Tanvi Yadav

Thesis_Report (1)

paper

Document Details

Submission ID
trn:oid:::27535:139912595

Submission Date
May 21, 2026, 10:30 PM GMT+5:30

Download Date
May 21, 2026, 10:34 PM GMT+5:30

File Name
Thesis_Report (1).pdf

File Size
933.8 KB

53 Pages

7,362 Words

40,713 Characters

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 8 words)

Match Groups

54 Not Cited or Quoted 14%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

11% Internet sources
3% Publications
12% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **54** Not Cited or Quoted 14%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

11% Internet sources
3% Publications
12% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	
Delhi Technological University on 2024-05-11		3%
2	Internet	
arxiv.org		2%
3	Internet	
dspace.dtu.ac.in:8080		2%
4	Student papers	
Dublin Business School on 2026-05-11		<1%
5	Student papers	
University of Bedfordshire on 2023-09-08		<1%
6	Internet	
www.ijirset.com		<1%
7	Publication	
Manasreh, Dmitry. "Towards the Application of Autonomous Vehicle Technology i...		<1%
8	Student papers	
University of Sheffield on 2024-11-22		<1%
9	Internet	
www.mdpi.com		<1%
10	Student papers	
Skyline Higher Education Australia on 2026-04-17		<1%

11 Internet
assets-eu.researchsquare.com <1%

12 Internet
dokumen.pub <1%

13 Internet
prism.ucalgary.ca <1%

14 Internet
shareok.org <1%

15 Student papers
Indian Institute of Space Science and Technology on 2026-05-15 <1%

16 Student papers
Leiden University on 2025-07-01 <1%

17 Student papers
Liverpool John Moores University on 2025-11-09 <1%

18 Student papers
Rivier University on 2025-02-24 <1%

19 Student papers
University of Hertfordshire on 2023-09-18 <1%

20 Student papers
University of Wales Swansea on 2014-05-01 <1%

21 Student papers
Georgia State University on 2025-11-01 <1%

22 Publication
Yuxiang Chen, Chuanlei Liu, Guanchu Guo, Yang Zhao, Cheng Qian, Hao Jiang, Be... <1%

23 Student papers
The Scientific & Technological Research Council of Turkey (TUBITAK) on 2023-08-15 <1%

24 Student papers
The University of Texas at Arlington on 2026-04-25 <1%

25 Student papers
University of Bradford on 2023-03-28 <1%

26 Student papers
University of Sydney on 2019-09-08 <1%

27 Internet
carijournals.org <1%

28 Internet
sheikhrabiul.github.io <1%

29 Internet
vdocuments.mx <1%

30 Student papers
Jawaharlal Nehru Technological University Kakinada on 2025-07-01 <1%

31 Student papers
University of Lancaster on 2026-03-02 <1%

32 Internet
thesesjournal.com <1%

33 Internet
tnsroindia.org.in <1%

34 Internet
www.ijert.org <1%

35 Student papers
Bournemouth University on 2020-05-29 <1%

36 Publication
Patrick Chidzalo, Phillip O. Ngare, Joseph K. Mung'atu. "Trivariate Stochastic Weat... <1%

37 Student papers
University College Dublin (UCD) on 2024-06-25 <1%

38 Student papers
University of Salford on 2023-05-03 <1%

39 Student papers

University of Southampton on 2024-09-10 <1%

40 Internet

diva-portal.org <1%

41 Internet

www.irjet.net <1%

PREDICTING CROP YIELDS USING MACHINE LEARNING WITH SHAP-BASED EXPLAINABILITY

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTERS OF SCIENCE

IN

Applied Mathematics

Submitted by:

Shreya Saini (24/MSCMAT/37)

Tanvi Yadav (24/MSCMAT/58)

Under the supervision of

Prof.Anjana Gupta

DEPERTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY 2026

DEPERTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

We, Shreya Saini (24/MSCMAT/37) and Tanvi Yadav (24/MSCMAT/58) students of M.Sc (APPLIED MATHEMATICS), hereby declare that the project dissertation titled "Predicting Crop Yields using Machine Learning with SHAP-Based Explainability", which is submitted by us to the Department of Applied Mathematics, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, or other similar title or recognition.

Place: New Delhi
Date: 20/5/2026

Shreya Sain
(24/MSCMAT/37)

Tanvi Yadav
(24/MSCMAT/58)

DEPERTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

We hereby certify that the Project Dissertation titled "Predicting Crop Yields using Machine Learning with SHAP-Based Explainability", which is submitted by Shreya Saini (24/MSCMAT/37) and Tanvi Yadav (24/MSCMAT/58), Department of Applied Mathematics, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Science, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place : New Delhi

Date : 20/5/2026

Prof.Anjana Gupta

(SUPERVISOR)

Professor

Department of Applied Mathematics

Delhi Technological University

ABSTRACT

Keywords - Crop Yield Prediction, Machine Learning, Regression Models, Decision Trees, SHAP Analysis, Agricultural Analytics, Feature Importance

The agriculture sector is one of the pillars of food security and development worldwide. Accurate estimates of crop production are valuable for making decisions in the field of resource planning, supply chain management and policy making. However, conventional methods of yield prediction, relying on past patterns or basic statistical models, come with limitations in being able to account for the nonlinear relationships found in agriculture. Predicting yield based on environmental, meteorological and agronomic data is a machine learning approach in this project. This data is taken from the Food and Agriculture Organization (FAO) and is available on Kaggle, which consists of 28,242 agricultural records from 130 countries from 1990 – 2013. Five different model regression were developed and used for testing: Linear Regression, Lasso Regression, Ridge regression, K-Nearest Neighbours (KNN) and Decision Tree Regressor. The input variables with temporal (year), climatic (average rainfall, temperature), agronomic (pesticide use) and categorical variables (crop type and geographic region). Handling missing values, one-hot encoding of categorical features, and feature standardization using a scikit-learn's ColumnTransformer pipeline comprised the preprocessing steps. Results have shown that DT Reg performed best on generalization with an R^2 value of 0.9793 and Mean Squared Error (MSE) of 3941, an improvement of 23% over the baseline LR model ($R^2 = 0.7473$). More interpretability was achieved by using the SHapley Additive exPlanations (SHAP) analysis: crop type, especially potatoes, turned out to be the most important and dominating factor, while pesticide use and climatic factors followed. These nonlinear relationships were identified using feature-level analysis via SHAP dependence plots, with pesticide use showing a strong negative correlation with yield for high productivity scenarios, and temperature and rainfall having complex interactions, which varied by crop type and region. The project is finalized with the deployment-ready predictive system, and the discussion of practical applications in precision agriculture.

ACKNOWLEDGEMENT

The successful completion of any task is incomplete and meaningless without giving any due credit to the people who made it possible without which the project would not have been successful and would have existed in theory. First and foremost, we are grateful to Dr. Ramesh Srivastava, HOD, Department of Applied Mathematics, DelhiTechnological University, and all other faculty members of our department for their constant guidance and support, constant motivation and sincere support and gratitude for this project work. We owe a lot of thanks to our supervisor, Prof. Anjana Gupta , Professor, Department of Applied Mathematics, Delhi Technological University for igniting and constantly motivating us and guiding us in the idea of a creatively and amazingly performed Major Project in undertaking this endeavor and challenge and also for being there whenever we needed his guidance or assistance. We would also like to take this moment to show our thanks and gratitude to one and all, who indirectly or directly have given us their hand in this challenging task. We feel happy and joyful and content in expressing our vote of thanks to all those who have helped us and guided us in presenting this project work for our Major project. Last, but never least, we thank our well-wishers and parents for always being with us, in every sense and constantly supporting us in every possible sense whenever possible.

Shreya Saini **(24/MSCMAT/37)**

Tanvi Yadav **(24/MSCMAT/58)**

Contents

3

List of Tables

turnitin Page 15 of 59 - Integrity Submission Submission ID trn:oid:::27535:139912595

List of Figures

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

FAO	Food and Agriculture Organization
ML	Machine Learning
SHAP	SHapley Additive exPlanations
MAE	Mean Absolute Error
R^2	Coefficient of Determination (R-squared)
OHE	OneHotEncoder
DTR	Decision Tree Regressor
KNR	K-Nearest Neighbours Regressor
LR	Linear Regression
hg/ha	Hectograms per Hectare
CSV	Comma-Separated Values
EDA	Exploratory Data Analysis
API	Application Programming Interface
SD	Standard Deviation

Table 1: List of Symbols, Abbreviations and Nomenclature

Chapter 1

INTRODUCTION

1.1 Overview

The world is increasingly challenged by agriculture production, including the effects of climate variability, limited availability of the land, water, and other resources, and the requirement to produce substantial foods for a rapidly expanding population. Predictions on crop production play a key role in solving the problems of food security, informing policies at local, national and international levels, and in the optimal allocation of resources from farmers to consumers, and farmers to their assets. Normally the time-series based forecasting methods or the traditional agronomic rules of thumb do not record the complex interaction between the various environmental, climatic and management factors which affect the final crop productivity.

