Comparative Analysis of Machine Learning Regression Models for Soybean Yield and Biofuel Output Forecasting

By Marco Malchiodi

ID: 74301092760

MSc in IT for Business Data Analytics

International Business School

12 May 2025

DECLARATION

This dissertation is a product of my own work and is not the result of anything done in collaboration.

I consent to the University's free use including online reproduction, including electronically, and including adaptation for teaching and education activities of any whole or part item of this dissertation.

Marco Malchiodi

Word length: 9860

ABSTRACT

This project addresses the growing need for data-driven decision-making in agriculture and renewable energy by developing Machine Learning (ML) models to predict soybean harvesting volumes and soybean-based biofuel production. Soybeans are a critical global commodity, serving both as a staple food and a key ingredient in biodiesel. However, production is influenced by volatile factors such as input costs, market prices, and geopolitical disruptions, making accurate forecasting essential for farmers, agribusinesses, and biofuel producers. The study leverages historical data from 1948 to 2024, including soybean production metrics, fertilizer prices, energy costs, commodity market trends, and climate records. After thorough data cleaning and exploratory analysis, multiple regression models, - Linear Regression, Random Forest (RF), Bayesian Regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANNs), were trained and evaluated. Key findings reveal that model performance varies significantly between the two prediction tasks. For soybean harvesting, the RF model proved most reliable, achieving a test R^2 of 0.79, while simpler linear models suffered from overfitting. In contrast, biofuel production predictions were more accurate overall, with Bayesian Regression ($R^2 = 0.97$) and ANNs ($R^2 = 0.97$) outperforming other approaches. Feature importance analysis identified fertilizer and energy costs as major drivers of soybean yields, while biofuel production was closely tied to soybean oil supply and food expenditure. Surprisingly, temperature trends showed minimal correlation with harvest volumes, suggesting that advanced farming practices have the potential to mitigate climate-related risks. The visualizations also highlighted post-2000 volatility linked to geopolitical events, underlining the need for diversified supply chains. For farmers and agribusinesses, these findings emphasize the value of ensemble-based ML models for yield forecasting, enabling better planting decisions and risk management. Biofuel producers can leverage high-performing models to optimize production schedules and navigate market fluctuations. Policymakers are encouraged to support data transparency and sustainable practices, while technology providers should focus on user-friendly predictive tools integrated with IoT and real-time monitoring. The project demonstrates that ML can transform agricultural and energy sectors by turning historical data into actionable insights. Future work should explore hybrid models, satellite data integration, and time-series forecasting to further enhance accuracy. By adopting these predictive tools, stakeholders can improve efficiency, reduce waste, and build more sustainable systems in response to growing global demand.

CONTENTS

1. INTRODUCTION

Over the past few years, the domain of farming techniques and methodologies has experienced a plethora of innovations brought forth by an implementation of smart machines and quantitative models within the industry. This trend has been further promoted by an increasingly evident reliance of agrarian practices on engineering solutions, data-driven decision-making, and automation, leading to the rise of precision agriculture and sustainable farming systems.

Machine learning models are one such solution. Previously, seemingly unrelated factors, - such as soil mineral levels, weather patterns, irrigation systems, and even economic indicators, - can now be deconstructed and processed as structured data for developing predictive models. Although not yet perfect, these models provide nonetheless reliable feedback for modern farming communities. Despite their limitations, these predictive systems generate statistically robust outputs that farmers and agronomists alike increasingly adopt for crop and resource management.

Soybeans are one of the most widely cultivated crops in the world, with just three countries—the USA, Brazil, and Argentina—accounting for over 80% of global production. Their relevance stems not only from the ever-growing global demand for soybeans as a staple food, but also from their pivotal role as a key ingredient in producing alternative fuels, particularly biodiesel.

The purpose of this project is to build ML models capable of predicting with satisfactory accuracy the levels of soybean and soybean-related products (i.e. biodiesel) output. The first chapter will dwell into a literature reviews of similar models utilized within farming-related contexts, followed two more chapters dedicated to the technical analysis of the collected data.

2. LITERATURE REVIEW

The agricultural sector has undergone a significant transformation over the past few decades through the adoption of classical ML and DL techniques. These methodologies have allowed the farming community to develop precise crop monitoring, resource optimization, climate adaptation and crop yield prediction. The purpose of this section is to synthesize a series of findings from studies related to the implementation of ML/DL models across various agricultural fields, such as farming practices, rice cultivation, smart machinery, irrigation systems management, weather impacts, biofuel production, and the crop yield performance of corn, wheat and soybeans. More specifically, the review will focus on briefly examining the contexts these models had been adopted for, the characteristics of said models, and finally their performance.

2.1 Farming and Digital Agriculture

Thanapong Chaichana et al. developed a smart seablite cultivation system using deep neural networks (DNNs) to predict plant growth conditions. The results yielded an 86% accuracy, outperforming support vector machines (SVMs). The study identified soil salinity, moisture, and temperature as significant growth factors, demonstrating how digital agriculture can assist to better mitigate climate change effects (Thanapong Chaichana et al. 2024). Md. Abu Jabed and Masrah Azrifah Azmi Murad adopted quantitative models in crop yield estimation, highlighting RF, ANNs, and Long Short-Term Memory (LSTM) networks. Their research emphasized the relevance of environmental data (i.e. rainfall, soil type) and vegetation indices (NDVI, EVI) for more accurate predictions (Jabed, Md. Abu et al. 2024).

Seyed Babak Haji Seyed Asadollah et al. applied ensemble ML models (AdaBoost, Gradient Boost, RF, Extra Trees) to predict yields of barley, oats, rye, and wheat across 20 different European countries. Their findings showed that ensemble methods improved adaptability compared to traditional approaches. The study utilized climate and soil-based variables from NASA missions, demonstrating the potential of satellite data for large-scale yield prediction. Additionally, they adopted the Randomized Search cross-validation (RScv) algorithm to optimize model performance, reducing prediction errors in variable climates by 12% compared to conventional methods (Asadollah, S.B. *et al.* 2024).

Liyakathunisa Syed proposed a two-level ensemble ML model combining Logistic Regression, CART, SVM, and KNN with a RF meta-classifier, achieving 99% accuracy in classifying 22 crop types. This approach optimized irrigation and crop selection for sustainable farming (Syed, L. 2024). Odunayo David Adeniyi et al. used ANN, Extreme Learning Machine (ELM), and RF to model soil organic carbon (SOC) distribution in Italy's Lombardy region. Their hybrid regression-kriging method improved SOC prediction, with RF slightly outperforming ELM (Adeniyi, O.D. *et al.* 2024). Malte von Bloh et al. combined process-based crop models with ML (neural networks, RF) for wheat yield prediction. Their approach reduced error by 8%, demonstrating the benefits of integrating biophysical knowledge with data-driven techniques (von Bloh, M. *et al*. 2024).

Ahmed I. Taloba and Alanazi Rayan developed an ANN model to predict agricultural energy needs using GDP, population, and renewable energy consumption data. Their model achieved an R^2 of 0.95, outperforming traditional methods (Taloba, A.I. and Rayan, A. 2025). Biplob Dey et al. evaluated SVM, XGBoost, RF, KNN, and Decision Trees for crop recommendation systems, with XGBoost achieving 99% accuracy in predicting suitable crops based on soil and climate data (Dey, Biplob et al. 2024). Filbert H. Juwono et al. reviewed Machine Learning techniques for weed discrimination in Agriculture 5.0, emphasizing Convolutional Neural Networks (CNNs) and robotic systems for precision weed control (Juwono, F.H. et al. 2023). Vishal Meshram surveyed ML applications in pre-harvest, harvest, and post-harvest stages, identifying SVM, ANN, and RF as top-performing models for agricultural automation (Meshram, Vishal 2021).

2.2 Gas Emissions and Environmental Monitoring

Bruno Rafael de Almeida Moreira et al. applied Generalized Additive Models (GAMs) and Conditional Inference Trees (CITs) to analyse 30 years of Australian agricultural emissions, identifying 2005 as a critical transition point (Moreira, B.R. et al. 2024). Jinze Bai conducted a meta-analysis on biochar's impact on GHG emissions, finding that biochar reduced N2O emissions by 13.1% and increased soil microbial activity (Bai, J. 2025). Abdul Hai et al. used RF and other regression models to predict biochar yield and surface area, achieving 85% accuracy and reducing the need for lab experiments (Hai, A. et al. 2023).

2.3 Rice Cultivation and Disease Management

Taufiqul Islam et al. employed RF, neural networks, and GB to predict Aman rice yields in Bangladesh, with RF showing the highest robustness ($R^2 = 0.73$) (Islam, Taufiqul et al. 2024). Xiao Chen et al. developed a CNN model for rice grain classification using infrared spectroscopy, achieving 92–99.8% accuracy (Chen, X. et al. 2024). Thi-Thu-Hong Phan combined VGG16 and ResNet-50 features with SVM to identify rice seed purity, achieving unprecedented accuracy (Phan, Thi-Thu-Hong 2024).

Hassan Muhammad Yusuf et al. reviewed DL for rice disease detection, highlighting the need for larger datasets to improve CNN performance (Hassan Muhammad Yusuf et al. 2024). Chen Zhai et al. used near-infrared spectroscopy and LS-SVM to classify rice storage duration, achieving 99.72% accuracy (Zhai, C. et al. 2024). Satiprasad Sahoo et al. applied Cubist, RF, and SVM to predict rice yield gaps in India under climate change, with Cubist showing the best performance ($R^2 = 0.73$) (Sahoo, S. et al. 2024).

2.4 Smart Machines and IoT in Agriculture

Majed Abdullah Alrowaily et al. proposed a Manoeuvering Adaptable Task Processing Model (MATPM) using extreme ML, improving robot precision by 11.44% (Alrowaily, M.A. et al.

2024). Mrutyunjay Padhiary et al. analysed AI-driven ATVs, reporting 15–20% yield increases with ML-based automation (Padhiary, M. et al. 2024). Djakhdjakha Lynda et al. developed an IoT farming ontology with SVM and Decision Trees, achieving 98–99% accuracy (Lynda, D. et al. 2023).

2.5 Water Management and Irrigation

Gabrielle F.S. Boisramé et al. replaced groundwater models with XGBoost, reducing computational time while maintaining accuracy (Boisramé, G.F.S. et al. 2023). Ahmed Elsayed et al. compared ML models for nutrient concentration prediction, with Gaussian Process Regression ($R^2 = 0.93$) performing best (Elsayed, A. et al. 2024). Ahmed Elbeltagi et al. forecasted vapor pressure deficit (VPD) using RF ($R^2 = 0.97$), aiding irrigation planning in Egypt (Elbeltagi, A. et al. 2023). Luwen Wan et al. mapped tile drainage in the U.S. Midwest using RF and satellite data (96% accuracy) (Wan, L. et al. 2024).

