1 What does the chosen bias detection method reveal about its impact on AI-based yield predictions for food security?

The evaluation starts with an analysis of the sub-region use case, focusing on model performance and bias detection metrics. The discussion then moves to the crop type use case, where the model performance and bias detection metrics are examined accordingly.

Sub-region use case

Model performance In regard to the sub-region use case, the performance of the Random Forest regressor shows a noticeable tendency: it performs optimally for S1 Northern Europe (highest yield), somewhat less so for S2 Southeast Asia (median yield) and deviates the most for S3 Sub-Saharan Africa (lowest yield). However, these differences remain minimal, underscoring the robustness of the model. At the same time, the Random Forest classifier remains consistent across all scenarios, demonstrating its resilience to regional biases.

Bias detection In the assessment of bias detection across the three subregions, the models show different degrees of sensitivity to the introduced biases. The accuracy metric points to the models maintaining consistent predictive capabilities across S1 Northern Europe, S2 South-eastern Asia, and S3 Sub-Saharan Africa, regardless of the specific regional biases. For *Disparate Impact*, Northern Europe's value is close equal outcomes between privileged and unprivileged groups. This could potentially be attributed to the fact that Northern Europe is privileged by virtue of its high yield. In contrast, South-eastern Asia's metric is notably indicating an advantage for this unprivileged group, suggesting that South-eastern Asia is experiencing

more favorable outcomes than its privileged counterpart, despite being a medium yield sub-region. Conversely, Sub-Saharan Africa presents a bias against it. This sub-region, which offers the lowest yield compared to all other sub-regions, seems to bear the brunt of the introduced bias, thereby suggesting less favorable outcomes compared to the privileged group. Almost the same results are achieved for *Statistical Parity Difference*. Northern Europe's value is again close to zero, thereby balanced between the privileged and unprivileged groups. In contrast, South-eastern Asia's high positive value demonstrates a bias in favor of the unprivileged group, while Sub-Saharan Africa's negative value suggest a bias towards the unprivileged group. The *Equal Opportunity Difference* and the *Average Odds Difference* across the scenarios display negligible disparities. Their minimal magnitudes suggest limited practical implications, yet they serve as complementary information in the broader context of bias detection.

In summary, while Northern Europe's bias remains largely undetected, possibly due to its privileged status, Sub-Saharan Africa clearly exhibits the introduced bias, consistent with its unprivileged position. South-eastern Asia offers a surprising deviation, tending toward unexpectedly favorable outcomes for its unprivileged group. Overall, the AIF360 toolkit proves useful in detecting bias in two of the three scenarios, underscoring its effectiveness and relevance in such analytical efforts.

Crop type use case

Model performance With respect to the crop type use case, all Random Forest regression metrics show the best performance for S4 potatoes, the crop with the highest yield in the dataset. Interestingly, S6 soybeans, in which bias is introduced against the crop with the lowest yield, ranks second in terms of performance. In contrast, S5 plantains, in which bias is introduced against the crop with medium yield, ranks lower than the other two, although the performance differences between the scenarios are comparatively small. When evaluating the Random Forest classifier, the performance across all three scenarios - S4 potatoes, S5 plantains, and S6 soybeans - is found to be fairly consistent.

Bias detection When evaluating bias detection in the crop types, the models again show different degrees of sensitivity to the introduced biases. Across all three scenarios, the accuracy is consistent, indicating that the general predictive capability of the model is comparable across all three crop