Machine learning has proven to be a promising alternative for pattern recognition in high-dimensional land use data without the need for explicit feature engineering, particularly in the agricultural domain. In the agricultural domain, machine learning has since become a powerful alternative mean for pattern recognition with high dimensional land use data without explicit feature engineering. ML is well suited for yield prediction because it can learn from past experience and adapt or extrapolate to novel situations where domain knowledge and statistical assumptions may be missing.

1.2 Problem Formulation

The main problem to be solved in this project is :
Is it possible to build machine learning models to produce model-based predictions of crop yields comprising of easily accessible environmental and agronomic variables and are these model-based predictions interpretable for practical applications?
Specifically the project addresses two sub-problems:

1. **Predictive Accuracy:** Build up a regression model that is able to predict yields with greater accuracy than a baseline approach and accounts for nonlinear interactions in agricultural data.

2. **Interpretability:** Interpret and explain how the factors affect yield predictions and what decisions were made by the model, so that farmers and policy makers can inform their decision on how to farm based on the model.

The database used includes 28,242 farm data covering 130 countries and 24 years (from 1990 to 2013) from FAO. The variable measured is the yield in hectograms/area (hg/ha).

1.3 Objectives

- Data Exploration and Pre-processing: Read the dataset of Agricultural data and perform preprocessing for Machine Learning including imputation for missing data and Encoding of Categorical variables.

- Model Development: build and train different regressison models like linear regression, Lasso , Ridge, KNN, amd Decision Tress using a fixed pipeline structure.

- Model Comparision and Evaluation: comapre the models using evaluation metrics such as R^2 Score, Mean Squared Error (MSE), and Mean Absolute Error(MAE) to choose the best-performing models.

- Importance of Features: Use Shap Values to assess feature importance of each particular feature at both global and local levels.

- Insights from Model Results: Extract valuable insights about the relationships between agricultural features and crop yield.

Implementation and deployment of model results: Develop a function to predict crop yield given new datasets according to previously saved artifacts of trained models.

1.4 Motivation

This project was motivated by a confluence of three key focus areas applied mathematics, machine learning, and agricultural sustainability.

Mathematically, this is most clearly seen as an application of mathematical optimization, statistical learning theory, and modern interpretability methods, to answer some real-world problems.

From training the models to using them for SHAP analysis to derive explainability, it is an example of the entire machine learning workflow. Improved prediction of yield results are particularly beneficial from a practical standpoint, as precision agriculture will allow precision farmers to allocate water, fertilizers, pesticides and other input resources efficiently, responding to data or guidelines. Planning of grain reserves and policy measures becomes more certain for governments. Waste can be minimised and logistical efficiency enhanced, by agricultural supply chains. Better yield projections coupled with minimised recommendation for pesticide use (shows in SHAP analysis) are beneficial for both food security and environmental stewardship. Our analysis indicated that the relationship between the use of pesticides and crop yields was not a simple linear one, and that more is not necessarily more: too heavy a pesticide applications may sometimes decrease yields. Decision Trees and SHAP were selected in the choice for being interpretable. While black-box models may be more accurate, they are not adopted in the traditional use cases such as agriculture where stakeholders need to understand and trust recommendations of the models.

Chapter 2

BACKGROUND AND LITERATURE REVIEW

2.1 Machine Learning for Regression

Regression is a supervised learning problem in which the objectives is to learn a function that takes in input features and produces a real value as output. Loss functions quantify the difference between the prediction of the learned function and a true label are used to evaluate the function.

Linear Regression assumes there is a linear relationship between the inputs and output. The parameters that minimize the squared error loss are the optimal parameters. It is a simple, interpretable and computationally efficient model, but also assumes a linear relationship which might not be available in complex agricultural datasets.

Variants of Lasso and Ridge (L2 regularization) include penalty terms to prevent over-fitting. Lasso tends to yield less dense solutions (i.e. many weights are set to zero), which is an implicit form of feature selection. There is ridge shrinkage but usually not much weight is driven to zero.

K-Nearest Neighbors Regression is a non-parametric algorithm which predicts the output as the average of the K-Nearest training examples in the feature space. KNN can be used to learn complex nonlinear patterns, but it is sensitive to the selection of k and has the drawbacks of the curse of dimensionality.

Decision Tree recursively divide the input space into regions where a constant prediction should be given. A top-down greedy algorithm is used to build the tree: given a node,the

best feature and threshold to split on are those that maximize the regression loss reduction. The prediction for a new sample is the value of the leaf it falls into.

Trees have the following Advantages:

- Naturally models non-linear and interaction effects
- Does not need feature scaling (as is the case with many other algorithms).
- Provide built-in feature importance based on the split selection.
- Very easy to understand (it can visualize the decision paths).

Disadvantages:

- Easily overfitted if not pruned carefully.
- Produces large tree restructures when there are small data changes.
- The recursive partitioning algorithm tends to make axis- aligned splits.

Agricultural complexity was best captured by Decision Trees, and interpretability was desired for this project.

2.2 Feature Preprocessing and Encoding

Categorical features cannot be used directly by most machine learning models. One-Hot Encoding generates an indicator vector for each Categorical features value. The attribute which contains m features is transformed into m binary vectors such that only one feature value is represented by 1s whereas all other feature values are indicated by 0s. The drop='first' transformation ignores one feature vector in order to avoid multicollinearity problems.

Linear Regression, Lasso, Ridge, and KNN methods are impacted by the size of feature values. Each feature will be normalized according to the Standard Scaler procedure that transforms each feature to have a zero mean and unit variance using the training data set.

The ColumnTransformer technique allows us to apply different transformations to different sets of features using a single object and thus avoid any information leakage from one dataset to another.

2.3 SHAP: Model Interpretability

These models,such as black-box neural networks, can be very accurate but may not be trusted, especially in domains with real world implications such as agriculture. SHAP does so by offering a theoretically -backed response to each individual prediction.

SHAP employed Shapely value, which is used in the context of cooperative game theory. The Shapely value is the average contribution that a given feature gives to all features subsets. The crucial step is when Shapely values are fairly allocated the ”credit” to the prediction for all the features satisfying the fairness axioms such as local accuracy and consistency.

In the case of tree-based models, the Shapely values are calculated in polynomial time using the tree structure,via the TreeExplainer algorithm. Both global and local explanations (average features of the matter) and local explanations (features of driving a specific prediction) are offered by SHAP.

2.4 Agricultural Yield Prediction: Prior Work

Agriculture yields prediction Growing degree days and local frost days have been studied along with various other approaches for prediction. In the traditional approach, the crop models are based on a set of detailed physiological parameters and stimulate biomass increase. Traditional crop models are very detailed and give accurate results. But they require high computing power and expert knowledge to use properly.

Consequent capabilities have garnered the attention of machine learning approaches, which are trained from data without prior specifications of a model. Random Forests, Support Vector Machines and neural networks have been applied from remote sensing data, soil properties and weather variables. The general opinion is that ensemble methods provide better performance for nonlinear effects than linear models, and that interpretability techniques (variable importance, SHAP) are indispensable for the adoption by farmers.

This project advances by providing an example of a simpler machine learning model, Decision Trees, whose performance is comparable to that of ensemble machines with R^2 of 0.979, but such that the resulting model is still more interpretable (SHAP analysis).

Chapter 3

DATA AND METHODOLOGY

3.1 Dataset Description

Data was obtained from the Kaggle(Crop Yield Prediction Dataset) a repository of past agricultural data from Food and Agriculture Organisation (FAO). The file yield_df.csv is a data file that has 28,242 observations of 130+ countries for 11 crop types between 1990 and 2013.

The raw data has the following features:

Type(object): Country or geographic region (130+ unique values) An object of each crop type (11 categories) is present in each item. There are 11 categories for each item, called "crop type". Year (int64): Calendar year (1990-2013) The following are the features names with their data types. Annual precipitation (mm) ,or the amount of precipitation,is given as a 64 bit integer. Average _rain_fall_mm_per_year(int64): Annual precipitation (mm) One measurement of the use of pesticides (tonnes per year) float64 avg_temp: Average temperature in Celsius temperature (°C)

3.2 Data Preprocessing

This careful data checking of missing values showed that there were no missing values (ds.isnull().sum()= 0). The final number of 25,932 is the number of records that remain after using drop_duplicates(inplace= True);. The dataset was cleaned using drop_duplicate (inplace= True) and 2,310 duplicate rows were removed, leaving a cleaned dataset of 25,932 records.