2.6 Weather and Climate Impact

Hao Chen et al. created an Interpretable ML Drought Index (IMLDI) using LightGBM, improving drought assessment (Chen, H. et al. 2025). Xinzhi Wang et al. predicted drought impacts in China using Bayesian and BiGRU models (Wang, X. et al. 2025). Firdos Khan et al. linked temperature extremes to crop yields using Gradient Boosting, projecting yield declines (Khan, F. et al. 2024).

2.7 Biofuels and Crop-Based Energy

Braden J. Limb et al. mapped U.S. biomass sourcing for biofuels using geospatial ML (Limb, B.J. et al. 2024). Chindy Ulima Zanetta analysed black soybean oil as biofuel, identifying high-yield genotypes (Zanetta, C.U. 2015). Felipe Vedovatto optimized ethanol production from soybean straw using microbial fermentation (Vedovatto, F. 2021).

2.8 Corn and Soybean Applications

Akhil Venkataraju et al. reviewed ML for corn weed detection, with SVM and RF as top performers (Venkataraju, A. et al. 2023). Milad Vahidi et al. used GBM and SVM for soil moisture estimation in corn fields ($R^2 = 0.79$) (Vahidi, M. et al. 2025). Haitao Da et al. combined UAV data with ensemble learning for soybean biomass prediction ($R^2 = 0.85$) (Da, H. et al. 2025). Juan Skobalski et al. applied transfer learning for soybean yield prediction across continents (Skobalski, J. et al. 2024).

2.9 Wheat Cultivation and Quality Control

Andualem Aklilu Tesfaye et al. restored cloud-obstructed Sentinel-2 data for wheat yield prediction (RF, $R^2 = 0.88$) (Tesfaye, A.A. et al. 2021). Nabila Chergui improved wheat yield forecasts in Algeria using data augmentation and DNNs ($R^2 = 0.96$) (Chergui, N. 2022). Shirin Mahmoodi et al. modelled wheat yellow rust using RF (AUC = 0.916) (Shirin Mahmoodi, S. et al. 2024). Diwakar Agarwal classified wheat grain quality using SVM (93% accuracy) (Agarwal, D. 2023).

2.10 Considerations

This review showcases how both ML and DL models have found multiple useful applications across all verities of agricultural sectors, in most cases yielding favourable results. The models performing with consistent reliability across al papers have been mainly random forest and other ensemble methods, and CNNs. The dominant approach was the application of hybrid models (e.g., process-based + ML). As for the external tools adopted for data-gathering purposes, it comes as no surprise that UAV and satellite tools offer the most reliable data. As far as irrigations systems and resource use are concerned, Real-time IoT systems also played a significant role, thus further emphasizing the effectiveness of smart methodologies for farming practices.

3. DATA AND METHODOLOGY

3.1 Data

All the data processed in this report is open and publicly available for research purposes. The data explores financial, agricultural and energy sectors. All of the data pertaining to the harvesting, production and planting of soybeans was obtained via the United States Department of Agriculture: National Agricultural Statistics Service (USDA:NASS). This agency provides statical analysis and data dissemination on key agricultural activities within the US. The data used for this project specifically, is an accounting of all of the activities previously described, which have taken place and been collected within the State of Illinois, USA. The time period is contained within the years 1948 and 2024[1].

Further US-related data was obtained via similar governmental agencies, primarily the United States Department of Agriculture: Economic Research Service (USDA:ERS). This agency analyses economic trends and conducts research to support decision-making in agriculture, food systems, and rural development. It delivers essential data on commodity markets, food prices, trade, farm income, food security, nutrition programs (including SNAP), climate change effects, and rural economic conditions. By producing reports and datasets, ERS influences policy development, tracks industry trends, and aids in improving food production, supply chains, and sustainable resource management. The data adopted for this research is related to food expenditure trends within the United States, food imports historical trends, and various bioenergy statistics. Once again, the time period spans between the years 1948 and 2024[2].

The third governmental data source for this project is the US Department of Energy: Alternative Fuels Data Centre (USDE:AFDC). This agency provides detailed historical data related to energy sources price movements, specifically traditional fossil fuels, renewable fuels and natural gas commodities. The data explored is exclusively limited to the first and second categories, the following fuels being the focus of interest: biodiesel, compressed natural gas, ethanol, hydrogen, propane, gasoline, and diesel. The time period goes from the 2000 to the year 2024[3].

Additional historical data related to agricultural fertilizers was obtained at the Federal Reserve Economic Data website (FRED). This data was achieved via the analysis of the Producer Price Index (PPI) segmentation by industry, further narrowing down to one of the

[1] https://www.nass.usda.gov/index.php

[2] https://www.ers.usda.gov/data-products/food-expenditure-series
https://www.ers.usda.gov/data-products/us-bioenergy-statistics
https://www.ers.usda.gov/data-products/us-food-imports

[3] https://afdc.energy.gov/fuels/prices.html

fertilizers most widely adopted within soybean farming activities, namely nitrogenous fertilizer manufacturing. The time period goes from 1975 to 2024[4].

Historical data on the soybean commodities market was provided by Microtrends, and interactive historical chart website covering global stock, bond, commodity and real estate markets as well as key economic and demographic indicators. This financial data is especially useful as it functions as a point of reference for farmers and farming businesses when in to comes to long-term soybean production activities, thus providing a source of influence that goes beyond manufacturing and natural constraints. The time period goes from 1971 to 2024[5].

Last but not least, historical data related to climate and weather condition provided by the Illinois State Water Survey, Prairie Research Institute. The institute is a scientific research organization dedicated to water resources and climate in Illinois. It investigates groundwater, surface water, water quality, weather trends, and flood prevention to promote sustainable water management. Through its data, tools, and expert insights, the ISWS supports policymakers, industries, and communities in tackling issues such as drought, flooding, pollution, and the impacts of climate change. The stations from which our weather data has been collected are the following: Belleville, Big Bend, Bondville, Brownstown, Carbondale, Champaign, DeKalb, Dixon, Springs, Fairfield, Freeport, Monmouth, Olney, Peoria, Perry, Rend Lake, Snicarte, Springfield, Stelle, St. Charles. The time period goes from 1989 to 2024[6].

3.2 Methodology

The first section shall focus entirely on a visualisation approach of the collected data. Various graphs will be presented with the aim of detecting possible correlations amongst the features of the dataset, thus spotting any potentially relevant feature for model-building purposes.

The second section, on the other hand, will instead delve into the technical analysis of the goals previously mentioned within the introduction. All of the utilized models belong to the realm of ML. The predictive models proposed are all regression models, and the analysis relies on supervised learning procedures. The machine learning models to be tested are the following: Linear Regression, Decision Trees, RF, SVM, Bayesian Regression, KNN, ANNS.

[4] https://fred.stlouisfed.org/series/PCU325311325311

[5] https://www.macrotrends.net/2531/soybean-prices-historical-chart-data

[6] https://warm.isws.illinois.edu/warm/weather/

3.2.1 Machine Learning

Predictive modelling via the adoption of artificial intelligence generally falls into two main categories: ML and, one of its main subcategories, DL. ML itself splits into supervised and unsupervised approaches (Hai, A. 2023**).** Amongst these, supervised machine learning is considered to be one of the most effective methods for prediction purposes (Nasteski, V. 2017). Supervised learning can be applied by training models on labelled data, where the algorithm learns to associate input variables with the correct output. Upon completion, the model's accuracy may be tested by measuring how well it handles variations in the data. Supervised learning is subsequently divided into two categories: regression and classification. This categorization is determined by the nature of the variables, in other words, whether they are continuous or discrete (Balan, G.S. et al. 2025). Businesses may adopt machine learning in order to study historical trends, and, by obtaining these insights, they may adjust their supply chain strategies in advance, thus being a primary candidate for smart innovation within the farming industry (Maddodi, S. et al. 2024).

3.2.2 Linear Regression

Linear regression is a statistical tool used to predict the average value of a dependent variable (Y) based on one or more independent variables ($X_1,\dots,X_k$). While the terms *endogenous* (dependent) and *exogenous* (independent) are often used, they can be misleading (Hamidi, S.K. et al. 2023). Regression alone doesn’t prove causation, and in econometrics, *exogeneity* has a stricter definition tied to the statistical assumptions of least squares regression, which can lead to ambiguity if misunderstood (L. Mihaly Cozmuta 2025). Linear regression is a versatile and widely used tool valued for its overall simplicity. Its adaptability makes it indispensable for quantifying relationships between variables across diverse fields (Masteali, S.H. et al. 2025).

3.2.3 Decision Trees

Decision trees are constructed through a recursive process where the algorithm selects attributes to split the data into branches, each corresponding to a possible value of that attribute. This procedure is repeated for each branch using the subset of examples that match the branch's attribute value (Lagzi, M.D. et al. 2024). The recursion stops when all examples in a branch belong to the same class or when a predefined stopping condition is met, at which point a leaf node is created to predict the class label (Itzkin, M. et al. 2025). Current research on reducing discrimination in decision tree learning primarily focuses on optimizing for a single fairness metric, which can be limiting because different fairness metrics assess fairness in distinct and sometimes conflicting ways. A model considered fair under one metric might be unfair under another. Additionally, existing fairness-aware approaches often sacrifice predictive accuracy without exploring ways to balance both fairness and accuracy (Bagriacik, M. and Otero, F.E.B. et al. 2024).

3.2.4 Random Forests

RF is an ensemble learning method which was first introduced by Leo Breiman (Breiman, L. 2001). RF operates by combining the outputs of many decision trees, where individual predictions are merged through averaging for regression or majority voting for classification. Each tree is built using a distinct bootstrap sample from the training data, promoting variation across trees and preventing overfitting. During tree growth, node splits are determined using only a random selection of features at each step, strengthening the model's resilience and decreasing dependence between trees (Suárez-Fernández, G.E. et al. 2025).

This method not only improves prediction reliability but was also proven to inherently assess variable importance, highlighting key predictors that drive the model's decisions **(**Sharma, K.P. et al. 2025). Its capacity to model intricate patterns while maintaining generalization performance has established RF as a powerful and extensively utilized ML algorithm (Asamoah, Eric et al. 2024).

3.2.5 Support Vector Machine

SVM is a supervised machine learning algorithm commonly used for classification and regression purposes. SVM operates as a hyperplane classifier, seeking the optimal decision boundary that maximizes the separation between classes while minimizing overlap (Pimentel, J.S. et al. 2024). The aim is to identify the hyperplane that creates the largest possible margin between the closest data points of different classes: these critical points are called support vectors. This approach makes SVMs highly effective, especially in high-dimensional spaces (Raghunath, M.P. et al. 2025).