types. For S4 potato, which represents the crop with the highest yield in all Sub-regions, the model shows relatively fair treatment, as reflected by a *Disparate Impact* value near the ideal threshold of 1. This means that the model does not significantly favor or discriminate against potatoes compared to other crops. Conversely, S5 plantains, a medium yield crop, has a *Disparate Impact* value well below 1, indicating that this crop type is less likely to obtain a favorable outcome compared to potatoes. S6 soybeans, the crop with the lowest yield across all sub-regions, has a *Disparate Impact* value above 1, implying that despite having the lowest yield in the dataset, soybeans are more likely to receive favorable outcomes than potatoes. This rather counterintuitive result suggests that the model may overcompensate for the inherent low yield of soybeans. Again, almost the same results as for the *Disparate Impact* are mirrored in the *Statistical Parity Difference*. The potatoes' value suggests an almost equal rate of favorable outcomes for the privileged and unprivileged groups. Contrasting this, plantains negative value suggests a bias towards the unprivileged group, while soybeans' positive *Statistical Parity Difference* suggests that this crop may be receiving more favorable results than expected relative to their actual yield. The distinctions in these values reflect the challenges in achieving model fairness, especially when there are inherent disparities like differing crop yields. The values for both, *Equal Opportunity Difference* and the *Average Odds Difference* are again rather modest across all three scenarios in the crop type use case, suggesting that the differences in receiving favorable outcomes between the privileged and unprivileged groups is minimal. Thereby, these metrics fail to detect the introduced bias, falsely suggesting the models across all three crops are fair.

In essence, while model accuracy is consistent across crop types, the bias detection tools in the AIF360 toolkit highlight potential differences in predictions, particularly for the plantains and soybeans scenarios.

6.2 The Applied Bias Mitigation Method

RQ1.2 How effective are the selected bias mitigation techniques in reducing bias in AI-based food security yield prediction models?

The evaluation starts with the sub-region use case, investigating first the *Reweighing* algorithm and then the *Disparate Impact Remover*. A similar analysis is then done for the crop type use case with both algorithms.

Sub-region use case

Upon applying the *Reweighing* algorithm from the AIF360 toolkit to mitigate bias, the following can be observed. The accuracy metrics for all scenarios remain robust, suggesting that the *Reweighing* process does not compromise the general predictive capability of the model. In addition, the *Disparate Impact* for S1 Northern Europe marginally improves, showing that the mitigation process has been slightly effective in mitigating the bias for this region. Similarly, for S2 South-eastern Asia and S3 Sub-Saharan Africa, the post-mitigation results still reveal a significant bias. In fact, for S2 South-eastern Asia, the bias appears to minimally increase, while for S3 Sub-Saharan Africa, the bias decreases only slightly. This suggests that *Reweighing* may not always be effective as a bias mitigation strategy, especially when the initial bias is strong. After *Reweighing*, the values for both *Equal Opportunity* and *Average Odds Differences* remain small across all scenarios, indicating that the model's fairness in terms of positive predictions (both true and false) between the privileged and unprivileged groups remains consistent post-mitigation. While *Reweighing* appears to preserve model accuracy here, its effectiveness in mitigating bias varies across scenarios. For S1 Northern Europe, the mitigation process seems relatively successful, yet for S2 South-eastern Asia and S3 Sub-Saharan Africa, the results are mixed. The challenge to effectively address more pronounced biases remains, particularly in scenarios with high *Disparate Impacts*.

Likewise, the application of the *Disparate Impact Remover* from the AIF360 toolkit as a means of bias mitigation results in several notable observations. The model's accuracy decreases significantly across all scenarios after implementing the *Disparate Impact Remover*. This might imply that there has been a trade-off in attempting to remove bias, which negatively affects the overall predictive ability of the model. However, the post-mitigation values of the *Disparate Impact* also demonstrate concerning developments: the values for all three scenarios have shifted further away from the ideal value of 1. For S1 Northern Europe and S3 Sub-Saharan Africa, the *Disparate Impact* scores decrease further below 1. And while the metric for S2 South-eastern Asia

already clearly exceeded the value 1 before, it rises even further after mitigation. Strikingly, the values continue to shift in the undesirable direction, pointing to an increase in bias for all three scenarios after applying the *Disparate Impact Remover*. This illustrates that the *Disparate Impact Remover* may not be a universal panacea and may not necessarily achieve the desired fairness effects in all scenarios. Similarly, a decrease across all three scenarios is evident for the *Equal Opportunity* and *Average Odds Differences* after applying the *Disparate Impact Remover*. Particularly, S1 Northern Europe and S3 Sub-Saharan Africa record a larger bias.

In conclusion, the intent of using the *Disparate Impact Remover* was to promote fairness. Yet, the results indicate that it's a difficult endeavor. Usually, bias mitigation algorithms require a delicate balancing act between achieving fairness and maintain model effectiveness. However, in this case, both have deteriorated after applying the *Disparate Impact Remover*.