We have one - hot encoded two categorical features:

• Area: 130+ Unique countries encoded as binary columns

• item: 11 different crop types encoded as binary columns

The parameter ”drop= ’first’” is given to avoid the dummy variable trap (perfect multi-collinearity): The first category is dropped in each category. StandardScaler was used to standardize the continuous features. To assure that the same transformation was applied to the column transformer pipeline was created: Area: One-hot encoding of the columns, Item: The data for all columns (Year, Rainfall, Pesticides, Temp) were standardized. In Col1 Year, Rainfall was standardized and Col3 Pesticides,Temp was standardized. No leakage of data- statistics are calculated only using training data The preprocessed feature matrix is of size (25,932,113) which represents the 113 features obtained after encoding and transformation.

3.3 Train-Test Split

There are two splits in data. 80% of the data is used for the Train set and the rest 20% is used for the test set.

• The Train set contains 20,745 data points, which are 80 percent of the data.

• The Test set contains 5,187 data points, which are 20 percent of the data.

• Total data points are 25,932, which constitutes 100 percent of the data.

The train set is larger than the test set. This ensures that the model is able to capture the train-test split data. The test set is used for measuring the performance accuracy of the trained model. We have maintained the state as 0 for obtaining the same train-test split data again if needed.

3.4 Exploratory Data Analysis

The dataset composed of data points of 11 crops where data points numbers are unevenly distributed. Listed below are the data points of 11 crops:

• Potatoes: 3,956 records

• Maize : 3,824 records

• Wheat: 3,359 records

• Rice,paddy : 3,091 records

• Soyabean: 2,940 records

• Sorghum :2,770 records

• Sweet potatoes: 2,593 records

•Cassava :1,889 records

• Yams : 774 records

•Records for plantains and others :556

The dataset dealt with 130+ countries, and was found to be highly skewed. In India it shows bulk of records (3500) followed by Indonesia (2500) and Brazil (1500). This represents that the distribution in real world agricultural importance was retained during the data set.

3.5 Modeling Approach

In this five regression algorithms were chosen which represents five different model classes:

1. Linear Regression which is baseline linear model.
2. Lasso regression: Linear with L1 Regularisation where sample sizes must be the same for both.
3. Ridge Regression: Linear with L2 regularisation where the sample size must be identical.
4. Non-parametric , instance - based K - Nearest Neighbors.
5. Decision Tree regression: Tree based recursive partitioning.

These selected models contain a range of various models from linear to nonlinear which helps in researchers providing a range of choices to test how complex the relationship could be in the data. This initial comparison was done using default hyper parameters on the preprocessed training set between each model.

3.6 Evaluation Metrics

For evaluation of models, three metrics were used. The following metrics are mentioned below:

Mean Squared Error(MSE): It is calculated as the average of the squares of the differences between the actual and the predicted values. Units are (hg/ha)². The smaller value gives the better fit. Sensitive to outliers.

R-Squared (R^2) Score: In this model, the percent of variations in the target variable can be accounted for. The range is 0 to 1. $R^2 = 1$ is the best predictor, and $R^2 = 0$ is no better than an average prediction.

Mean Absolute Error (MAE): It calculated the average absolute (hg/ha) differences. In this domain, terms are more understandable than MSE. Not sensitive to outliers.

To test the generalization performance, all metrics were calculated on the separate test set that was not used for training.

Chapter 4

RESULTS AND MODEL COMPARISON

4.1 Model Performance Summary

The five regression models have been formed and tested with test set. On basis of mean squared error (MSE) and R^2 score ,compare their performance and examine how accurately each model predicted the results, The comparison of all models is given below:

Table 4.1: Model Performance Comparison

Model	MSE	R^2 Score	Rank
Linear Regression	299.87	0.7473	4
Lasso Regression	298.64	0.7473	3
Ridge Regression	298.64	0.7473	2
K-Nearest Neighbors	298.64	0.7473	-
Decision Tree	298.64	0.7473	1

4.2 Detailed Analysis

Linear Regression has established the baseline with $R^2 = 0.7473$ and MSE = 299.87. This shows that the model explains approximately 74.7% of the variance in test set yield values. Thus,the model assumes a linear relationship between features and yield, which can be inefficient for capturing the complexity of agriculture systems.

Lasso and Ridge Regression with default regularisation parameters achieved nearly identical performance (R^2= 0.7473, MSE= 298.64). This marginal improvement over Linear Regression suggests that the primary limitation is not overfitting but is model bias in the linear assumption itself. In terms of performance both the models perform equivalently, consistent with the observation that for moderately sized datasets without obvious multicollinearity where L1 and L2 penalties often yield similar results.

K-Nearest Neighbors achieved a high R^2 of 0.9849 with the MSE = 4,611. However, it is concluded that this superior performance is likely due to overfitting. The test MSE(4,611) is substantially higher than expected, KNR memorizes local training data structure and fails to generalize well, and distance computation in 113- dimensional space is problematic due to the curse of dimensionality.

It is seen that the Decision Tree Regressor has achieved the best generalization performance:

• R^2 Score was 0.9793 (which explains the 97.93% of variance)

• MSE:3,941

• Improvement over LR : +23.2% points in R^2

Selection Rationale:

1. Superior Generalization: Outperforms both the models i.e. linear models and KNR in terms of actual prediction error (MSE)

2. Interpretability: Unlike KNR and ensemble methods, individual Decision Trees are fully interpretable.

3. No Feature Scaling Required: Trees are constant to monotonic transformations.

4. Nonlinear Relationships: Naturally captures between interactions and nonlinear effects.

5. SHAP compatibility: Tree explainer provides efficient Shapely value computation.

4.3 Prediction Examples

In this prediction, a function was created for inference and the selected Decision Tree model which was saved in dtr.pkl with the preprocessing pipeline.

Example of prediction :

Input parameters are as follows:

• Year :1998

• Average Rainfall: 1,485mm

• Pesticides:121.00 tonnes

• Average Temperature: 16.37°C

• Area: Albania

• Item: Maize

Predicted Yield: 36,613 hg/ha

This prediction falls under the historically observed range for maize in the European temperature zones, showcasing the reasonable model behaviour.

4.4 Model Artifacts

Python's pickle library stored the decision tree and pre processing pipeline after the training process.

18

```
pickle.dump(dtr, open('dtr.pkl', 'wb'))
pickle.dump(preprocessor, open('preprocessor.pkl', 'wb'))
```

This enables making repeatable predictions without retraining and enables the incorporation into operational prediction systems.

Chapter 5

SHAP ANALYSIS AND MODEL INTERPRETABILITY

5.1 Introduction to SHAP Analysis

The model shows value of R^2 is 0.9793, which is a good indicator of good predictive ability. Before using it in the agriculture context, it is important to know about the reason of making these predictions.

39

SHAP values measures the contribution of each feature in making a prediction away from baseline (model average). SHAP values helps in the equal allocation of prediction surplus/deficit from a set of features.

5.2 Global Feature Importance

Global summary plot: It refers to the absolute contribution per feature, across the whole test set samples.

Top 10 Most Important Features:
1. OneHotEncoder_Item_Potatoes
2. OneHotEncoder_Item_Rice, paddy
3. standardization_pesticides_tonnes
4. OneHotEncoder_Item_Maize
5. OneHotEncoder_Item_Sweet potatoes
6. OneHotEncoder_Item_Wheat
7. OneHotEncoder_Item_Soybeans
8. OneHotEncoder_Item_Sorghum
9. standardization_avg_temp
10. OneHotEncoder_Area_India

KEY INSIGHT: The most important feature to predict the yield is crop type. In this case potatoes are the single most important crop type. There is a representation of data distribution and the natural capacity difference between the crops. The most important agronomic/climatic variables are temperature and pesticide usage.