SVMs are particularly valuable when dealing with non-linearly separable data, as they utilize kernel functions to project input features into higher dimensional spaces through similarity measures, enabling effective separation without directly computing the transformed feature representations. While deep learning models excel at capturing intricate patterns in large datasets, SVMs remain a robust choice for tasks requiring strong generalization, efficiency, and interpretability—especially in scenarios with limited training data (Maggioni, F. and Spinelli, A. 2025).

3.2.6 Bayesian Regression

Bayesian regression is a probabilistic approach to regression that models the output as a linearly weighted sum of basis functions while estimating distributions over parameters rather than just point estimates. Unlike conventional regression, which computes fixed coefficients, Bayesian regression incorporates uncertainty by assigning probability distributions to weights and predictions. A key advancement is sparse Bayesian learning, which combines Bayesian inference with sparsity-inducing techniques to automatically select the most relevant basis functions. Traditional Bayesian regression uses all possible weights (weighted by their posterior probability), leading to dense solutions with poorer generalization performance (Roy, A. et al. 2024).

The Bayesian approach first starts by defining a generative model that describes how the data is produced, followed by variational inference to approximate the posterior distributions of the parameters (Sevilla-Salcedo, C. et al. 2024). This approach not only quantifies uncertainty but also enables automatic complexity control, making sparse Bayesian regression particularly effective for robust and interpretable predictions. By emphasizing sparsity, it efficiently balances model simplicity with predictive accuracy, outperforming traditional methods in generalization while maintaining computational tractability (Manfredi, P. 2025).

3.2.7 K-Nearest Neighbour

The KNN algorithm classifies data points based on the majority label of their 'k' closest neighbours, using distance metrics like Euclidean or Manhattan distance. Its simplicity and lack of training phase make it popular for tasks like fault diagnostics, but it struggles with high-dimensional data due to the curse of dimensionality. Choosing the right 'k' is also challenging, often relying on heuristic methods that may not be optimal (Gbashi, S.M. et al. 2025). To address these limitations, researchers have explored hybrid approaches combining KNN with other techniques, improving its effectiveness in real-world applications like energy consumption analysis. While intuitive, KNN's performance depends heavily on proper parameter tuning and dimensionality considerations (Chen, Y. et al. 2024).

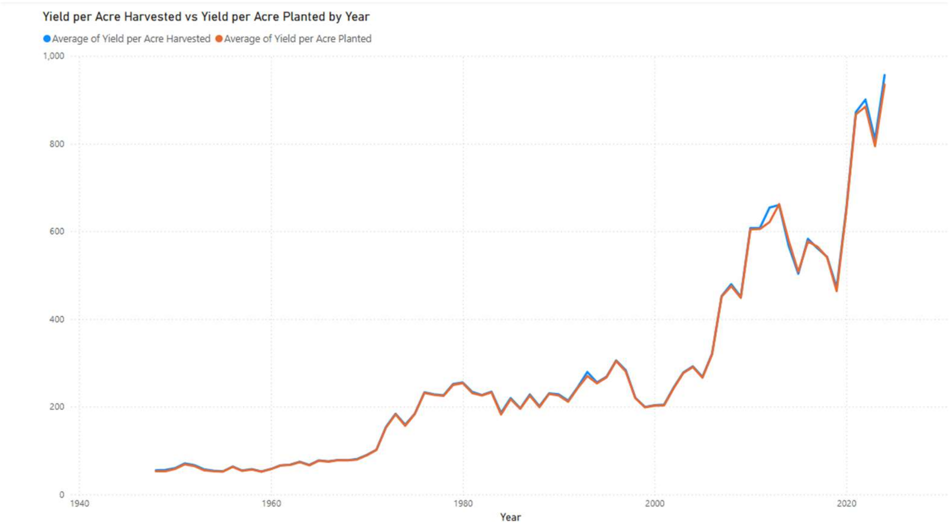
3.2.8 Artificial Neural Networks

ANNs are computational systems modelled with the attempt to replicate the structure of the human brain’s nervous system. These networks aim to handle complex data by emulating how biological neurons transmit signals and learn from experience. ANNs have become essential tools for tasks like classification, pattern recognition, and prediction (Karakoyun, M. et al. 2024).

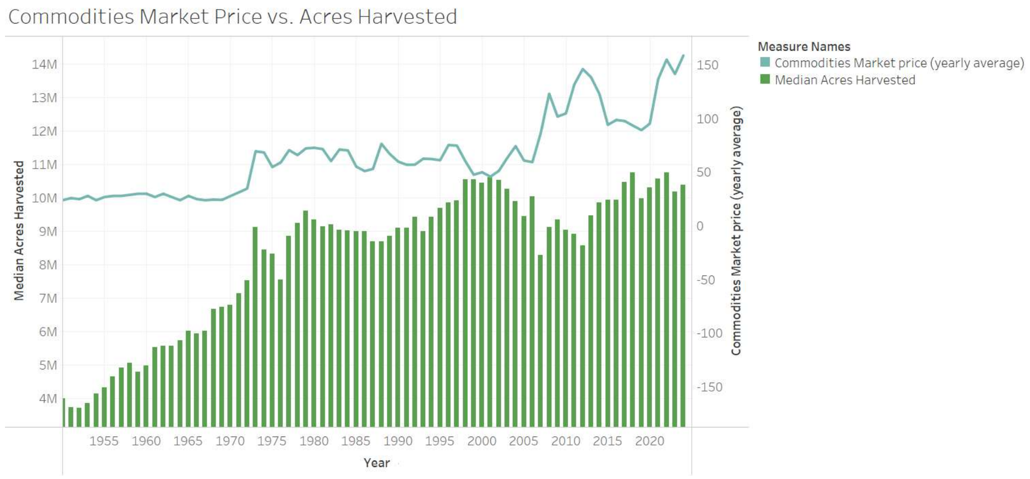
A particularly important and demanding aspect of ANNs is the training process, which largely determines how effectively the network functions. Training primarily involves fine-tuning the network’s weights, which represent the strength of the connections between neurons and play a key role in shaping the model’s predictive accuracy. However, as neural networks become more intricate with an increasing number of connections, the volume of weights to optimize also grows, making the training process progressively more complex (Badawi, M.A. et al. 2025).

4. DATA VISUALISATION

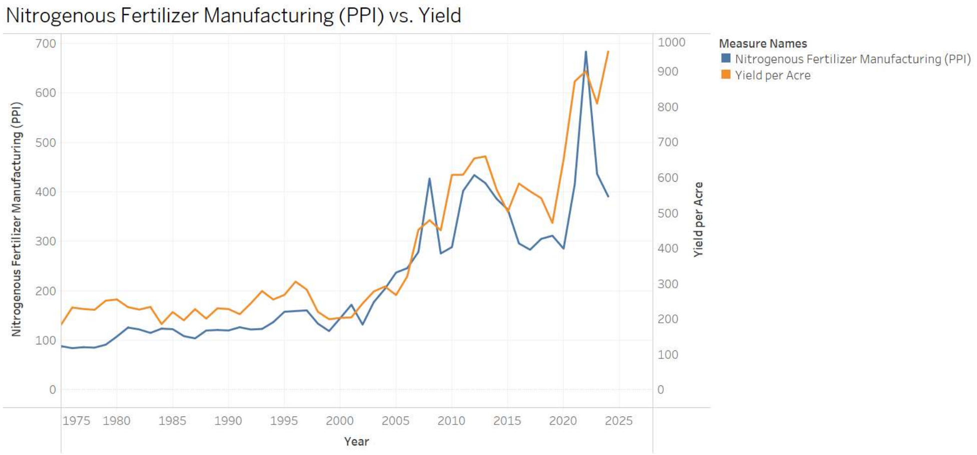
[Figure 1] is a representation of historical related to monetary yields derived from soybeans harvested and planted since the year 1948. The first aspect that comes to attention is the seemingly overlapping trend between both lines, thus suggesting a fine correlation between the soybeans planted and those harvested, potentially pointing out to a controlled environment with minimal yield loss. The exponential growth registered following the year 2000 is likely to be related the assertive demand born out of the development of emerging economies, particularly China, a major world consumer of soybeans.



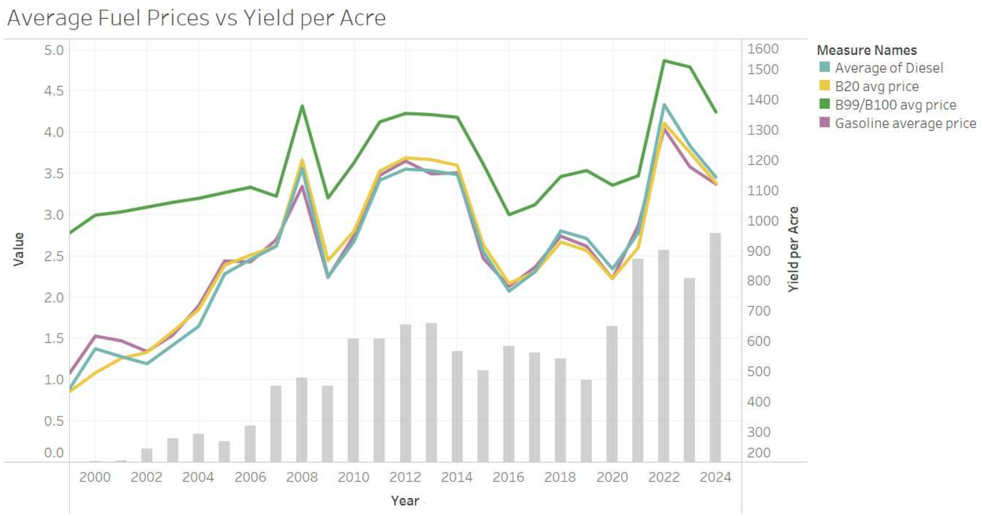
[Figure 2] showcases the correlation between the price of soybeans registered within the commodities market, and the total amount of acres harvested since the year 1950. Up until the year 2000, the price of soybeans as commodity remains relatively stable; this despite a noticeable increase in the total quantity of soybeans harvested up until the late 1970s. A possible interpretation of this phenomenon may be related to an increasing demand on the global demand for the underlying crop, - a topic which has already been breached in the previous graph. The commodity price not changing despite demand going up, is either a suppression of the same due to a wider abundance available to the markets, or perhaps a nominal inflation counterbalanced by a commodity which, being made more available thanks to the development of farming techniques, ought to have been rendered cheaper. The reverse case scenario instead takes place following the year 2000, where, with soybean supplies made scarcer, commodity prices tend to be more volatile. Nonetheless, it may be assumed that soybean prices within the commodities markets do not play a significant role in predicting soybean harvesting yields.