Crop type use case

Applying the *Reweighing* algorithm from the AIF360 toolkit to mitigate bias based on crop type results in the following observations. The accuracy metrics across all three scenarios stay robust, signifying that the *Reweighing* process does not significantly affect the models' overall predictive ability. After *Reweighing*, the *Disparate Impact* for S4 potatoes as well as S5 plantains deteriorates slightly, suggesting that the bias for these crops has not been effectively addressed. However, for S6 soybeans, there is a modest improvement. Overall, this observation suggests that while *Reweighing* can at times help mitigate bias, its effectiveness is not uniform across all crop types. After *Reweighing*, the values for both the *Equal Opportunity Difference* and the *Average Odds Difference* remain modest for all scenarios. In particular, these values are lowest for S6 soybeans, reflecting a small improvement in fairness for this crop. As the changes are minimal across all crop types, however, it could be argued that these metrics have not experienced significant shifts post-mitigation. In general, the *Reweighing* method offers mixed results for bias reduction depending on the crop type. Mitigation efforts are less productive for S1 potatoes and S2 plantains, but there is a slight bias reduction for S3 soybeans. The challenge of using *Reweighing* to effectively address biases in different crop yields remains evident.

Furthermore, the application of the *Disparate Impact Remover* from the AIF360 toolkit as a means of bias mitigation leads to a number of remarkable observations. Just like in the sub-region use case, the accuracy of the model predictions drops in all scenarios after applying the *Disparate Impact Remover*. S4 potatoes experience the most severe loss in accuracy, followed by S5 plantains and then S6 soybeans. This particular mitigation process might have altered the feature distributions in a way that affects the overall predictive performance of the model. The post-mitigation values of the *Disparate Impact* for S4 potatoes also deteriorate considerably, demonstrating that the bias for this crop increases. In comparison, S5 plantains shows some reduction in bias while S6 soybeans exhibits a significant improvement in bias mitigation as shown by the improved *Disparate Impact* score. Again, the varying results across crop types highlight the complexities in addressing biases using the *Disparate Impact Remover*. The *Equal Opportunity Difference* still presents a notable disparity for S5 plantains and S6 soybeans, indicating a growing disparity between the privileged and unprivileged groups in terms of receiving favorable outcomes, while the change for S4 potatoes is minimal. The *Average Odds Difference* decreases for all crops, with soybeans seeing the most significant drop, highlighting fairness challenges.

Overall, in the crop type use case, the *Disparate Impact Remover* again shows mixed results in its bias reduction efforts based on crop type. While the results are less favorable for potatoes and modest for plantains, the bias is almost completely mitigated for soybeans. Nevertheless, the complexity of bias mitigation becomes clear, especially when trying to balance fairness with predictive accuracy.

All in all, it is interesting to highlight that in the sub-region use case, S1 Northern Europe stands out. This sub-region, notable for its highest yield, reacts most effectively to the *Reweighing* algorithm, showing a visible reduction in bias. Similarly, in the crop type use case, S6 soybeans stands out. Despite being the crop with the lowest yield, the *Disparate Impact Remover* proves to be very effective in reducing the inherent bias for this particular crop. This contrast between the two use cases reinforces the nuanced interplay between data characteristics and the effectiveness of specific mitigation tools.

6.3 The Impact of Bias on Food Security Models

RQ1.3 How do potential biases in AI-based yield prediction models affect food security forecasts?

In the scope of this research, one region stands out: Sub-Saharan Africa. Not only is this region marked by the lowest yield within the dataset, but as stated in Chapter 2, it also faces extensive poverty and starvation.

Understanding the underlying dynamics in the context of ML is imperative. Typically, in real world scenarios, datasets do not provide an equal representation of all segments, groups, or regions. This inherent imbalance can unintentionally affect the model's predictions, often to the detriment of smaller, underrepresented groups. For Sub-Saharan Africa, the limited presence in the dataset could magnify the bias when introduced, potentially making the model's predictions less accurate or even misleading for this specific region. Moreover, the external realities of Sub-Saharan Africa, characterized by widespread poverty and hunger, may inadvertently classify it as an unprivileged group in ML terminologies. The predictions produced by the model appear to reflect these real-world disparities, serving as a strong reminder of the potential pitfalls of ML. In an area as critical as food security, such biases can have immediate consequences, leading to misinformed strategies or misdirected resources, and possibly even human deaths.