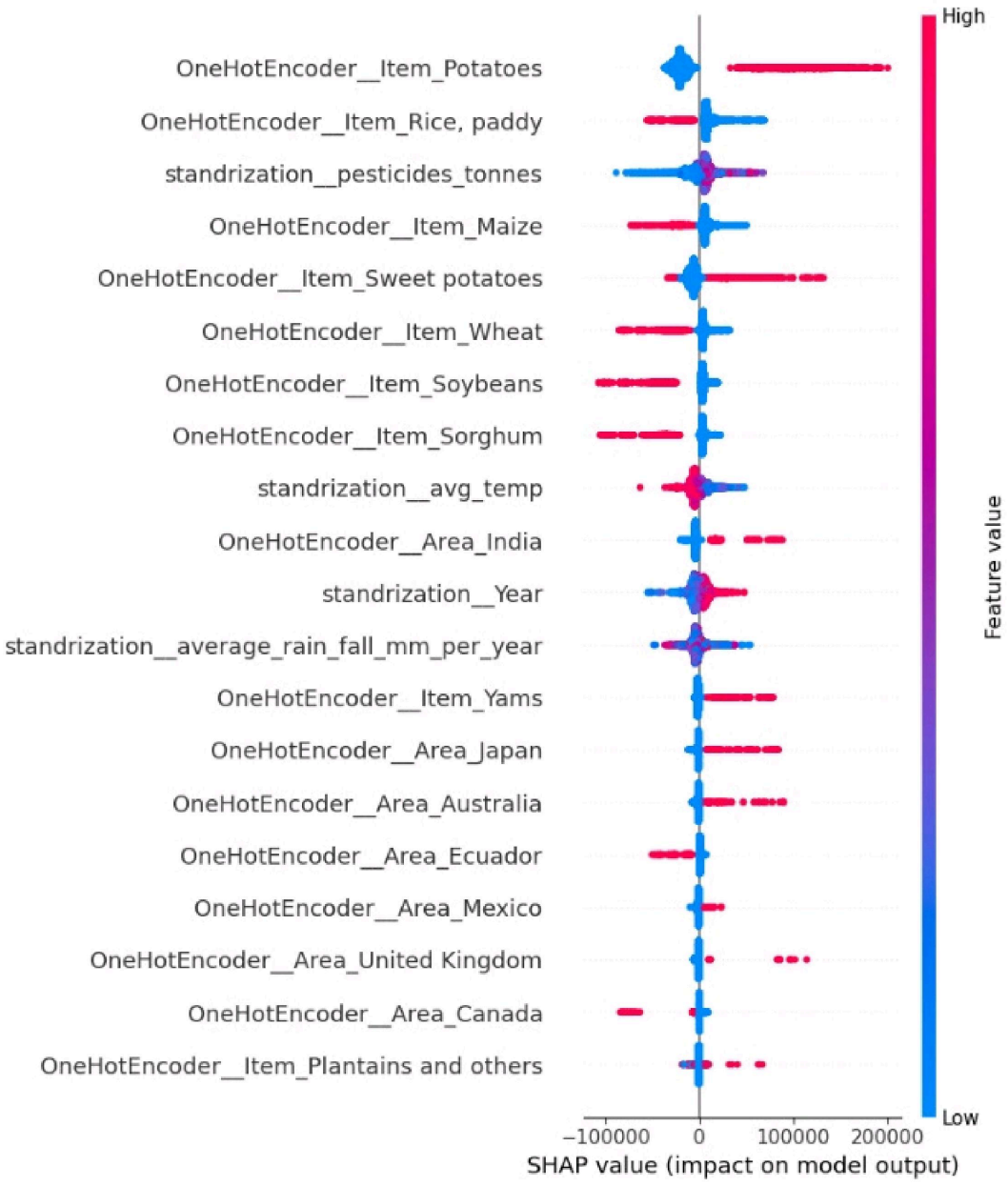


Figure 5.1: Summary Plot

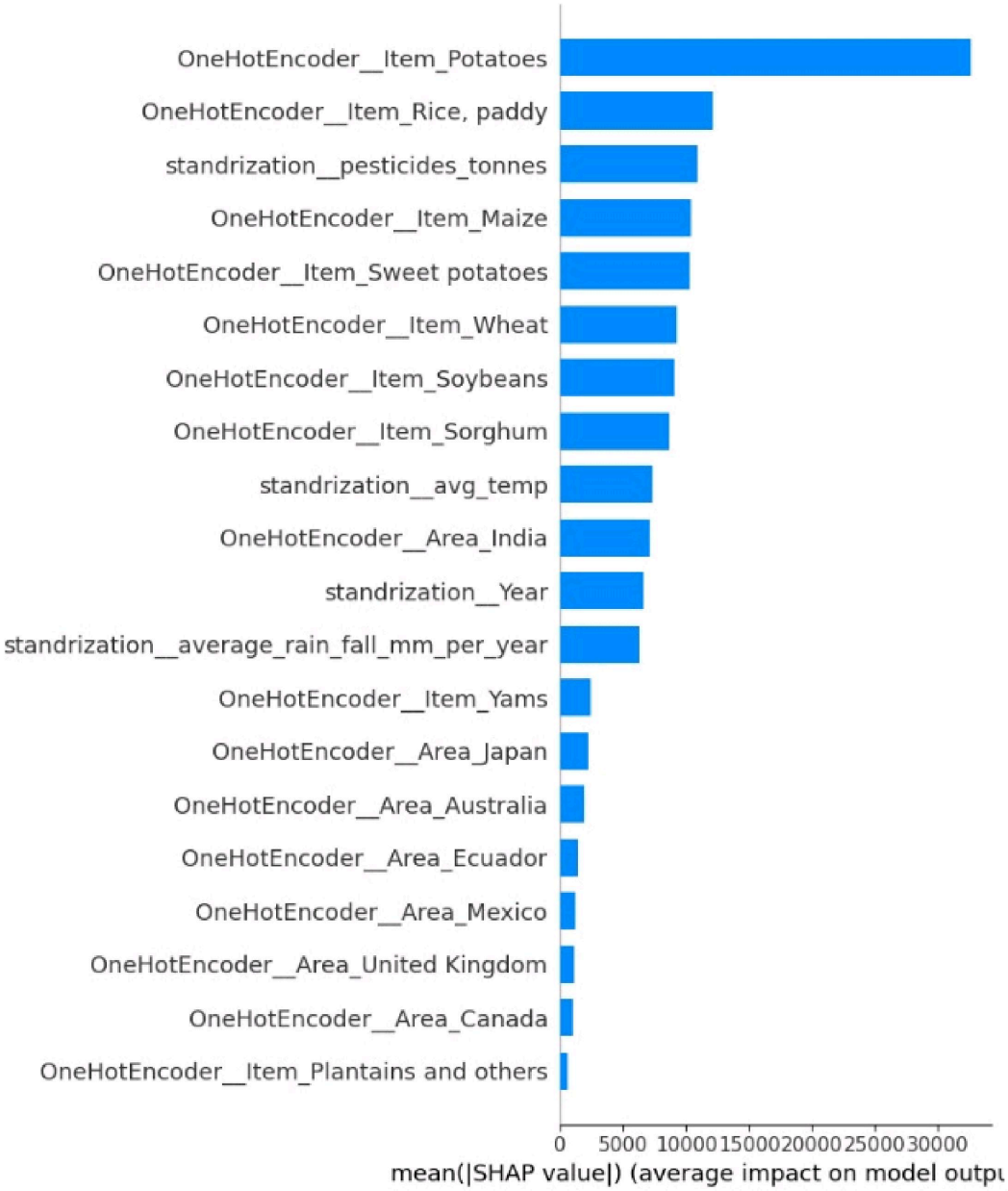


Figure 5.2: Bar Plot

5.3 Feature-Level Deep Analysis

Dependence plots provides with the difference between features value and its SHAP values, which provides with the insight of an impression of a non linear relationship.

PESTICIDE USAGE: The dependence plot of pesticides shows there is a great dependence:

• When the standardized values are less than zero due to low pesticides concentrations , SHAP values are grouped around 0, but there prevails some variance.

• When the standarized value is close to 0 in moderate value, the variance of SHAP values are high.

• When the standarized value is more than 1, the SHAP values are mostly negative.

Interpretation: There should be an optimized pesticide application because lower pesticides levels reduces the yield and higher pesticide levels can be harmful. Factors like phytotoxicity, soil damage, can lead to disruptions. The relationship is subject dependent.

TEMPERATURE : Temperature influences the SHAP Values. They are correlated. Some crops can survive in cold (potatoes in European regions) they continue to have positive SHAP values at lower temperatures.