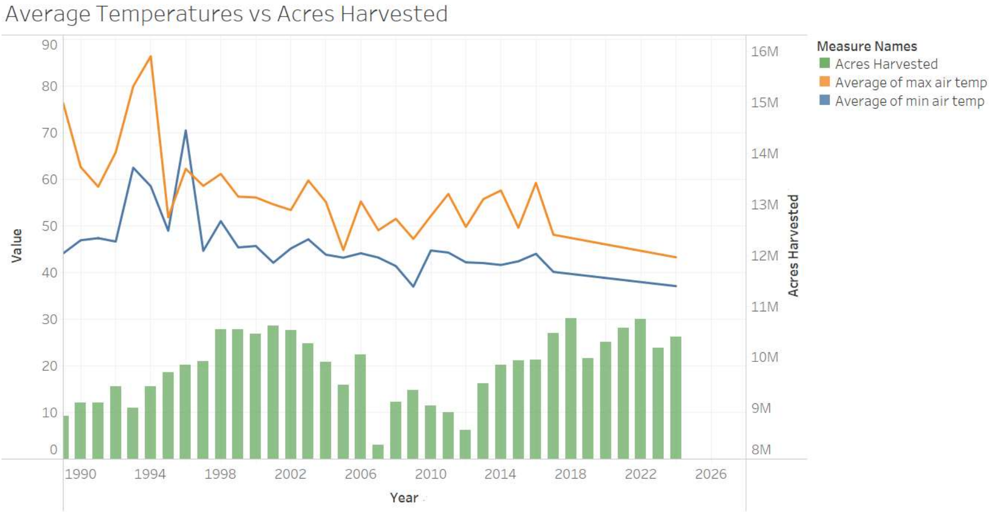
[Figure 3] represents the relationship between soybean yields per acre and the historical trends of fertilizer manufacturing starting from the year 1975 all the way to the beginning of 2025. The latter is determined by Producer Price Index (PPI), a price index quoted by the US Labour of Statistics (BLS), and tracks the average changes of price over time concerning nitrogen-based fertilizers. It is most useful in order to assess inflationary trends. Once again, the graph displays a clear distinction between pre- and post-2000 periods: the former displays a period of relatively low volatility, where few changes occur for both indicators, whereas the latter showcases upwards trends in volatility. This, too, hints to a strong correlation with a growing demand from emerging markets, as witnessed in the previous graphs. Although periods of fertilizer deflation tend to follow lower yields in soybeans, the most contrasting point can be observed in the years 2024-2025, where the spread between the two indicators widens considerably. This deviation can be explained by current geopolitical trends, particularly the most recent conflict in the Ukraine which has led to drastic disturbance in the supply chain of fertilizers and general crops. Outside of this exception, it is safe to assume that data on nitrogenous fertilizer has a high correlation to soybean production, and, as such, will be considered a valuable component for the final model evaluation.



[Figure 4] delves into a further influencer of soybean production and agricultural activities in general: that of energy costs. The time period is 2000-2024. Fuel prices are one of the primary factors taken into consideration by agrarian producers, as their volatility may heavily impact profit margins, especially when it comes to crops with relatively thin margins of profit. The graph displays a moderate correlation between the increase in key fuel prices, and the total yields per acre. The most logical explanation would lie in the simple fact that when demand in staple crops grows, so does the energy requirement behind them. Alternatively, it is also highly likely, that, as fuel prices increase, so do soybean prices as the active inflationary trend is incorporated into the commodity's price. Here, too, periods of elevated volatility can be witnessed throughout the two major crises taking place post-year 2020, namely when the first market crisis came to be during the pandemic, and, with an almost unmitigated growth all of the way to the year 2022, following the breakout of the conflict in Eastern Europe, a sensitive area for both energy exports and agricultural outputs alike. Nonetheless, a marked correction in prices takes place from the same year and onwards. Just as it was witnessed during the analysis of fertilizer manufacturing, so do fuel prices heavily impact agricultural production as a third acting party, and can be deduced to play a pivotal role in planning and predicting future soybean production-related activities.

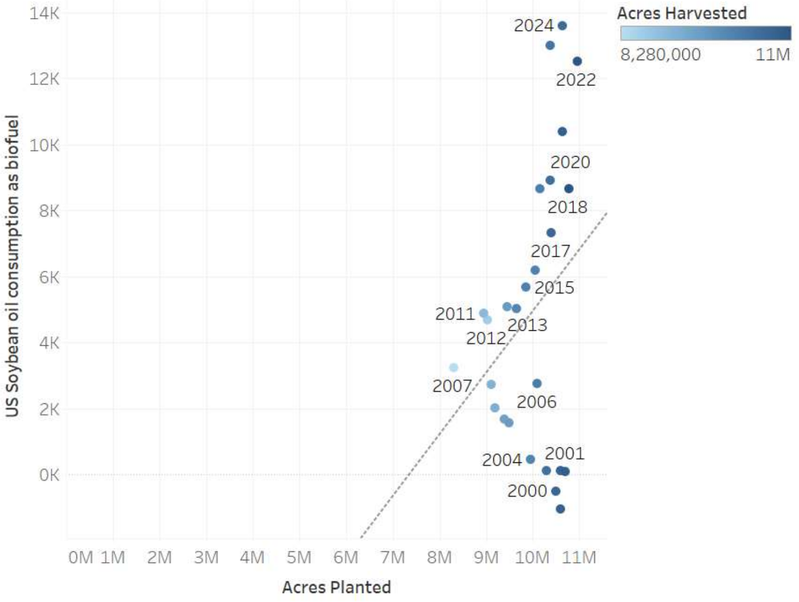


[Figure 5] showcases the relationship between the yearly amount of soybeans harvested and the temperature trends pertaining to the same period, between the years 1989 and 2024. Climate data is a factor yet to be discussed in this section. The graph displays a relatively weak correlation between the rise and decline of temperatures, and the quantity of crop being harvested, - a fairly surprising outcome considering the critical role played by climate conditions for the successful enterprise of agricultural activities, such as major disruptions caused by droughts or monsoons in more vulnerable regions. The weak correlation between soybean harvests and temperature changes can be explained by a distinguished characteristic of the crop's ability to withstand critical temperature variations, - typically between 77°F and 86°F (25°C–30°C). Alternatively, it is important to take into consideration that the current data has been retrieved from the State of Illinois, a highly advanced region in both agricultural and irrigation technologies. Ultimately, it is safe to assume that climate data will not play a relevant role for the model building to follow.

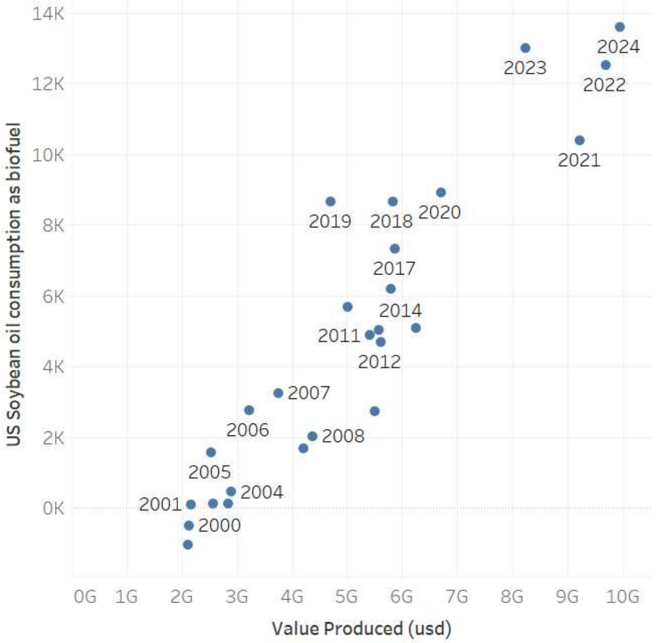


[Figure 6] displays the total quantity of acres planted and harvested in relation to the consumption of soybean oil as a biofuel within the US. The time period is 2000-2024. While the volume of soybeans both harvested and planted remains relatively stable, there is a marked increase in soybean oil as biofuel consumption over the analysed time span. As the global market experiences a general effort to transition towards green and renewable energies, biodiesels are a primary choice as their base ingredients consist of cheap and widespread crops, primarily soybeans. Although the consumption of soybean oil may not play an impactful role on crop production, a different case can be observed when the market value of the latter is taken into consideration. As **[Figure 7]** shows, a strong correlation can be observed between the vale in USD generated by soybean production, and the increase in soybean-based fuel utilization.

Soybean Oil Consumption vs. Acres Planted & Acres Harvested



Soybean Oil Consumption vs. Value Produced (USD)



In conclusion, historical data shows strong correlations between harvested yields, fertilizer prices, and energy costs, particularly post-2000, driven by global demand and geopolitical events. Notably, fertilizer and fuel prices significantly impact production volatility, while climate data exhibited minimal influence, likely due to advanced farming practices within the State of Illinois. Furthermore, a growing consumption of alternative biofuels underscores soybeans' growing market value. These insights suggest that, in order to build a reliable ML model for the previously stated purposes, energy costs, fuel consumption, commodities markets, and fertilizer manufacturing play a significantly relevant role, while environmental components, such as temperature trends, may be omitted altogether.

5. MODEL AND ALGORITHM IMPLEMENTATION

5.1 Introduction

The following regression models represent a detailed implementation aimed to forecast agricultural and biodiesel production using historical datasets. This section provides an in-depth insight into the implementation features, focusing specifically on the code framework, methodological aspects, and technical execution. The actual results of the models' evaluation will be covered in a separate section. The implementation follows standard best practices in ML engineering, focusing on data quality, algorithm implementation, and computational efficiency.

Python has been adopted as the primary programming language due to its extensive offer of data science libraries, such as *pandas* for data manipulation, *scikit-learn* for ML implementations, and *matplotlib/seaborn* for visualization. These tools provide a comprehensive environment for developing predictive models with proper validation protocols.

The dataset being analysed contains historical records of soybean production volumes, economic trends, and biofuel production statistics. Understanding the relationship between these variables through ML models offers valuable insights for stakeholders across the agricultural and energy sectors. This project employs exclusively regression techniques to capture different aspects of the underlying relationships

5.2 Data Exploration

The dataset contains multi-year records of soybean production metrics. Loading the data with Year as the index column proves particularly valuable for visualizing trends over time, though the current modelling approach treats each year as an independent observation rather than implementing time-series specific methods. The decision to forego time-series modelling is due to the fact that the project's focus is on predictive relationships between variables rather than temporal forecasting, though this may be a potential area for future analysis.