In contrast, regions that already enjoy a more privileged status, both in terms of dataset representation as well as real-world benefits, tend to do better. Northern Europe, the region with the highest yield in the dataset, does not only demonstrate superior model performance, but also a higher degree of resilience to bias. This dual advantage emphasizes the inherent benefits of being a "majority" or "privileged" group within a dataset. The model's ability to provide more accurate predictions for Northern Europe, even in the face of introduced biases, illustrates how societal privilege can be inadvertently reflected in ML results.

In conclusion, while ML offers transformative potential for sectors such as agriculture and food security, it is important to approach these models with a nuanced understanding. Without rigorous checks and balances, models can unwittingly perpetuate and even exacerbate existing societal inequalities. This underscores the need for robust bias detection and mitigation strategies, especially when addressing vital issues such as food security for vulnerable populations.

Metrics	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Bias Detection Metrics						
Accuracy	X	X	X	X	X	X
Disparate Impact	-	X	X	-	X	X
Statistical Parity Difference	-	X	X	-	X	X
Equal Opportunity Difference	-	-	-	-	-	-
Average Odds Difference	-	-	-	-	-	-
Bias Mitigation Strategy 1: Reweighing						
Accuracy after Reweighing	X	X	X	X	X	X
Disparate Impact after Reweighing	X	-	-	-	-	X
Equal Opportunity Difference after Reweighing	-	-	-	-	-	X
Average Odds Difference after Reweighing	-	-	-	-	-	X
Bias Mitigation Strategy 2: Disparate Impact Remover						
Accuracy after DIR	-	-	-	-	-	-
Disparate Impact after DIR	-	-	-	-	-	X
Equal Opportunity Difference after DIR	-	-	-	-	-	-
Average Odds Difference after DIR	-	-	-	-	-	-

Table 6: Summary of Results

The overarching research question of this thesis project, **How can bias be detected in AI based yield predictions for food security?** is resolved through the exploration and resolution of its constituent sub-questions. Each sub-question not only contributes to this overarching research question, but also expands the overall comprehension of the topic. For a concise overview of the findings of this thesis project, all results are consolidated and presented in Table 6.

The models consistently maintain high accuracy across all scenarios, despite the introduction of bias. This robust performance persists after applying the 1. bias mitigation strategy, *Reweighing*. However, the accuracy drops significantly when applying the 2. bias mitigation strategy, the *Disparate Impact Remover*.

In terms of bias detection, the fairness metrics fail to detect bias in privileged groups, specifically S1 Northern Europe in the sub-regional use case and S4 potatoes in the crop yield use case. In contrast, scenarios of the less privileged groups such as S2 South-eastern Asia, S3 Sub-Saharan Africa, S5 plantains, and S6 soybeans demonstrate detectable bias, primarily through the *Disparate Impact* and *Statistical Parity Difference* metrics. The *Equal Opportunity Difference* and *Average Odds Difference* metrics do not detect bias in any of the six scenarios.

Concerning bias mitigation, the *Reweighing* strategy improves the *Disparate Impact* metric for the S1 Northern Europe and S6 soybeans scenarios. However, the for the other scenarios, the fairness metrics essentially remain very similar or even deteriorate after *Reweighing*. Only the S6 soybeans scenario displays an improvement in the *Equal Opportunity Difference* and *Average Odds Difference* metrics.

Finally, after applying the *Disparate Impact Remover* bias mitigation strategy, only the *Disparate Impact* fairness metric for the S6 soybeans scenario shows improvement, while other metrics for the other scenarios either remain very similar or deteriorate.

6.4 Limitations

While the research offers important and useful insights, there are several limitations to be considered when interpreting the results. In particular, the study is subject to data limitations. While a broad range of variables were analyzed, critical factors such as nutrient deficiencies, which are essential to achieving optimal yields, were not included. This exclusion may potentially compromise the accuracy and depth of the predictions. Likewise, the prediction model took a simplified approach to crop yield prediction. Influential elements such as soil conditions, floods and droughts, which play a key role in determining yields, were not taken into account, which could affect the robustness of the model's predictions.