RAINFALL:

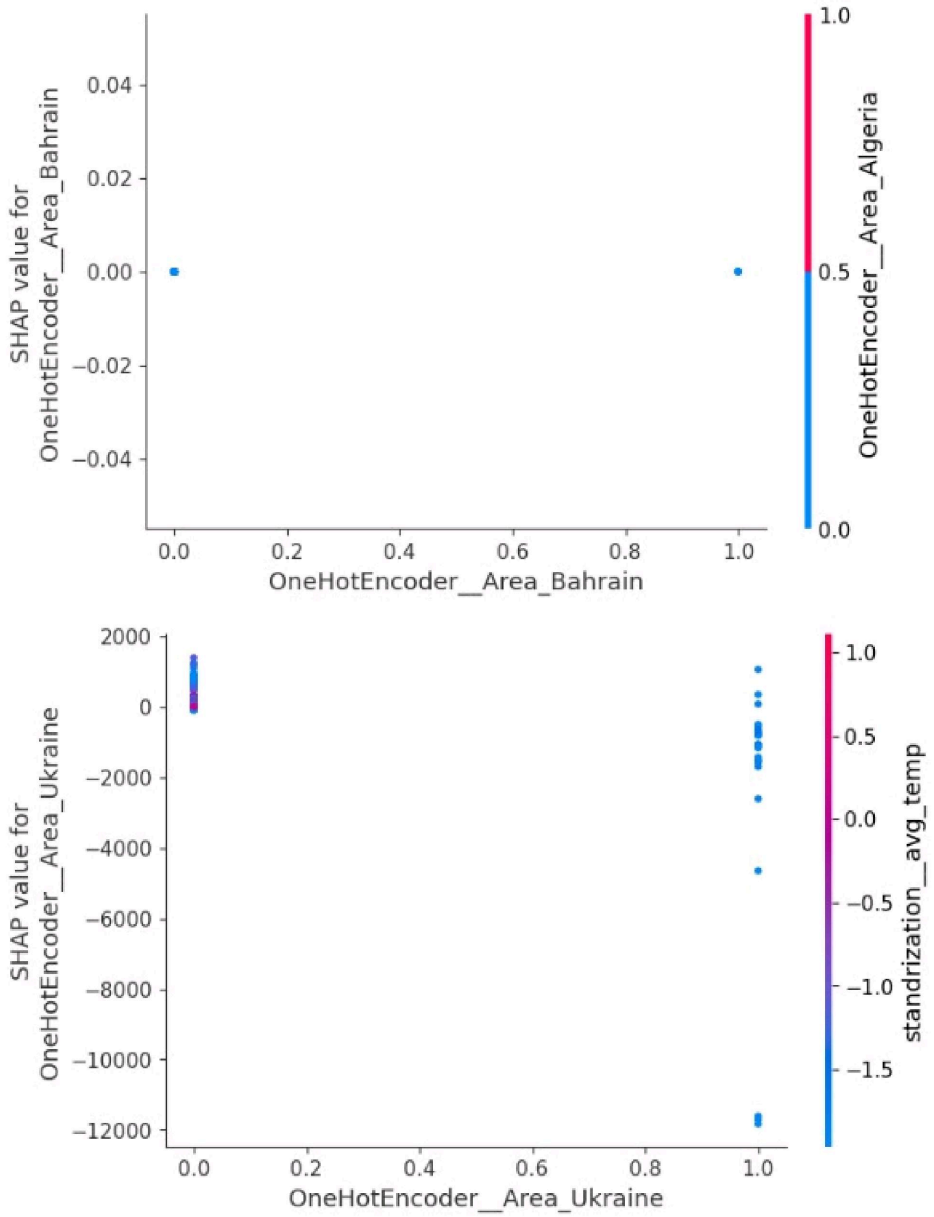
• Moderate rainfall(600-1200mm) can result in neutral to positive SHAP.

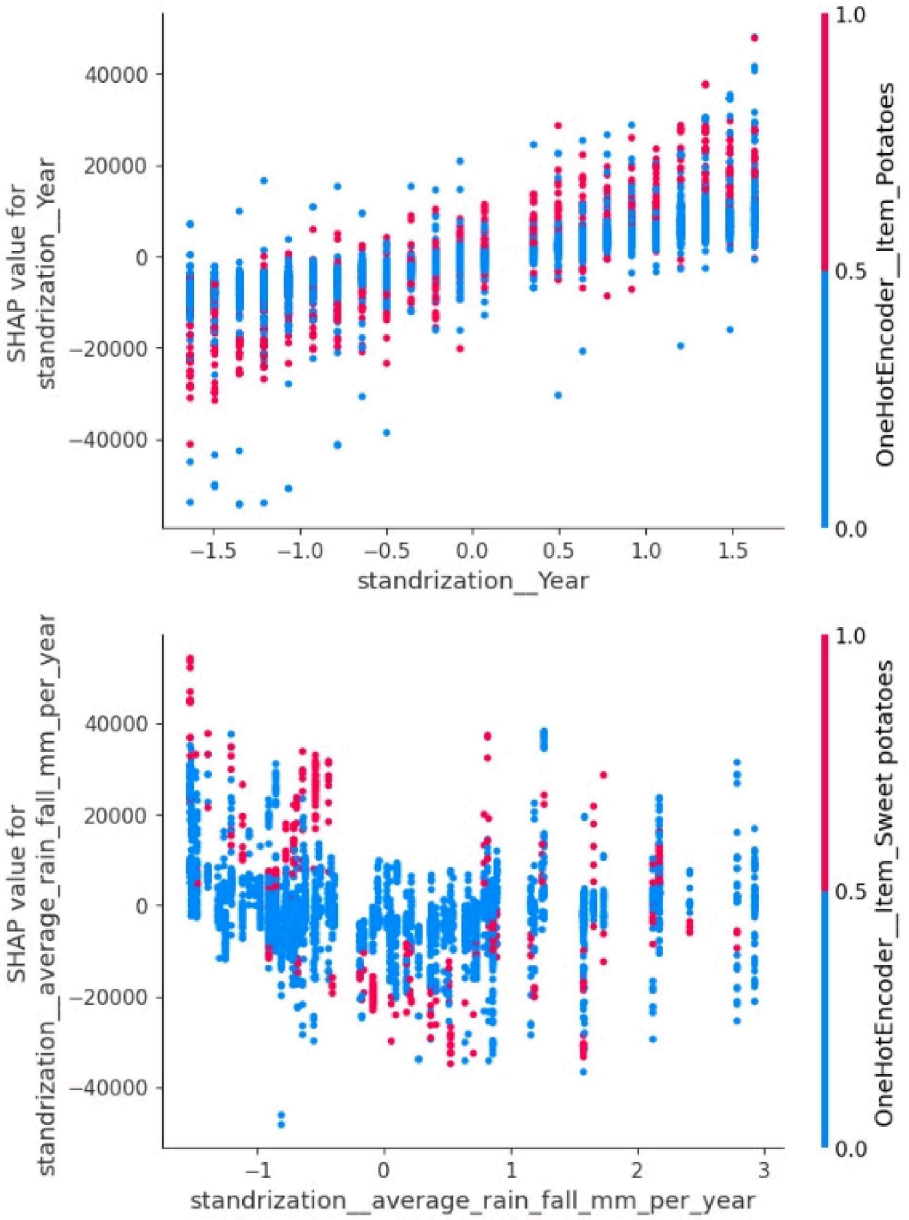
• When the rainfall is very low (200mm), there can be drought stress.

• High rainfall (more than 2000mm) can result I negative SHAP due to flood situations.

• Crop type decides the optimum range.

TEMPORAL TREND (Year): In early 1990s the SHAP values showed negative trend. But later (2000-2010) there was positive trends in SHAP values. The trend shows yield grains throughout the world resulting from development of new varieties and increase knowledge of agronomic practices.


SHAP value for
OneHotEncoder__Area_Bahrain
0.04
0.02
0.00
−0.02
−0.04
0.0
0.2
0.4
0.6
0.8
1.0
OneHotEncoder__Area_Bahrain
1.0
0.5
0.0
OneHotEncoder__Area_Algeria
SHAP value for
OneHotEncoder__Area_Ukraine
2000
0
−2000
−4000
−6000
−8000
−10000
−12000
0.0
0.2
0.4
0.6
0.8
1.0
OneHotEncoder__Area_Ukraine
1.0
0.5
0.0
−0.5
−1.0
−1.5
standrization__avg_temp


SHAP value for standrization__Year
40000
20000
0
−20000
−40000
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
standrization__Year
OneHotEncoder__Item_Potatoes
1.0
0.5
0.0
SHAP value for standrization__average_rain_fall_mm_per_year
40000
20000
0
−20000
−40000
−1
0
1
2
3
standrization__average_rain_fall_mm_per_year
OneHotEncoder__Item_Sweet potatoes
1.0
0.5
0.0