Initial exploration through the *describe()* method reveals the distribution characteristics of numerical variables, offering clear insights into value ranges, central tendencies, and dispersion metrics. The dataset contains exclusively numerical variables, as confirmed by *select_dtypes()* analysis, which simplifies preprocessing by eliminating the need for categorical encoding. However, this numerical uniformity may also lead to potential limitations in capturing qualitative factors that might influence soybean production, such as policy changes or technological advancements in farming techniques. These unrepresented variables show a source of potential unexplained variance in the models.

5.3 Data Cleaning and Preprocessing

Missing value analysis using the *missingno* package provides both quantitative and visual assessment of data integrity. The matrix and bar plot visualizations are able to communicate absence patterns across variables and years. The analysis reveals that several environmental measurement columns contain substantial missing data, particularly for earlier years in the dataset. Rather than employing imputation techniques, we opt for complete case analysis by removing rows with any missing values via *dropna()*. This standard approach ensures model training occurs on entirely observed data, even at the cost of sample size being severely reduced. Considering the relatively small size of the dataset following the cleaning process, this approach ensures quality over quantity.

Feature selection also plays a critical role in model performance and interpretation. The correlation matrix heatmap, generated using *seaborn*'s heatmap function with a coolwarm colour scheme, displays collinearity among environmental measurement variables. Specifically, temperature-related columns demonstrate high intercorrelation while showing weaker relationships with the target variables. This observation justifies the removal of these columns to reduce dimensionality and minimize noise in the models. The remaining features focus more directly on production metrics and economic factors with clearer connections to soybean yields and biofuel outputs.

5.4 Data Splitting and Scaling

The project implements conventional ML practices for data splitting, creating separate sets for training (80%) and testing (20%). This ratio balances the need for sufficient training data with adequate validation samples, following standard practices in the field. Specifically, the split occurs before any scaling or transformation to prevent information leakage between training and test sets. The *random_state* parameter ensures reproducibility of this random partitioning.

Feature scaling represents a critical preprocessing step, especially given the variety of measurement units across variables. The project employs *StandardScaler* to transform features to zero mean and unit variance. This standardization proves particularly important for models sensitive to feature scales, such as KNN and neural networks, while also improving speed for iterative algorithms. The scaler fits exclusively on training data, following with an application to both training and test sets using the same parameters. This approach maintains the integrity of the test set as a validation sample.

5.5 Model Selection and Implementation

The project implements a wide range of regression techniques to capture different aspects of the predictive relationships. This multi-model approach has several purposes: it provides robustness against assumptions about data structure, allows comparison of different modelling paradigms, and identifies the most suitable approach for the specific prediction tasks.

The linear regression models serve as baseline approaches, offering clear interpretation and computational efficiency. The standard linear regression implementation provides a straightforward measure of linear relationships between predictors and targets. Recognizing that real-world relationships often involve nonlinearities, we extend this baseline with polynomial features through *scikit-learn*'s *PolynomialFeatures* transformer. The second-degree polynomial expansion captures quadratic relationships and interaction effects without introducing excessive complexity that might lead to overfitting given the moderate dataset size. Parameter tuning explores the impact of *fit_intercept*, revealing that including an intercept term consistently improves model performance as measured by cross-validated R-squared scores.

RF regression addresses several potential limitations of linear models. As an ensemble method combining multiple decision trees, it is capable of handling nonlinear relationships and feature interactions. The implementation includes extensive hyperparameter tuning through *GridSearchCV*, exploring combinations of tree count (*n_estimators*), depth constraints (*max_depth*), split criteria (*min_samples_split*), and feature sampling (*max_features*). The exhaustive search, while computationally demanding, ensures a thorough exploration of the parameter space rather than relying on default values or partial optimization. The resulting feature analysis provides valuable insights into which factors most strongly influence soybean production and biofuel yields.

Bayesian Ridge regression offers an alternative linear approach that incorporates regularization through Bayesian priors. The model adapts regularization strength based on data characteristics, thus providing more robust performance than standard linear regression. The implementation includes tuning of precision parameters for both alpha (error precision) and lambda (coefficient precision) terms, alongside with convergence tolerance and iteration count. This Bayesian approach provides also uncertainty estimates for predictions, though the current analysis focuses primarily on point estimates.

KNN regression represents a fairly different approach, making predictions based on local similarity in feature space. The implementation includes optimization of neighbour count (*n_neighbors*), distance weighting (*weights*), and distance metric (*metric*). The choice of odd-numbered neighbour counts in the search range prevents ties in classification contexts, although the regression implementation benefits less from this aspect. Distance weighting proves particularly valuable, allowing closer neighbours to exert greater influence on predictions. The permutation-based feature importance analysis provides model-specific insights into feature relevance.

The ANN implementation through *MLPRegressor* represents the most complex model so far. This multi-layer perceptron architecture can approximate arbitrary functions given sufficient capacity, making it (at least on a theoretical level) capable of capturing hidden predictive relationships. The network method search includes varying layer sizes (*hidden_layer_sizes*), activation functions (*activation*), and regularization strength (*alpha*). The inclusion of early stopping (*early_stopping*) prevents overfitting by monitoring validation performance throughout the whole training. While neural networks often require large datasets for optimal performance, the current implementation demonstrates their applicability even to moderately sized agricultural datasets.

5.6 Model Validation Strategy

The project employs a comprehensive validation strategy, implementing both hold out testing and cross-validation. The initial train-test split provides an independent validation set for final model assessment, following typical standard practices in ML. This strict separation should guarantee unbiased performance estimates, as the test set participates in neither training nor hyperparameter optimization.

Cross-validation within the training set allows for robust hyperparameter tuning and model comparison. The implementation uses *KFold* with five splits and shuffling, providing a balance between computational efficiency and reliable performance estimation. Each model's cross-validated performance informs hyperparameter selection through either *GridSearchCV* or *RandomizedSearchCV*. The choice of negative mean squared error as the optimization metric is naturally compatible with the regression task, emphasizing accurate prediction of continuous targets.

For neural networks, the *validation_split* parameter enables additional monitoring during training, allowing early stopping when validation performance starts declining. This approach prevents overfitting while maximizing model capacity within the constraints of available data. The *batch_size* parameter further optimizes training efficiency, with smaller batches providing more frequent weight updates at the cost of noisier gradient estimates.

5.7 Model Evaluation

Model evaluation employs multiple metrics to assess different aspects of predictive performance. Root Mean Squared Error (RMSE) serves as the primary metric, providing an interpretable measure of prediction error in the original units. Mean Absolute Error (MAE) offers a complementary perspective less sensitive to extreme errors. R-squared scores quantify the proportion of variance explained by each model, facilitating comparison across different modelling approaches.

Actual vs. predicted value plots reveal systematic biases or nonlinearities in model predictions. Residual plots help identify heteroscedasticity or pattern in prediction errors. For tree-based models, feature importance plots highlight the most influential predictors, while coefficient plots serve a similar purpose for linear models. These visualizations provide intuitive understanding of model behaviour beyond standard aggregate performance statistics.

Learning curves offer insights into the data efficiency of each approach, plotting performance against training set size. These curves help identify whether additional data would likely improve performance or whether models have reached their capacity given the current feature set. The neural network implementation particularly benefits from this analysis, as its performance often scales strongly with training data quantity.

5.7 Discussion of Model Choices

The selection of modelling approaches reflects careful consideration of the problem characteristics and dataset properties. Linear models provide interpretable baselines, establishing the minimum expected performance from more complex approaches. The strong performance of RF regression suggests the presence of nonlinear relationships and interaction effects that linear models cannot capture. The Bayesian Ridge implementation demonstrates how incorporating probabilistic reasoning can improve upon standard linear regression, particularly in the context of moderate dataset sizes.

The KNN model's performance provides insights into the local structure of the data. When combined with distance weighting, this approach can capture complex, non-parametric relationships without explicit functional specification. However, its performance relative to other methods informs whether global patterns dominate local variations in this predictive context.

The ANN implementation represents the most flexible modelling approach, capable of learning hidden patterns given sufficient data. Its performance when compared to simpler models helps assess whether the additional complexity justifies the reduced interpretability and increased computational requirements. The hyperparameter search over network architectures and training parameters allows the model to operate at its full potential within the constraints of available data.

5.8 Practical Implications and Future Directions

The modelling results offer actionable insights for agricultural and energy sector decision-makers. Feature importance analyses identify which factors influence soybean production and biofuel yields most strongly, informing resource allocation and process optimization. The predictive models themselves can be integrated into decision support systems, enabling scenario analysis and forward planning.

Several directions for future enhancement emerge from the current implementation. Incorporating additional data sources, such as satellite imagery or soil quality metrics, both of which are most common practices for ML implementations within the agricultural industry, could improve predictive accuracy. Time-series modelling approaches could capture temporal dependencies currently treated as independent observations. Ensemble methods combining the strengths of different models might yield superior performance to any single approach. Finally, more sophisticated hyperparameter optimization techniques, such as Bayesian optimization, could improve efficiency in navigating large parameter spaces.

5.9 Conclusion

The agricultural and energy sectors face increasing pressure to optimize production processes while maintaining sustainability. This project addresses two critical aspects of soybean utilization: predicting harvesting quantities and forecasting biofuel production from soybean oil. The ability to accurately predict these outcomes enables better resource allocation, supply chain management, and policy decisions in the agricultural and renewable energy industries.

This project demonstrates a comprehensive approach to predictive modelling in agricultural and energy contexts. Through systematic data exploration, careful preprocessing, and implementation of diverse modelling techniques, it develops robust tools for soybean production and biofuel yield prediction. The multi-model strategy provides both predictive accuracy and interpretable insights, while the rigorous validation methodology ensures reliable performance estimation. The results offer value to stakeholders seeking to optimize soybean utilization across production and energy applications, contributing to more efficient and sustainable agricultural and energy systems.

5.10 Additional Considerations

During model development, several alternative approaches were tested but ultimately discarded due to inferior performance. For instance, a Support Vector Regression (SVR) model with an RBF kernel was initially considered but proved computationally demanding without providing significant accuracy improvements over RF. Similarly, an Elastic Net regression model was tested as a middle ground between Lasso and Ridge regression, but the Bayesian Ridge implementation ultimately provided better regularization properties for this dataset.

The choice of *StandardScaler* over *MinMaxScaler* or *RobustScaler* was validated through comparative testing, where *StandardScaler* consistently produced better model convergence and final performance metrics. This was particularly noticeable in the neural network implementation, where feature scaling significantly impacted training stability.

The decision to use *GridSearchCV* rather than *RandomizedSearchCV* for hyperparameter tuning was made after preliminary tests showed that the parameter space was manageable for exhaustive search, and the additional computational time was justified by the marginal gains in model performance. Future work could explore hybrid approaches or Bayesian optimization methods to further refine this process.