The thesis' emphasis on specific sub-regions and crop types further questions the generalizability of its findings. While the gained insights are relevant to the sub-regions and crops studied, extrapolating these findings to other sub-regions or different crops may not be straightforward.

Another clear limitation stems from tool dependency. The study relies strongly on the AIF360 toolkit for bias detection and mitigation. While AIF360 is robust, the need for binary classification could lead to potential loss of information, and other tools might offer alternative perspectives or mitigation techniques. Following this tool-centric approach also exposed bias mitigation challenges, as specific mitigation strategies were only successful in certain scenarios.

External factors, which often lie beyond the scope of data-driven models, can have a significant impact on food security. While not directly related to agricultural yields, factors such as political stability and economic standing, play an indispensable role in ensuring food security. Excluding them may provide an incomplete picture of the complex interplay of factors that affect food security.

Finally, the temporal scope of the study is a crucial point to emphasize. The study represents a snapshot in time based on current data and prevailing conditions. As the global landscape evolves, impacted by a myriad of factors ranging from climate change to geopolitical shifts, the model's predictions and the biases identified could change, impacting the long-term applicability of this thesis' findings.

7 Conclusion

The vulnerability of the world's food systems is undeniable. This fragility shows that even small perturbations can lead to significant consequences. With this in mind, the role of predictive models, particularly those based on ML, become increasingly vital in the agricultural sector. While potentially transformative, these models come with their own set of challenges, most notably the inadvertent introduction of biases.

This thesis project highlights the multifaceted impact of biases in machine learning models. While tools like AIF360 have proven effective in uncovering these biases, subsequent efforts to mitigate them do not always yields consistent results. This is evidenced by the mixed results obtained when deploying mitigation strategies like *Reweighing* and the *Disparate Impact Remover*. For the sub-region use case, the bias detection tools, particularly the *Disparate Impact* metric, highlighted existing discrepancies between sub-regions. Northern Europe, with a privileged position in the dataset, experienced positive results from the *Reweighing* algorithm. However, the same cannot be said for South-eastern Asia and Sub-Saharan Africa. The application of the *Disparate Impact Remover* proved to be a double-edged sword, at times reducing model accuracy and, in some scenarios, exacerbating inherent biases. In the crop type use case, while the models demonstrated consistent performance across different crop types, biases were still evident, particularly for plantains and soybeans. The *Reweighing* approach showed mixed effectiveness, with soybeans showing the largest benefit. In contrast, the *Disparate Impact Remover* again posed a conundrum. While it improved bias for soybeans, it exacerbated the situation for potatoes.

These results underline an important point. While ML holds great promise for improving food security, its widespread implementation presents a myriad of challenges. Biases, both subtle and obvious, can distort predictions. Left unchecked, these biases can have tangible consequences, from financial impacts over misallocation of critical resources to death. For these reasons, it is crucial that more research is conducted in this field. There are multiple promising paths for improving the scope and accuracy of ML models in agricultural predictions and food security in general.

Firstly, in light of the indisputable impact of climate change on agriculture, future research should place greater emphasis on the understanding and incorporation of its effects. This could mean to not only considering direct influence on crop yields, but rather the wider environmental and socioeconomic impacts. In particular, environmental impacts could encompass changes shifts in growing seasons and water availability, while socioeconomic implications could include shifts in food prices and land use patterns.

Furthermore, the integration of richer datasets has the potential to significantly improve the predictive accuracy of these models. Critical data elements, such as in-depth soil data and information on floods and droughts, could provide invaluable context and depth to predictions, enabling a more grounded and practical understanding of future agricultural trends.

Lastly, while existing methodologies primarily stress individual steps, whether it's data acquisition, modeling, or prediction, there is a growing recognition of the need for more holistic and exhaustive methods. Such end-to-end bias detection methodologies could encompass the entire ML process, thereby allowing for more comprehensive and nuanced insights, as highlighted by Mehrabi et al. [57].