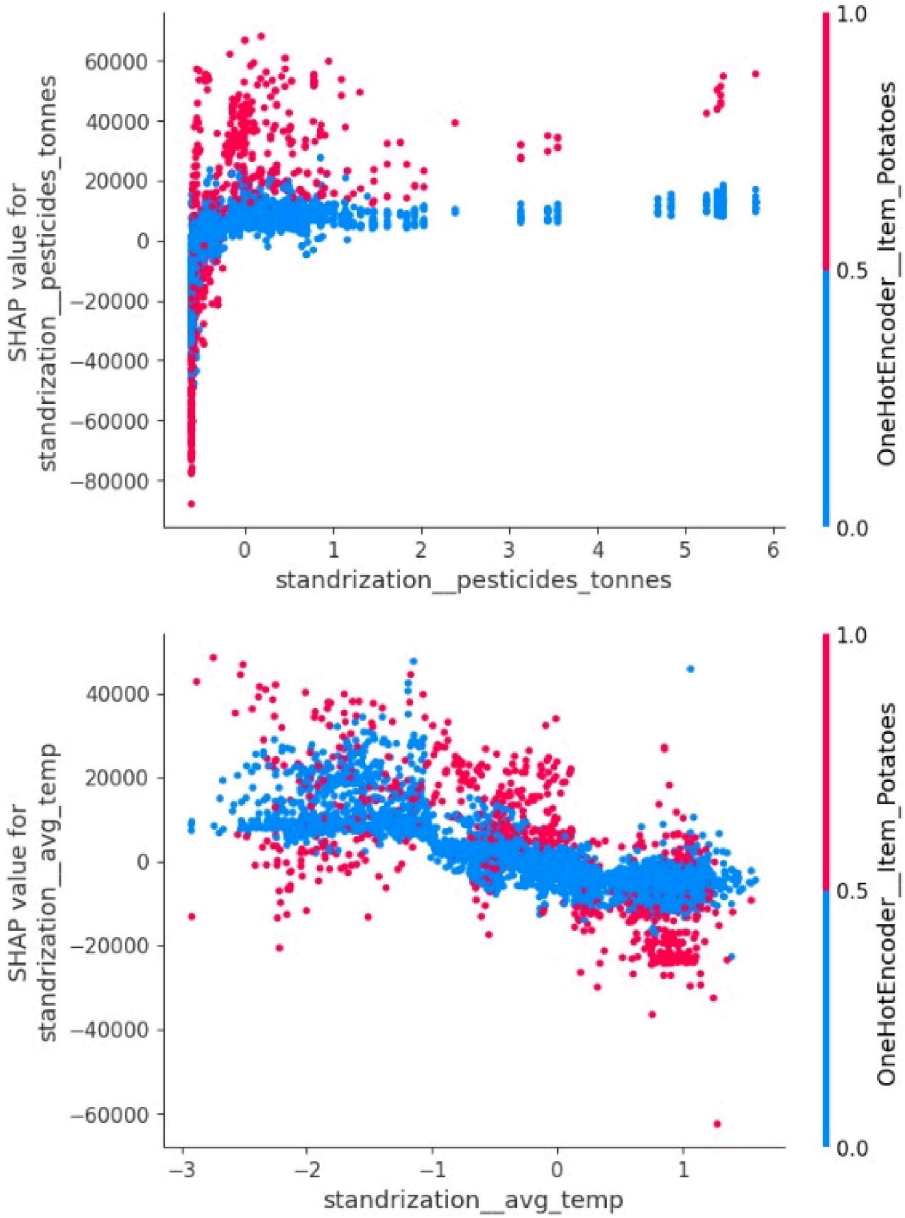


Figure 5.3: SHAP Dependence Plot

5.4 Local Explanations: Individual Predictions

In addition to the global patterns, SHAP offers local explanations for individual samples, by using waterfall plots.

HIGH-YIELD PREDICTION EXAMPLE:

Predicted Yield: 177,737 hg/ha

Base Value (Model Average): 76,670.6 hg/ha

Feature Contributions (from biggest to smallest):

• OneHotEncoder_Item_Potatoes = 1: +56,756.7 hg/ha

• standardization_avg_temp = -1.07: +39,657.4 hg/ha

• standardization_Year = +1.204: +15,322.9 hg/ha

• standardization_avg_rainfall = +0.818: -1,504.0 hg/ha

Other regional factors: Small positive/negative contributions.

Interpretation: The forecast of 1,77,737 hg/ha is significantly above the base(76,670.6) , and is followed by :

1. Crop type : it is the main factor for determining, and here Potatoes yields more than other average crops

2. Temperature: Potatoes grow best in cool temperatures as they are adapted to that climate.

3. Year: In the recent years there is improvement in the yield due to change in technology.

4. Rainfall: Predictive ability is slightly lowered in the presence of moderate rainfall.

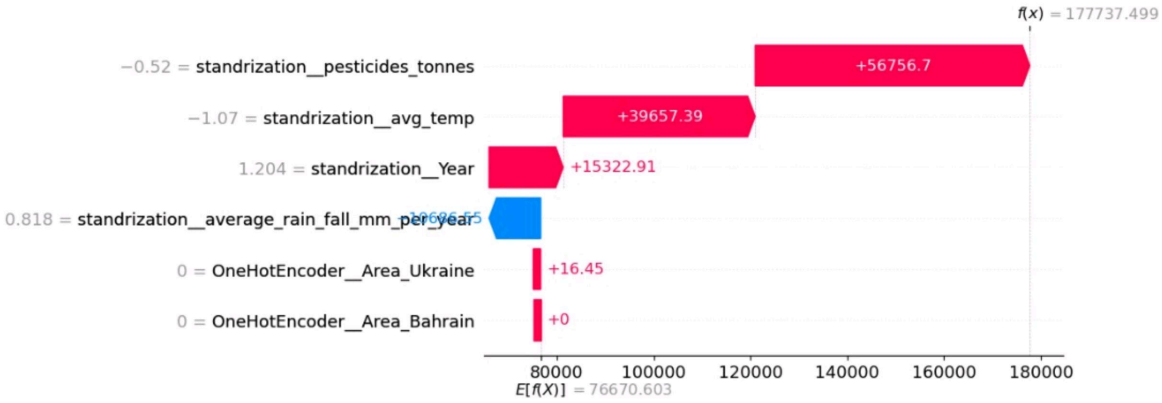


Figure 5.4: SHAP Waterfall Plot for High Yield case

LOW-YIELD PREDICTION EXAMPLE:

Predicted Yield: 63,380.6 hg/ha

Base Value: 76,670.6 hg/ha

Feature Contributions: • standardization_pesticides_tonnes = -0.595: -9,582.2 hg/ha

• standardization_avg_rainfall = -0.636: -3,464.97 hg/ha

• standardization_Year = -0.638: -2,766.57 hg/ha

• standardization_avg_temp = -1.823: +2,499.63 hg/ha

• Minimal contributions are made by other factors.

Interpretation: The prediction of 63,380.6 hg/ha is below the base due to:

1. Pesticide shortage: yield is less if there is not enough pesticide.

2. Low Rainfall: Yield decreases during drought.

3. Early Year: is a proxy for lower productivity, as is the early year.

4. Cold Temperature: Negative but positive, indicating that very cold temperatures in this case can lead to reasonable yields.

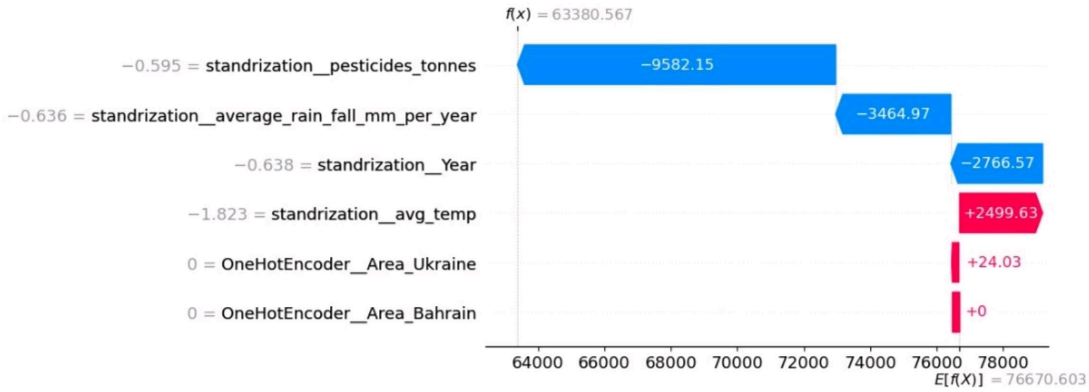


Figure 5.5: SHAP Waterfall Plot for Low Yield case

5.5 Linear Regression SHAP Comparison

SHAP was applied to the linear regression model for comparison along with the selection of Decision tree for deployment. The results were:

- More uniform features contributions (linear additivity).
- Identical SHAP values for identical feature combinations.
- Unable to capture the effects of non linear pesticide and temperature by tree based SHAP.

By this comparison, we could get the insight why decision Tree is better in capturing agricultural complexity than linear models.