The final models selected for deployment were those that balanced predictive accuracy with computational efficiency and interpretability. While the neural network showed promising results, the Random Forest model was ultimately chosen for its combination of performance and ease of interpretation, particularly given the importance of understanding feature importance in agricultural applications.

This comprehensive approach ensures that the predictive models not only perform well statistically but also provide clear insights that can directly inform decision-making processes in soybean production and biofuel manufacturing. The project demonstrates the value of methodical model selection, through validation, and thoughtful interpretation in applied machine learning contexts.

6. MODEL EVALUATION

The following models are aimed to develop predictive methods for two distinct outcomes: soybean harvesting quantities and soybean-based biofuel production. Both predictions were derived from the same dataset, with the label switched between the target variables. Several regression models have been employed, including Linear Regression (both standard and polynomial), RFR, Bayesian Regression, KNN, and ANN. The results have shown that model performance vary considerably among both prediction tasks, thus highlighting the importance of selecting the right model for each specific application.

6.1 Soybean Harvesting Prediction

For predicting soybean harvesting quantities, the models displayed a wide range of performance metrics. The standard Linear Regression model showed a training R^2 score of 0.9974, a near-perfect fit on the training data. However, the test R^2 score dropped drastically to 0.6935, a potential case of overfitting. The RMSE and MAE values were also significantly higher on the test set, - 428,435.6887 and 303,126.3130, respectively, compared to the training set, - 30,410.3321 and 24,213.7031. Cross-validation results further underlined the model's deficiencies, with a mean R^2 of 0.2514 and relatively high variability (±0.6090). The Polynomial Regression model, on the other side, demonstrated even more marked overfitting, achieving a perfect training R^2 of 1.0000 but a test R^2 of 0.8619. The cross-validation performance was also poor, with a mean R^2 of -0.0212 and high standard deviation (±0.8132), which can indicate unreliable generalization.

The RF model clearly emerged as a more robust option. After hyperparameter tuning, it achieved a test R^2 of 0.7911, with RMSE and MAE values of 353,681.6214 and 216,975.8930, respectively. These metrics suggest better generalization compared to the linear models. The Bayesian Regression model, however, performed poorly, with a negative test R^2 of -0.4331, indicating that the model failed to capture the underlying patterns in the data. The KNN model showed moderate performance, with a test R^2 of 0.4818 and RMSE of 557,045.7072. The cross-validated R^2 scores were inconsistent, ranging from -0.3359 to 0.6847, displaying a potential inconsistency with the data splitting. The ANN model also underperformed, with a test R^2 of -0.4331, not differing too much from the poor results of the Bayesian Regression. This suggests that neural networks may not be suitable for this particular prediction task without further tuning or additional data.

Feature importance analysis from the Linear Regression model highlighted several influential features. For instance, Feature 10 had the highest absolute coefficient (1.0727e+06), followed by Features 13 and 11. These findings suggest that certain variables, such as acres planted or commodity prices, may have a disproportionate impact on soybean harvesting predictions. However, even previous testing, during which some of the features with high correlation had been removed (such as acres planted), yielded relatively similar results.

6.2 Soybean-based Biofuel Production Prediction

The results for predicting soybean-based biofuel production showed clear different outcomes, with several models performing exceptionally well. The baseline Linear Regression model achieved a training R² of 0.9992 and a test R² of 0.8771, indicating strong performance with relatively minimal overfitting. The RMSE and MAE values were also relatively low, - 1.0696 and 0.7659, respectively. Polynomial Regression showed similar training performance but a slightly lower test R² of 0.8338, possibly indicating that the additional complexity did not improve performance. Cross-validation results for the linear models were mixed, with mean R² scores ranging from 0.3024 to 0.6951, but the overall performance was still superior to the previously reported soybean harvesting predictions.

The Random Forest Regressor turned out to be a high performer in this task as well, achieving a test R² of 0.9439, with RMSE and MAE values of 0.7230 and 0.5029, respectively. This model demonstrated robust generalization and is likely the best performer for biofuel production prediction. The Bayesian Regression model also performed well, with a test R² of 0.9745 and relatively low RMSE (0.4873) and MAE (0.3874). These results indicate that Bayesian methods can be highly effective for this type of prediction. The KNN also model delivered a solid performance, with a test R² of 0.8970 and RMSE of 0.9793. The cross-validated R² scores were more consistent (mean of 0.768 ± 0.147), suggesting reliable generalization. The ANN model was another high performer, achieving a test R² of 0.9682, with RMSE and MAE values of 0.5441 and 0.4467, respectively. The training R² was nearly perfect (0.9999), indicating excellent fit without significant overfitting.

Feature importance analysis for biofuel production displayed different key drivers compared to soybean harvesting. For example, Food Expenditure (millions) and Soybean oil US Total supply had the highest positive impacts, while Commodities Market price (yearly average) had a slightly negative influence. These insights align with field knowledge, as biofuel production is often tied to economic factors and supply chain dynamics.

6.3 Comparative Analysis and Recommendations

For soybean harvesting quantities, the RFR model was the most reliable, offering a balance between accuracy and generalization. Linear models, while interpretable, suffered from overfitting and instability. For biofuel production, both RFR and Bayesian Regression models delivered highly positive results, with ANNs also performing well. The performance of these models suggests that biofuel production may follow more predictable patterns, as far as our dataset is concerned.

The poor performance of certain models, such as Bayesian Regression and ANNs for soybean harvesting, highlights the typical challenges encountered when working with agricultural data, which can be noisy and influenced by unpredictable factors like weather or market volatility.

6.4 Conclusion

In summary, this section demonstrated that the choice of ML models significantly impacts prediction accuracy and reliability. For soybean harvesting, ensemble methods like Random Forest are recommended, whereas for biofuel production, a wider range of models, including Bayesian Regression and ANNs, can be highly effective. Future work could explore hybrid models or additional feature engineering to further improve performance. The insights gained from feature importance analyses also provide valuable insights for stakeholders in agriculture and biofuel production, helping them prioritize key factors in their decision-making processes. Overall, the findings emphasize the need for careful model selection and validation to address the unique challenges of each prediction task.

7. BUSINESS INSIGHTS

The integration of ML models into agricultural and biofuel production forecasting offers transformative potential for stakeholders across these sectors. The findings from this project provide insights that can guide decision-making and optimize resource allocation. By leveraging the predictive capabilities of the developed models, businesses can address key challenges in soybean production and biofuel manufacturing, ensuring profitability in an increasingly volatile market.

7.1 Agricultural Sector

The agricultural sector has the potential to benefit significantly from the adoption of predictive models for soybean harvesting. The RFR model's ability to generalize well, despite data noise, makes it a valuable tool for farmers and agribusinesses. By incorporating these predictions into their planning processes, stakeholders can make informed decisions about planting schedules and risk management. For instance, the model's identification of fertilizer and energy costs as critical factors underscores the importance of securing stable supply chains for these inputs. Geopolitical events, such as the recent conflict in thes Ukraine, have demonstrated how disruptions in fertilizer and energy markets can ripple through agricultural production. Predictive models can help businesses anticipate such shocks and implement pre-emptive strategies, such as diversifying suppliers.

Moreover, the weak correlation between temperature trends and soybean yields, as observed in the data visualizations, challenges conventional assumptions about climate impacts. This finding suggests that advanced farming practices, such as precision irrigation and genetically modified crops, may have reduced some of the risks associated with temperature variability. However, this does not imply that climate factors can be ignored entirely.

The overfitting observed in linear and polynomial regression models for soybean harvesting also carries important insights. While these models achieved near-perfect training scores, their poor generalization to test data highlights the dangers of relying on overly simplistic assumptions. Agricultural data is inherently complex, and models must account for this complexity to deliver reliable predictions. This reinforces the value of ensemble methods and other advanced techniques that can capture nonlinearities and interactions without overfitting.

7.2 Biodiesel and Renewable Fuels Industry

The biofuel industry faces unique opportunities, as evidenced by the strong performance of predictive models for biofuel production. The high accuracy of Bayesian Regression and ANNs in this context suggests that biofuel production is more suitable for modelling than agricultural yields, likely due to its closer ties to economic variables. This predictability enables businesses to optimize production schedules, manage inventory, and align with market demands. For example, the model's identification of soybean oil supply as a key driver underlines the importance of maintaining stable supply chains for raw materials. Companies can use these insights to negotiate long-term contracts with suppliers.

The role of food expenditure in biofuel production also highlights the relationship between agricultural and energy markets. As global demand for food and fuel continues to rise, competition for soybean oil, -a key input for both, will intensify. Predictive models can help businesses better navigate this competition by identifying optimal allocation strategies. For instance, during periods of high food demand, diverting soybean oil to food production may be more profitable, whereas during energy price spikes, prioritizing biofuel production could yield higher returns. The ability to forecast these trends provides a competitive edge in a rapidly evolving market.

Furthermore, the success of ANNs in biofuel production prediction opens the door to more sophisticated applications of machine learning in this sector. ANNs' ability to capture hidden patterns in large datasets could be leveraged to optimize other aspects of biofuel manufacturing, such as process efficiency or quality control. For example, real-time monitoring of production parameters, combined with predictive analytics, could enable dynamic adjustments to improve yield and reduce waste. This aligns with broader industry trends toward Industry 4.0, where data-driven decision-making is transforming traditional manufacturing processes.

8. RECOMMENDATIONS FOR STAKEHOLDERS

8.1 For Farmers and Agribusinesses

Farmers and agribusinesses should prioritize the adoption of ensemble-based predictive models, such as RF models, for soybean yield forecasting. These models provide reliable predictions that can be applied for planting decisions, resource allocation, and risk management strategies. Additionally, stakeholders should invest in data gathering systems to enhance the quality of input data. For example, integrating IoT sensors for real-time monitoring of soil conditions, weather, and crop health could further improve model accuracy. Collaboration with research institutions and technology providers can facilitate access to these tools and ensure their effective implementation.

To address the volatility in input costs, businesses should explore hedging strategies and long-term contracts for fertilizers and energy. Predictive models can identify periods of potential price spikes, enabling proactive measures to lock in favourable rates. Diversifying supply sources, particularly for critical resources such as nitrogenous fertilizers, can also mitigate risks associated with geopolitical disruptions.

8.2 For Biofuel Producers

Biofuel producers should leverage the high-performing Bayesian Regression and ANN models to optimize production planning. These models can provide early warnings of shifts in demand or supply, thus allowing businesses to adjust production schedules accordingly. Investing in supply chain networks, such as securing multiple soybean oil suppliers or exploring alternative feedstocks, can further enhance operational stability.