Summarized, this thesis project lays the groundwork for future exploration, highlighting the need for a more profound comprehension of bias mitigation tools and strategies, tailored to specific scenarios and challenges. In the larger scheme of things, as the agricultural sector becomes increasingly intertwined with ML, the search for a balance between fairness and accuracy becomes paramount. This study serves as both, a demonstration of the progress that has been made in this direction and a reminder of the journey that still lies ahead.

References

[1] The state of food insecurity in the world 2014. strengthening the enabling environment for food security and nutrition, 2014.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[4] Salem Alelyani. Detection and evaluation of machine learning bias. *Applied Sciences*, 11(14):6271, 2021.

[5] Fatima Zahra Bassine, Terence Epule Epule, Ayoub Kechchour, and Abdelghani Chehbouni. Recent applications of machine learning, remote sensing, and iot approaches in yield prediction: a critical review. *arXiv preprint arXiv:2306.04566*, 2023.

[6] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[7] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR, 2018.

[8] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[9] Salah Bouktif, Ali Fiaz, Ali Ouni, and Mohamed Adel Serhani. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, 11(7):1636, 2018.

[10] Reginald Bryant, Celia Cintas, Isaac Wambugu, Andrew Kinai, and Komminist Weldemariam. Analyzing bias in sensitive personal information used to train financial models. *arXiv preprint arXiv:1911.03623*, 2019.

[11] Derek Byerlee and Jessica Fanzo. The sdg of zero hunger 75 years on: Turning full circle on agriculture and nutrition. *Global Food Security*, 21:52–59, 2019.

[12] Yaping Cai, Kaiyu Guan, David Lobell, Andries B Potgieter, Shaowen Wang, Jian Peng, Tianfang Xu, Senthold Asseng, Yongguang Zhang, Liangzhi You, et al. Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches. *Agricultural and forest meteorology*, 274:144–159, 2019.

[13] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

[14] Daniel Caputo. Minority bias assessment in healthcare. 2022.

[15] RMSR Chamara, SMP Senevirathne, SAILN Samarasinghe, MWRC Premasiri, KHC Sandaruwani, DMNN Dissanayake, SHNP De Silva, WMTP Ariyaratne, and B Marambe. Role of artificial intelligence in achieving global food security: a promising technology for future. 2020.

[16] Phusanisa Charoen-Ung and Pradit Mittrapiyanuruk. Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning. In *Recent Advances in Information and Communication Technology 2018: Proceedings of the 14th International Conference on Computing and Information Technology (IC2IT 2018)*, pages 33–42. Springer, 2019.

[17] Chen Chen, Kenneth Frank, Tiancai Wang, and Felicia Wu. Global wheat trade and codex alimentarius guidelines for deoxynivalenol: A mycotoxin common in wheat. *Global Food Security*, 29:100538, 2021.

[18] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[19] Louis Cohen, Lawrence Manion, and Keith Morrison. *Research Methods in Education*. Routledge, 8th edition, 2018.

[20] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[21] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[23] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[24] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[25] Food and Agriculture Organization (FAO). An introduction to the basic concepts of food security. EC - FAO Food Security Programme, 2008.

[26] Food and Agriculture Organization of the United Nations (FAO). How to feed the world in 2050, 2009.

[27] Niketa Gandhi, Leisa J Armstrong, Owaiz Petkar, and Amiya Kumar Tripathy. Rice crop yield prediction in india using support vector machines. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE, 2016.

[28] Milot Gashi, Matej Vuković, Nikolina Jekic, Stefan Thalmann, Andreas Holzinger, Claire Jean-Quartier, and Fleur Jeanquartier. State-of-the-art explainability methods with focus on visual analytics showcased by glioma classification. *BioMedInformatics*, 2(1):139–158, 2022.

[29] T. George. What is action research? | definition & examples, 2023.

[30] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021.

[31] Ch. Haider, Chris Clifton, and Yan Zhou. Unfair ai: It isn't just biased data. *2022 IEEE International Conference on Data Mining (ICDM)*, pages 957–962, 2022.

[32] M Hardt, E Price, N Srebro, et al. Equality of opportunity in supervised learning, in 'advances in neural information processing systems'. 2016.