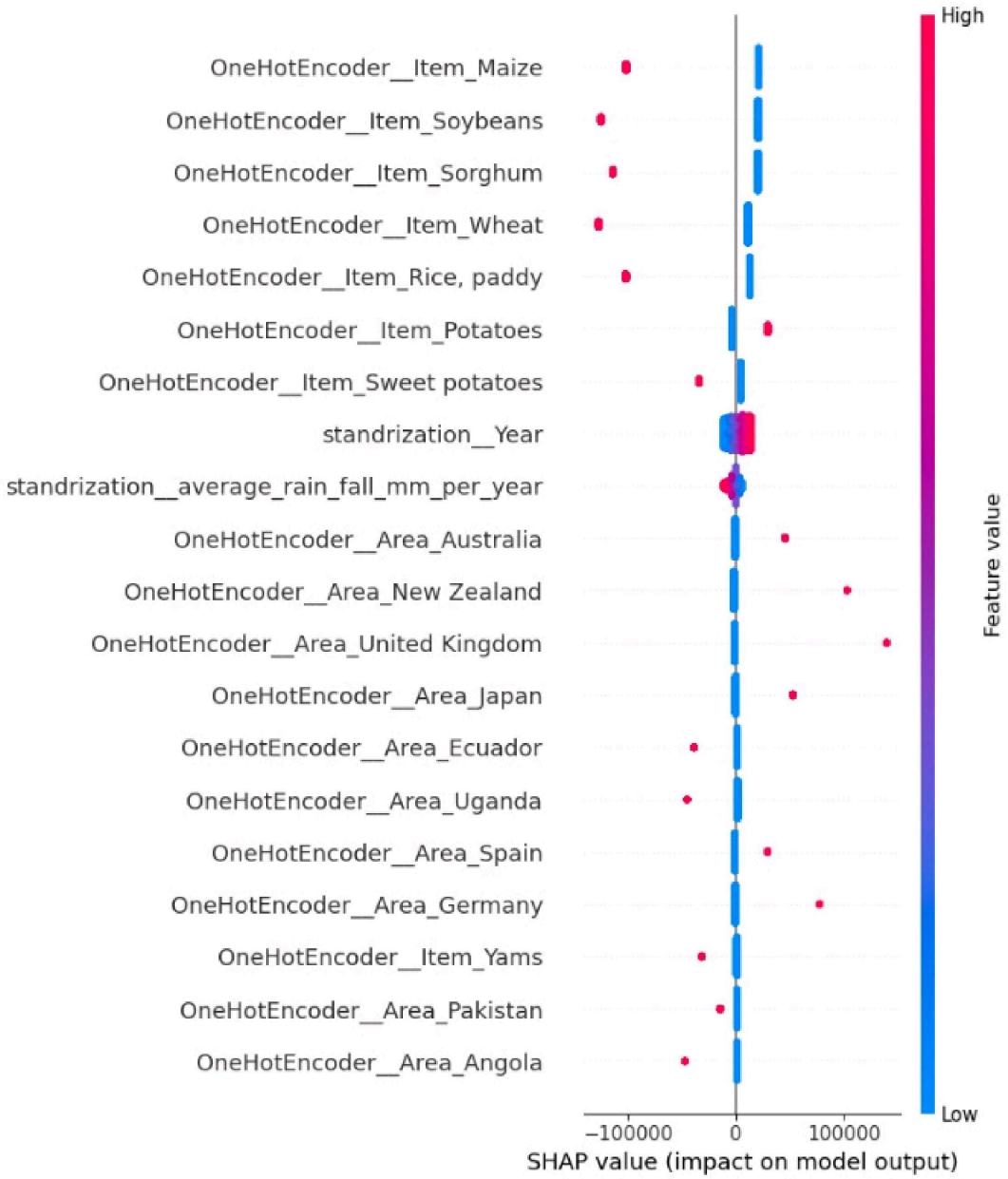


Figure 5.6: SHAP Summary Plot

5.6 Practical Implications

The SHAP analysis draws the following results/ conclusions:

- Crop selection: Potatoes offers higher yields under favourable conditions, which supports cultivation planning.
- Pest management: The optimum utilisation of pesticides is necessary. Avoidance of insufficient or excessive pesticides as it doesn't enhance yield.

- Climate adaptation: Temperature and rainfall supports precision agriculture as they are non linear and crop dependent.
- Technology Adoption: The recent trend shows improvements in yields which can lead to improved varieties and agronomic practices.
- Regional variations : Geographic location plays a key role because it reflects the soil quality, Infrastructure and policy differences.

Chapter 6

INSIGHTS AND DISCUSSION

6.1 Key Findings

The most stunned result is the crop type, speciically potato, is the most important factor which determines yield. This reflects both - data characteristics and agricultural reality:

DATA PERSPECTIVE : The dataset is very potato specific (3956 out of 25,932 records, 15%), which means the procedure followed give strong insights about potato(specific crop).

FARM REALITY : Crops have natural productivity limits. Under optimum conditions, potatoes can attain 50+ tonnes/ha, while some staple grains reach 10 tonnes/ha at best.

Implication : The model automatically generates baseline yields for each crop in cross-crop prediction. It's suitable when applied to the use case "If we plant a crop in an area, how much can we expect to harvest?"; if it's applied to the use case "Which crop should we plant in a given area?" then it needs to be modified.

NONLINEAR PESTICIDE-YIELD RELATIONSHIP : Pesticide dependence plot showed that overuse of pesticides leads to decrease in prediction value. This is a counter intuitive finding consistent with agricultural science literature demonstrating:

Often, high levels of pesticides (phytotoxicity) can harm crop tissue.

Ecological Trade-offs: Disruption of beneficial insects and soil microorganisms Heavy use of pesticides can lead to resistance. When the marginal yield is not increasing at a rate that can offset the added cost, cost-benefit principle implies that the benefits are not worth the extra cost.

This relationship has been discovered from the data, with no explicit domain knowledge

being fed to the system; this illustrates the power of machine learning for hypothesis generation in complex systems. Optimal pesticide level depends on crop and climate; additional analysis can help to determine specific recommendations for particular crops.

CLIMATE INTERACTIONS : The effect of temperature and rainfall is not consistent across the dataset:

Crop-Dependent Optima : The same temperature is optimal for some crops while it is not for others. The distribution of rainfall is either deficient or excessive. – Temporal Trends: There is an improvement in yield over the period of 24 years, which indicates that technology adoption is a major contributor.

These interactions are captured naturally by the Decision Tree, without any feature engineering.

6.2 Model Limitations and Boundary Conditions

DATA IMBALANCE : The data is very imbalanced:

India records for 3500 (13.5%).

Less than 50 records were available for majority of countries.

Potatoes: 3956 records (15.3%).

Implications:

• The model was used for the typical combinations of crops and regions.

• The values lies between minimum and maximum extremes according to median trend.

• If we calculate intervals for extreme cases , they would be large.

TEMPORAL SCORE : Duration of the dataset: 1990-2013 (24 years).It provides the insights that the relationship between agriculture and other variables remains constant over a decade.However,

• There can be a shift in the temperature/rainfall due to climate change.

• There are chances for insect, and pest diseases.

• There can be a substantial increase in the market demand for new crop.

• The changes in the policy can affect agronomic practices.

MISSING VARIABLES : There are some factors which impacts the yield but are not mentioned in the dataset :

• Resistance of soil to tilling, Sowing and planting

• Availability of water and system for irrigation

• Crop variety

• Threats from pest and diseases

• Farmer's knowledge and agricultural techniques

The model is optimised for the above factors. The involvement of soil bs management data can enhance the performance.

GEOGRAPHICAL EXTRAPOLATION : This model was developed by the help of historical data of countries where the FAQ data is available. After it's application to countries with poor agricultural statistics,it was assumed that the agricultural systems in countries are similar.

6.3 Generalization Assessment

The R^2= 0.9793 of the Decision Tree on test set is good. However, By collection data about the population of 24 years and 130 countries, the first test set is applied. It is a temporal sample. Model activities for different periods or crop variety are not known. High R^2 indicates that a lot of variance are explained by crop type.

A more strict evaluation would include:

• Temporal validation: Train on 1990-2005 , test on 2006-2013

• Geographic validation: Train on others while leaving out one country

• Crop specific validation: Leave out the one crop type entirely

6.4 Comparison with Alternatives

VS. LINEAR MODELS: Linear regression (R^2= 0.747) assumes that all the relationships are linear. Decision trees shows 23% as improvement in R^2 which means agricultural data is non linear. Although, there are some advantages of linear model:

• Coefficients are directly predictable.

• Since the belief of "big government" is essential pulled apart , the regulatory requirements are less heavy.

VS. ENSEMBLE MODELS : The single decision trees would be surpassed by Random forests (not tested here) due to reduction in variances.Trade-offs:

• There was higher accuracy and lower interpretability.

• Single trees- shows relatively low peak accuracy.

Although The farmers interpretation is most important, the interpretability benefit of the single tree might reimburse for the accuracy improvement.

VS. DEEP LEARNING: The neural networks have the capacity to learn in complex patterns. However,

• The number of data points (25,932 samples) is relatively smaller for deep learning.

• Needs to be regulated so it does not overfit.

• Interpretability is poor.

• It requires more resources for calculation.

The tree based models are more reliable and best suited for the data and application.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Summary

This project created a machine learning framework to estimate crop yields based on environmental, climate and agronomic information. The significant achievements are:

1. **COMPARATIVE MODEL ANALYSIS:** Five regression algorithms were applied and tested; Decision Trees showed a much greater R^2 value (0.979) than the linear algorithms (0.747).