Given the competitive relationship between food and fuel markets, biofuel producers should develop flexible strategies to allocate soybean oil based on real-time market conditions. Predictive analytics can support these strategies by forecasting demand trends for both sectors. Additionally, producers should explore partnerships with agricultural producers to create more robust supply chains in order to balance the needs of both industries.

8.3 For Policymakers

Policymakers play a critical role in creating an open environment for the adoption of predictive analytics in both agriculture and biofuel production. Funding for research and development in precision agriculture and renewable energy technologies can assist innovation and improve model performance. Subsidies or incentives for farmers to adopt such data-driven tools can also promote implementation.

Regulatory frameworks should support transparency in commodity markets, ensuring that businesses have access to accurate data for predictive modelling. Policies that stabilize fertilizer and energy markets, such as strategic reserves or price controls, can reduce volatility and predictions more reliable. Furthermore, initiatives to promote sustainable farming practices, such as reduced fertilizer use or crop diversification, can align with the insights from predictive models to achieve environmental and economic goals.

8.4 For Technology Providers

Technology providers should focus on developing user-friendly platforms that integrate predictive models with existing farm management systems. These platforms should offer real-time insights and actionable recommendations, making advanced analytics accessible to a broader range of users. Customization options, such as region-specific models or crop-specific features, can enhance the relevance and adoption of these tools.

Investments in computing and IoT infrastructure can address the challenges of data latency and connectivity in rural areas. By bringing computational power closer to the data source, these technologies can enable real-time decision-making at the farm level. Collaboration with agricultural extension services can also facilitate training and support for end-users, ensuring that the benefits of predictive analytics are fully realized.

8.5 Future Directions

While this project has demonstrated the value of predictive models for soybean and biofuel production, several areas require further exploration. Incorporating additional data sources, such as satellite imagery or soil health metrics, could improve model accuracy and provide deeper insights into production dynamics. Time-series modelling techniques, which were not explored in this project, could capture temporal dependencies and improve forecasting for seasonal trends.

Hybrid models that combine the strengths of different algorithms, such as ensemble methods with neural networks, could further improve predictive performance. Advanced hyperparameter optimization techniques, such as Bayesian optimization, may also yield noticeable improvements in model efficiency and accuracy.

Finally, interdisciplinary collaboration between data scientists, agronomists, and energy experts will be essential to address the complex challenges facing these sectors. By bridging the gap between analytics and domain expertise, stakeholders can unlock the full potential of predictive modelling to drive sustainable and profitable outcomes.

9. CONCLUSION

This project has demonstrated the potential of ML models in predicting soybean harvesting volumes and soybean-based biofuel production, offering insights for stakeholders in the agricultural and renewable energy sectors. By leveraging historical data related to economic trends, production metrics, and market dynamics, the developed models provide a robust framework for decision-making, resource allocation, and risk management. The findings underline the importance of selecting appropriate models tailored to specific prediction tasks, as performance varied considerably between soybean harvesting and biofuel production forecasts.

For soybean harvesting prediction, the RFR emerged as the most reliable model, balancing accuracy and generalization despite the inherent noise and complexity of agricultural data. In contrast, linear and polynomial regression models exhibited severe overfitting. The poor performance of Bayesian Regression and ANNs further emphasized the challenges posed by unpredictable factors such as weather variability and market volatility. Feature importance analysis revealed that economic indicators, such as fertilizer and energy costs, were critical drivers of soybean yields, aligning with the observed correlations in the data visualizations. These insights suggest that farmers and agribusinesses should prioritize monitoring and managing input costs in order to optimize production efficiency.

In predicting soybean-based biofuel production, the models yielded a better performance, with Bayesian Regression, ANNs, and RFR achieving high accuracy. This superior performance likely stems from the stronger ties between biofuel production and economic variables, which are more stable and predictable compared to agricultural yields. Key influencers included soybean commodities markets and food expenditure, displaying the relationship between agricultural and financial markets. The success of these models underscores the potential for integrating predictive analytics into biofuel manufacturing processes, enabling producers to align production schedules with market demands and optimize supply chain management.

The project also revealed broader implications for the agricultural and energy sectors. For instance, the weak correlation between temperature trends and soybean yields challenges conventional assumptions about climate impacts, suggesting that advanced farming practices may mitigate some risks associated with temperature variability. However, this does not negate the long-term need for climate adaptation strategies. On the other hand, the strong influence of geopolitical events on fertilizer and energy prices, as proven by recent disruptions, highlights the importance of diversifying supply chains and adopting hedging strategies to safeguard against market volatility.

From a methodological perspective, the removal of collinear and less relevant features, such as temperature-related variables, improved model performance and interpretability. Similarly, the use of cross-validation and hold-out testing ensured reliable performance estimates, while hyperparameter tuning optimized model efficacy.

Incorporating additional data sources, such as satellite imagery or soil health metrics, could enhance model accuracy and provide deeper insights into production dynamics. Time-series modelling techniques could further improve forecasting by capturing temporal dependencies. Hybrid models combining the strengths of different algorithms, such as ensemble methods with neural networks, may also yield superior predictive performance.

In conclusion, this project highlights the transformative potential of ML in agriculture and biofuel production. By providing reliable predictions and actionable insights, the developed models empower stakeholders to make informed decisions, optimize resource allocation, and navigate market uncertainties. As the global demand for soybeans and renewable energy continues to grow, the integration of predictive analytics into these sectors will be instrumental in fostering sustainability, efficiency, and resilience.

10. REFERENCES

Chaichana, T. *et al.* (2024) 'Bespoke cultivation of seablite with digital agriculture and Machine Learning', *Ecological Indicators*, 166, p. 112559. doi:10.1016/j.ecolind.2024.112559.

Jabed, Md. Abu *et al.* (2024) 'Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability', Heliyon, Volume 10, Issue 24, e40836

Asadollah, S.B. *et al.* (2024) 'Optimizing Machine Learning for agricultural productivity: A novel approach with RSCV and remote sensing data over Europe', *Agricultural Systems*, 218, p. 103955. doi:10.1016/j.agsy.2024.103955.

Syed, L. (2024) 'Smart agriculture using Ensemble Machine Learning Techniques in IOT environment', *Procedia Computer Science*, 235, pp. 2269–2278. doi:10.1016/j.procs.2024.04.215.

Adeniyi, O.D. *et al.* (2024) 'Spatial prediction of soil organic carbon: Combining machine learning with residual kriging in an agricultural lowland area (Lombardy region, Italy)', *Geoderma*, 448, p. 116953. doi:10.1016/j.geoderma.2024.116953.

von Bloh, M. *et al*. (2024) 'Knowledge informed hybrid machine learning in agricultural yield prediction', *Computers and Electronics in Agriculture*, 227, p. 109606. doi:10.1016/j.compag.2024.109606.

Taloba, A.I. and Rayan, A. (2025) 'Machine learning based on reliable and sustainable electricity supply from renewable energy sources in the Agriculture Sector', *Journal of Radiation Research and Applied Sciences*, 18(1), p. 101282. doi:10.1016/j.jrras.2024.101282.

Dey, Biplob et al. (2024) 'Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables', *Heliyon*, Volume 10, Issue 3, e25112.

Juwono, F.H. *et al.* (2023) 'Machine learning for weed–plant discrimination in agriculture 5.0: An in-depth review', *Artificial Intelligence in Agriculture*, 10, pp. 13–25. doi:10.1016/j.aiia.2023.09.002.

Meshram, Vishal *et al.* (2021) 'Machine learning in agriculture domain: A state-of-art survey', *Artificial Intelligence in the Life Sciences*, 1, p. 100010. doi:10.1016/j.ailsci.2021.100010.

Moreira, B.R. *et al*. (2024) 'Integrating machine learning methods for computing greenhouse gas emissions baselines in agriculture', *Journal of Cleaner Production*, 485, p. 144416. doi:10.1016/j.jclepro.2024.144416.

Bai, J. *et al.* (2025) 'Assessing biochar's impact on greenhouse gas emissions, microbial biomass, and enzyme activities in agricultural soils through meta-analysis and machine learning', *Science of The Total Environment*, 963, p. 178541. doi:10.1016/j.scitotenv.2025.178541.

Hai, A. *et al.* (2023) 'Machine learning models for the prediction of total yield and specific surface area of biochar derived from agricultural biomass by pyrolysis', *Environmental Technology &amp; Innovation*, 30, p. 103071. doi:10.1016/j.eti.2023.103071.

Islam, Taufiqul et al. (2024) 'A comparative study of machine learning models for predicting Aman rice yields in Bangladesh' *Heliyon*, Volume 10, Issue 23, e40764.

Chen, X. *et al.* (2024) 'Infrared microspectroscopy and Machine Learning: A novel approach to determine the origin and variety of individual rice grains', *Agriculture Communications*, 2(2), p. 100038. doi:10.1016/j.agrcom.2024.100038.

Phan, Thi-Thu-Hong (2024) 'A novel method for identifying rice seed purity using hybrid machine learning algorithms', *Heliyon*, Volume 10, Issue 14, e33941.

Dadashzadeh, M. *et al.* (2024) 'A stereoscopic video computer vision system for weed discrimination in rice field under both natural and controlled light conditions by machine learning', *Measurement*, 237, p. 115072. doi:10.1016/j.measurement.2024.115072.

Zhai, C. *et al.* (2024) 'Rapid classification of rice according to storage duration via near-infrared spectroscopy and machine learning', *Talanta Open*, 10, p. 100343. doi:10.1016/j.talo.2024.100343.

Sahoo, S. *et al.* (2024) 'Advanced prediction of rice yield gaps under climate uncertainty using machine learning techniques in eastern India', *Journal of Agriculture and Food Research*, 18, p. 101424. doi:10.1016/j.jafr.2024.101424.

Alrowaily, M.A. *et al.* (2024) 'Application of extreme machine learning for smart agricultural robots to reduce manoeuvring adaptability errors', *Alexandria Engineering Journal*, 109, pp. 655–668. doi:10.1016/j.aej.2024.09.062.

Padhiary, M. *et al.* (2024) 'Enhancing precision agriculture: A comprehensive review of machine learning and AI vision applications in all-terrain vehicle for farm automation', *Smart Agricultural Technology*, 8, p. 100483. doi:10.1016/j.atech.2024.100483.

Lynda, D. *et al.* (2023) 'Towards a semantic structure for classifying IOT agriculture sensor datasets: An approach based on machine learning and web semantic technologies', *Journal of King Saud University - Computer and Information Sciences*, 35(8), p. 101700. doi:10.1016/j.jksuci.2023.101700.

Boisramé, G.F.S. et al. (2023) 'Exploring climate-driven agricultural water shortages in a snow-fed basin using a water allocation model and machine learning', *Journal of Hydrology*, 621, p. 129605. doi:10.1016/j.jhydrol.2023.129605.