[33] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[34] Meng-Leong How, Yong Jiet Chan, and Sin-Mei Cheah. Predictive insights for improving the resilience of global food security using artificial intelligence. *Sustainability*, 12(15):6272, 2020.

[35] Ruixing Huang, Chengxue Ma, Jun Ma, Xiaoliu Huangfu, and Qiang He. Machine learning in natural and engineered water systems. *Water Research*, 205:117666, 2021.

[36] Economist Impact. The global food security index (gfsi) 2022, 2022.

[37] Nahina Islam, Md Mamunur Rashid, Santoso Wibowo, Cheng-Yuan Xu, Ahsan Morshed, Saleh A Wasimi, Steven Moore, and Sk Mostafizur Rahman. Early weed detection using image processing and machine learning techniques in an australian chilli farm. *Agriculture*, 11(5):387, 2021.

[38] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pages 4721–4732. PMLR, 2021.

[39] Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916, 2021.

[40] Jig Han Jeong, Jonathan P Resop, Nathaniel D Mueller, David H Fleisher, Kyungdahm Yun, Ethan E Butler, Dennis J Timlin, Kyo-Moon Shim, James S Gerber, Vangimalla R Reddy, et al. Random forests for global and regional crop yield predictions. *PloS one*, 11(6):e0156571, 2016.

[41] Kavita Jhajharia and Pratistha Mathur. A comprehensive review on machine learning in agriculture domain. *IAES International Journal of Artificial Intelligence*, 11(2):753, 2022.

[42] B. Manjula Josephine, K. Ruth Ramya, K.V.S.N. Rama Rao, Swarna Kuchibhotla, P. Venkata Bala Kishore, and S. Rahamathulla. Crop yield prediction using machine learning. *International Journal of Scientific & Technology Research*, 9(02), February 2020.

[43] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[44] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011.

[45] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pages 869–874. IEEE, 2010.

[46] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. Exploiting reject option in classification for social discrimination control. *Information Sciences*, 425:18–33, 2018.

[47] Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020.

[48] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018.

[49] Nesaretnam Barr Kumarakulasinghe, Tobias Blomberg, Jintai Liu, Alexandra Saraiva Leao, and Panagiotis Papapetrou. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 7–12. IEEE, 2020.

[50] Nadia Lambek. The un committee on world food security's break from the agricultural productivity trap. In *Transnational Food Security*, pages 241–255. Routledge, 2020.

[51] Erin Lentz and Joanna Upton. Benefits to smallholders? evaluating the world food programme's purchase for progress pilot. *Global Food Security*, 11:54–63, 2016.

[52] Xiaofeng Steven Liu. A probabilistic explanation of pearson's correlation. *Teaching Statistics*, 41(3):115–117, 2019.

[53] Allison Marie Loconto, Anne Sophie Poisot, and Pilar Santacoloma. *Innovative Markets for Sustainable Agriculture*. Food and Agriculture Organization of the United Nations and Institut National de la Recherche Agronomique, 2016.

[54] Yuchuan Luo, Zhao Zhang, Juan Cao, Liangliang Zhang, Jing Zhang, Jichong Han, Huimin Zhuang, Fei Cheng, and Fulu Tao. Accurately mapping global wheat production system using deep learning algorithms. *International Journal of Applied Earth Observation and Geoinformation*, 110:102823, 2022.

[55] Margaret Macherera and Moses J. Chimbari. A review of studies on community-based early warning systems. *Jamba*, 8(1):206, 2016.

[56] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[58] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 17–38. Springer, 2020.

[59] Sanjay Motia and SRN Reddy. Exploration of machine learning methods for prediction and assessment of soil properties for agricultural soil management: a quantitative evaluation. In *Journal of Physics: Conference Series*, volume 1950, page 012037. IOP Publishing, 2021.

[60] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping. Crop yield prediction with deep convolutional neural networks. *Computers and electronics in agriculture*, 163:104859, 2019.

[61] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[62] Kavir Osorio, Andrés Puerto, Cesar Pedraza, David Jamaica, and Leonardo Rodríguez. A deep learning approach for weed detection in lettuce crops using multispectral images. *AgriEngineering*, 2(3):471–488, 2020.