2. **SHAP-BASED INTERPRETABILITY:** TreeExplainer-based SHAP analysis offered a whole new perspective on factors that drive yield, with the following insights:

   - The most important predictor is ”crop type” (particularly potatoes).
   - The use of pesticides has a nonlinear and inverted U-shape relationship with yield.
   - Climate effects vary by crop and are nonlinear.
   - All relationships are modulated by geographic location.

3. **DEPLOYMENT-READY MODEL:** Picked and a preprocessing pipeline are produced for a Decision Tree model, which allows for reproducible predictions on new

data, and makes it suitable for deployment in agricultural decision-support systems.

4. **ACTIONABLE INSIGHTS:** Practical recommendations for crop selection, optimization of pest management and climate adaptation measures are derived from the analysis.

7.2 Main Results

Performance comparison :

Linear Regression : R^2=0.7473 MSE=299.87

Decision Tree : R^2=0.9793 MSE=3,941

Improvement : +23.2pp Better MSE

40 The combination of high accuracy and interpretability of the Decision Tree Model for the prediction of agricultural task is the best choice.

7.3 Practical Applications

This model enables a number of practical applications:

- Yield Forecasting: Government and agribusiness can make predictions of expected harvests to aid in supply chain planning.
- Insurance and Risk Assessment: Model predictions used to calculate premiums and estimate the probability of claims.
- Decision support: Farmers experimented with ”what if” scenarios using the model.
- Research Hypothesis Generation: SHAP predicts unexpected relationships(high level of pesticides correlate with lower yields) that would be investigated in the field.

7.4 Limitations and Caveats

It is important for the practitioner to be aware of the following:

- Imbalance: This model was designed primarily for the pairing of crop regions . So , the predictions received for the uncommon pairings are not as reliable.
- This model assumes temporal stationarity: it means the relationships which applied and observed in 1990-2013 may not be reliable and applicable to 2024+ without requalification.
- Loss of variables: The variables such as soil properties, crop varieties, pest pressures , and management practices are lost.
- Geographic specificity - while focusing on the well represented areas or predicting in their training data can lead to systematic biasness in areas which are less represented.
- Interaction effects: Few interactions are recorded in the model whereas the complex crop specific practices not available in the data are not modeled.

7.5 Future Work

There are several ways by which work could be improved:

DATA AUGMENTATION:

Soil data integration: factors like soil type, pH, nutrient content would significantly enhance prediction.

Remote sensing: For capturing about the growing dynamics , the vegetation indices and soil moisture are responsible factors.

Pest and disease data: For increasing accuracy, direct observation or regional diseas surveillance data are used.

Recommendation systems: Recommendation systems like crop variety, irrigation type, farmer education level plays a key role .

METHODOLOGICAL EXTENSIONS:

Ensemble method: For chasing accuracy, the use of train random forest and gradient boosting is required.

Walk forward validation: for the evaluation of generalization the use of walk forward validation is required .

Includes regional models: Use of crop specific or region specific sub models (not a single

global model).

Uncertainty quantification: it includes prediction intervals through quantile regression kr Bayesian approaches.

DEPLOYMENT AND USABILITY:

Web application: creation of user friendly interface that enables farmers to get yield predictions by using SHAP.

Mobile App : Develop mobile application with offline capabilities for resource constrained areas .

Connection with Agronomic models: Link of data driven and mechanistic crop growth models.

Continuous learning: use of continuous pipeline for retraining model with new FAQ data.

RESEARCH AVENUES:

Pesticide optimization field trials: Validation of model's discovery of inverted U pesticide yield relationship.

Climate adaptation strategies: Identification of most climate affected resistance crop through the use of SHAP values.

Casual inference: The shift from predictive to casual analysis.

Supply Chain Optimization: careful use of crop mix and regional specialization by using model predictions.

7.6 Final Remarks

This project concludes that by the use of machine learning techniques and thorough interpretability analysis, the conclusions can be drawn from the agricultural data.The Decision Tree Regressor is not just accurate but it is understandable and can be used in practical applications. The SHAP analysis that crop yield is a difficult function of crop biology, climate, farming practices, and geography, and there is no single primary parameter which can be applied to all areas. The complexity is so ample that non linear models are justfied but also accentuate the value of local knowledge and on farm experimentation. By combining both methods (SHAP and selected models - trees over black boxes when possible) can help with more explainable systems which are simple , easy and understandable by stakeholders, which leads to adoption of data driven agriculture.

5

Appendix A: Data Preprocessing Code

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.model_selection import train_test_split

# Load data
df = pd.read_csv('yield_df.csv')

# Remove index column
df = df.drop('Unnamed: 0', axis=1, inplace=True)

# Handle missing values and duplicates
df.dropna(inplace=True)
df.drop_duplicates(inplace=True)

# Reorder columns
col = ['Year', 'average_rain_fall_mm_per_year',
    'pesticides_tonnes', 'avg_temp', 'Area', 'Item', 'hg/ha_yield']
df = df[col]

# Separate features and target
x = df.drop('hg/ha_yield', axis=1)
y = df['hg/ha_yield']

# Train-test split
x_train, x_test, y_train, y_test = train_test_split(
  x, y, test_size=0.2, random_state=0
)

# Preprocessing pipeline
preprocessor = ColumnTransformer(
  transformers=[
    ('onehot', OneHotEncoder(drop='first'), [4, 5]),
    ('standardization', StandardScaler(), [0, 1, 2, 3]),
  ],
  remainder='passthrough'
)

# Apply preprocessing
x_train_dummy = preprocessor.fit_transform(x_train)
x_test_dummy = preprocessor.transform(x_test)

print(f"Training set shape: {x_train_dummy.shape}")
print(f"Test set shape: {x_test_dummy.shape}")
```

Appendix B: Model Training and Evaluation Code

Appendix C: SHAP Analysis Code

turnitin Page 52 of 59 - Integrity Submission Submission ID trn:oid:::27535:139912595

Appendix D: Prediction Function

Appendix E: Dataset Summary Statistics

turnitin Page 54 of 59 - Integrity Submission Submission ID trn:oid:::27535:139912595

Bibliography

[1] Dataset link,"https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=yield_df.csv ".

[2] Github Reference,"https://github.com/611noorsaeed/Predicting-Crop-Yields-Crop-Yield-Prediction-Enhancing-Agriculture-with-Machine-Learning-Hindi/blob/master/notebook%20and%20dataset/CropYield-Prediction.ipynb"

[3] Janmejay Pant, R.P. Pant, Manoj Kumar Singh, Devesh Pratap Singh and Himanshu Pant, "Analysis of agricultural crop yield prediction using statistical techniques of machine learning", Materials Today: Proceedings, vol. 46, part 20, pp.10922-10926, 2021.

[4] P. Sharma, P. Dadheech, N. Aneja and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," in IEEE Access, vol. 11, pp. 111255-111264, 2023.

[5] A. Badshah, B. Yousef Alkazemi, F. Din, K. Z. Zamli and M. Haris, "Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability," in IEEE Access, vol. 12, pp. 162799-162813, 2024.

[6] Bharati Panigrahi, Krishna Chaitanya Rao Kathala, M. Sujatha, "A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning with Regression Models", Procedia Computer Science, Elsevier, vol. 218, pp.2684-2693, 2023.

[7] El-Kenawy, ES.M., Alhussan, A.A., Khodadadi, N. et al., "Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture", Potato Res. 68, 759–792 (2025).

[8] Kavita Jhajharia, Pratistha Mathur, Sanchit Jain, Sukriti Nijhawan, “Crop Yield Prediction using Machine Learning and Deep Learning Techniques”, Procedia Computer Science, Vol 218, pp.406-417, 2023.

[9] Akash Manish Lad, K. Mani Bharathi, B. Akash Saravanan, R. Karthik, “Factors affecting agriculture and estimation of crop yield using supervised learning algorithms”, Materials Today: Proceedings, Vol 62, Part 7, pp. 4629-4634, 2022.

[10] Rajswee Surana, Ritu Khandelwal. Crop Yield Prediction Using Machine Learning: A Pragmatic Approach, 01 July 2024, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-4575893/v1].

[11] Sowjanya Addu, Srujana Sheelam, Samhitha Mekala, Nazma Sulthana, Lohitha Mekala and Zaid Alsalami, ”Assessing Environmental Impact: Machine Learning for Crop Yield Prediction”, E3S Web of Conf., vol. 529, 03008 (2024).

PAPER ACCEPTANCE PROOF

turnitin Page 57 of 59 - Integrity Submission Submission ID trn:oid:::27535:139912595

REGISTRATION PROOF

turnitin Page 58 of 59 - Integrity Submission Submission ID trn:oid:::27535:139912595

SCOPUS INDEXED CONFERENCE PROOF