Elsayed, A. *et al.* (2024) 'Machine learning models for prediction of nutrient concentrations in surface water in an agricultural watershed', *Journal of Environmental Management*, 372, p. 123305. doi:10.1016/j.jenvman.2024.123305.

Elbeltagi, A. *et al.* (2023) ‘Forecasting vapor pressure deficit for agricultural water management using machine learning in semi-arid environments’, *Agricultural Water Management*, 283, p. 108302. doi:10.1016/j.agwat.2023.108302.

Wan, L. *et al.* (2024) ‘Mapping agricultural tile drainage in the US Midwest using explainable random forest machine learning and satellite imagery’, *Science of The Total Environment*, 950, p. 175283. doi:10.1016/j.scitotenv.2024.175283.

Chen, H. *et al.* (2025) ‘A novel agricultural drought index based on multi-source Remote Sensing Data and interpretable machine learning’, *Agricultural Water Management*, 308, p. 109303. doi:10.1016/j.agwat.2025.109303.

Wang, X. *et al.* (2025) ‘Agricultural GDP exposure to drought and its machine learning-based prediction in the Jialing River Basin, China’, *Agricultural Water Management*, 307, p. 109265. doi:10.1016/j.agwat.2024.109265.

Khan, F. *et al.* (2024) ‘Assessing the impacts of temperature extremes on agriculture yield and projecting future extremes using machine learning and deep learning approaches with CMIP6 data’, *International Journal of Applied Earth Observation and Geoinformation*, 132, p. 104071. doi:10.1016/j.jag.2024.104071.

Limb, B.J. *et al.* (2024) ‘Estimating geographic origins of corn and soybean biomass for biofuel production: A detailed dataset’, *Data in Brief*, 54, p. 110291. doi:10.1016/j.dib.2024.110291.

Zanetta, C.U. *et al.* (2015) ‘Oil content and potential region for cultivation black soybean in Java as biofuel alternative’, *Energy Procedia*, 65, pp. 29–35. doi:10.1016/j.egypro.2015.01.025.

Vedovatto, F. *et al.* (2021) ‘Production of biofuels from soybean straw and hull hydrolysates obtained by subcritical water hydrolysis’, *Bioresource Technology*, 328, p. 124837. doi:10.1016/j.biortech.2021.124837.

Venkataraju, A. *et al.* (2023) ‘A review of machine learning techniques for identifying weeds in corn’, *Smart Agricultural Technology*, 3, p. 100102. doi:10.1016/j.atech.2022.100102.

Vahidi, M. *et al*. (2025) ‘Depth-specific soil moisture estimation in vegetated corn fields using a canopy-informed model: A fusion of RGB-thermal drone data and machine learning’, *Agricultural Water Management*, 307, p. 109213. doi:10.1016/j.agwat.2024.109213.

Da, H. *et al.* (2025) ‘Advancing soybean biomass estimation through multi-source UAV data fusion and machine learning algorithms’, *Smart Agricultural Technology*, 10, p. 100778. doi:10.1016/j.atech.2025.100778.

Skobalski, J. *et al.* (2024) ‘Bridging the gap between crop breeding and Geoai: Soybean yield prediction from multispectral UAV images with transfer learning’, *ISPRS Journal of Photogrammetry and Remote Sensing*, 210, pp. 260–281. doi:10.1016/j.isprsjprs.2024.03.015.

Tesfaye, A.A. *et al*. (2021) ‘Combining machine learning, space-time cloud restoration and phenology for farm-level wheat yield prediction’, *Artificial Intelligence in Agriculture*, 5, pp. 208–222. doi:10.1016/j.aiia.2021.10.002.

Chergui, N. (2022) ‘Durum wheat yield forecasting using machine learning’, *Artificial Intelligence in Agriculture*, 6, pp. 156–166. doi:10.1016/j.aiia.2022.09.003.

Mahmoodi, S. *et al.* (2024) ‘Modeling spatiotemporal distribution of yellow rust wheat pathogen using machine learning algorithms: Insights from Environmental Assessment’, *Environmental Technology &amp; Innovation*, 36, p. 103865. doi:10.1016/j.eti.2024.103865.

Agarwal, D. *et al*. (2023) ‘Machine Learning Approach for the classification of Wheat Grains’, *Smart Agricultural Technology*, 3, p. 100136. doi:10.1016/j.atech.2022.100136.

Hai, A. *et al.* (2023) ‘Machine learning models for the prediction of total yield and specific surface area of biochar derived from agricultural biomass by pyrolysis’, *Environmental Technology &amp; Innovation*, 30, p. 103071. doi:10.1016/j.eti.2023.103071.

Nasteski, V. (2017) ‘An overview of the supervised machine learning methods’, *HORIZONS.B*, 4, pp. 51–62. doi:10.20544/horizons.b.04.1.17.p05.

Maddodi, S. *et al*. (2024), “Market resilience in turbulent times: a proactive approach to predicting stock market responses during geopolitical tensions,” J. Cap. Mark. Stud., doi: 10.1108/JCMS-12-2023-0049.

Balan, G.S. *et al.* (2025) ‘Machine learning and artificial intelligence methods and applications for post-crisis supply chain resiliency and Recovery’, *Supply Chain Analytics*, 10, p. 100121. doi:10.1016/j.sca.2025.100121.

Mihaly Cozmuta, L. (2025) ‘The application of multiple linear regression methods to FTIR spectra of fingernails for predicting gender and age of human subjects’, *Heliyon*, Volume 11, Issue 4, e42815.

Masteali, S.H. *et al.* (2025) ‘Uncertainty analysis of linear and non-linear regression models in the modeling of water quality in the Caspian Sea Basin: Application of Monte-Carlo method’, *Ecological Indicators*, 170, p. 112979. doi:10.1016/j.ecolind.2024.112979.

Hamidi, S.K. *et al.* (2022) ‘Projected Biodiversity in the hyrcanian mountain forest of Iran: An investigation based on two climate scenarios’, *Biodiversity and Conservation*, 32(12), pp. 3791–3808. doi:10.1007/s10531-022-02470-1.

Bagriacik, M. and Otero, F.E.B. (2024) ‘Multiple fairness criteria in decision tree learning’, *Applied Soft Computing*, 167, p. 112313. doi:10.1016/j.asoc.2024.112313.

Itzkin, M. *et al.* (2025) ‘Developing a decision tree model to forecast runup and assess uncertainty in empirical formulations’, *Coastal Engineering*, 195, p. 104641. doi:10.1016/j.coastaleng.2024.104641.

Lagzi, M.D. *et al*. (2024) ‘A hybrid stochastic data envelopment analysis and decision tree for performance prediction in retail industry’, *Journal of Retailing and Consumer Services*, 80, p. 103908. doi:10.1016/j.jretconser.2024.103908.

Breiman, L. (2001) ‘Random Forests.’, *Machine Learning*, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

Suárez-Fernández, G.E. *et al*. (2025) ‘Enhancing carbon stock estimation in forests: Integrating multi-data predictors with Random Forest method’, *Ecological Informatics*, 86, p. 102997. doi:10.1016/j.ecoinf.2025.102997.

Sharma, K.P. *et al.* (2025) ‘Quantum behaved binary gravitational search algorithm with random forest for Twitter spammer detection’, *Results in Engineering*, 25, p. 103993. doi:10.1016/j.rineng.2025.103993.

Asamoah, Eric et al. (2024) ‘Random Forest machine learning for maize yield and agronomic efficiency prediction in Ghana’, *Heliyon*, Volume 10, Issue 17, e37065.

Pimentel, J.S. *et al.* (2024) ‘A novel fusion support vector machine integrating weak and sphere models for classification challenges with Massive Data’, *Decision Analytics Journal*, 11, p. 100457. doi:10.1016/j.dajour.2024.100457.

Maggioni, F. and Spinelli, A. (2025) ‘A novel robust optimization model for nonlinear support vector machine’, *European Journal of Operational Research*, 322(1), pp. 237–253. doi:10.1016/j.ejor.2024.12.014.

Raghunath, M.P. *et al.* (2025) ‘PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things’, *Measurement: Sensors*, 37, p. 101806. doi:10.1016/j.measen.2024.101806.

Roy, A. *et al.* (2024) ‘Seismic reliability analysis of nonlinear structures by active learning-based adaptive sparse Bayesian regressions’, *International Journal of Non-Linear Mechanics*, 165, p. 104817. doi:10.1016/j.ijnonlinmec.2024.104817.

Sevilla-Salcedo, C. *et al.* (2024) ‘Bayesian learning of feature spaces for multitask regression’, *Neural Networks*, 179, p. 106619. doi:10.1016/j.neunet.2024.106619.

Manfredi, P. (2025) ‘A hybrid polynomial chaos expansion – gaussian process regression method for Bayesian uncertainty quantification and sensitivity analysis’, *Computer Methods in Applied Mechanics and Engineering*, 436, p. 117693. doi:10.1016/j.cma.2024.117693.

Gbashi, S.M. *et al.* (2025) ‘Optimal feature selection for a weighted K-nearest neighbours for compound fault classification in wind turbine gearbox’, *Results in Engineering*, 25, p. 103791. doi:10.1016/j.rineng.2024.103791.

Chen, Y. *et al.* (2024) ‘K-means clustering method based on nearest-neighbour density matrix for customer electricity behaviour analysis’, *International Journal of Electrical Power &amp; Energy Systems*, 161, p. 110165. doi:10.1016/j.ijepes.2024.110165.

Badawi, M.A. *et al.* (2025) ‘Artificial Neural Networks analysis for non-Newtonian nanofluid flow with variable viscosity and MHD effects in wire covering processes’, *Results in Engineering*, 25, p. 103878. doi:10.1016/j.rineng.2024.103878.

Karakoyun, M. (2024) ‘Artificial Neural Network training using a multi selection artificial algae algorithm’, *Engineering Science and Technology, an International Journal*, 53, p. 101684. doi:10.1016/j.jestch.2024.101684.

11. APPENDIX

For direct access to the dataset and code being used in this project, please visit the following GitHub repository: https://github.com/MarcoMalchiodi/DataAnalyticsProject

11.1 Table of Definitions

ML	Machine Learning
DL	Deep Learning
DNN	Deep Neural Network
ANN	Artificial Neural Network
RF	Random Forest
RFR	Random Forest Regressor
KNN	K-Nearest Neighbors
MAE	Mean Absolute Error
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
RScv	Randomized Search Cross-Validation
ELM	Extreme Learning Machine
CNN	Convolutional Neural Network
GAM	Generalized Additive Model
SVR	Support Vector Regression
RMSE	Root Mean Square Error