[63] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

[64] Tiago Palma Pagano, Rafael Bessa Loureiro, Fernanda Vitória Nascimento Lisboa, Gustavo Oliveira Ramos Cruz, Rodrigo Matos Peixoto, Guilherme Aragão de Sousa Guimarães, Lucas Lisboa dos Santos, Maira Matos Araujo, Marco Cruz, Ewerton Lopes Silva de Oliveira, et al. Bias and unfairness in machine learning models: a systematic literature review. *arXiv preprint arXiv:2202.08176*, 2022.

[65] Xanthoula Eirini Pantazi, Dimitrios Moshou, Thomas Alexandridis, Rebecca L Whetton, and Abdul Mounem Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and electronics in agriculture*, 121:57–65, 2016.

[66] Rafael Pérez-Escamilla. Food security and the 2015–2030 sustainable development goals: From human to planetary health: Perspectives and opinions. *Current developments in nutrition*, 1(7):e000513, 2017.

[67] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[68] Soma Prathibha, SV Hariharan, G Shamini, Kaajal Krishnamurthy, Maheshwari Avudaiappan, et al. Ai-assisted farming for crop recommendation & virtual assistance application. In *2022 1st International*

Conference on Computational Science and Technology (ICCST), pages 90–95. IEEE, 2022.

[69] Alexander Y Prosekov and Svetlana A Ivanova. Food security: The challenge of the present. *Geoforum*, 91:73–77, 2018.

[70] Charan Reddy. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. 2022.

[71] Drew S. Roselli, Jeanna Neefe Matthews, and Nisha Talagala. Managing bias in ai. *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

[72] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31, 2018.

[73] Raí A Schwalbert, Telmo Amado, Geomar Corassa, Luan Pierre Pott, PV Vara Prasad, and Ignacio A Ciampitti. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. *Agricultural and Forest Meteorology*, 284:107886, 2020.

[74] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

[75] Mohsen Shahhosseini, Guiping Hu, and Sotirios V Archontoulis. Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, 11:1120, 2020.

[76] Abhinav Sharma, Arpit Jain, Prateek Gupta, and Vinay Chowdary. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873, 2020.

[77] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[78] Department of Economic United Nations and Population Division Social Affairs. World population prospects: The 2015 revision, key findings and advance tables. Working Paper No. ESA/P/WP.241, 2015.

[79] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020.

[80] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature communications*, 11(1):1–10, 2020.

[81] Fraol Gelana Waldamichael, Taye Girma Debelee, Friedhelm Schwenker, Yehualashet Megersa Ayano, and Samuel Rahimeto Kebede. Machine learning in cereal crops disease detection: a review. *Algorithms*, 15(3):75, 2022.

[82] Xinlei Wang, Jianxi Huang, Quanlong Feng, and Dongqin Yin. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of china with deep learning approaches. *Remote Sensing*, 12(11):1744, 2020.

[83] Patrick Webb, Erin Boyd, Saskia de Pee, Lindsey Lenters, Martin Bloem, and Werner Schultink. Nutrition in emergencies: Do we know what works? *Food Policy*, 49(Part 1):33–40, 2014.

[84] Shu Xu, Bo Lu, Michael Baldea, Thomas F Edgar, Willy Wojsznis, Terrence Blevins, and Mark Nixon. Data cleaning in the process industries. *Reviews in Chemical Engineering*, 31(5):453–490, 2015.

[85] Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 2021.

[86] zbenmo. Detecting and mitigating bias in machine learning models using shap and aif360. https://github.com/zbenmo/detecting-and-mitigating-bias-in-machine-learning-models-using-shap-and-aif360/blob/main/Copy_of_milestone_3_instructions.ipynb, 2020.

[87] Yukun Zhang and Longsheng Zhou. Fairness assessment for artificial intelligence in financial industry, 2019.

A Code Repository

This appendix provides information about the GitHub repository associated with this thesis project.
The GitHub repository contains the code, the data, the visualizations and other materials related to the research presented in this thesis project. It can be accessed via the following link: `https://github.com/Nicinicinici/Masterthesis`