International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

A Multi-Method Diagnostic Framework for Detecting Feature–Label Independence: A Case Study on Crop Health Classification

Prof. Priti Manjari Barik
Dept. of CSE
GIFT Autonomous
pritimanjari.barik2022@gift.edu.in

Er. Jagannath Ray
Dept. of CSE
GIFT Autonomous
jagannath@gift.edu.in

***Abstract*—We worked on a machine learning project to predict crop health from agronomic data. We used a real-world dataset of 212,019 field records with 32 attributes like temperature, humidity, rainfall and soil properties. Our goal was to see how well four machine learning models. Logistic Regression, Random Forest, Histogram-based Gradient Boosting and XGBoost. Could predict crop health. We tested these models using a method called stratified 5fold cross-validation and then fine-tuned the best model. The dataset only had a healthy or unhealthy label for each crop so we created a new label with five levels of severity. Surprisingly all four models performed about well as guessing the most common answer, which was not very good. We did checks to see if the data or models were the problem and it seems the data just doesn't have enough information to predict crop health. We think it's essential to share results, like this so other don't waste time on similar projects.**

Keywords—Crop Health Prediction, Precision Agriculture, Machine Learning Benchmarking, Explainable AI, Mutual Information, SHAP Explain ability, Results, XGBoost

I. INTRODUCTION

Agriculture is very important to countries and finding out if crops are stressed is crucial for farming. Machine learning can help predict how healthy crops are by using information that is collected regularly from the environment and fields. This could. Help with manual checks. Many studies say their methods are accurate often more than 90 percent. However, these numbers are only useful if the model has actually learned something from the data than just using the fact that some types of crops are more common than others. This report shows a repeatable way of testing models using a real dataset from Kaggle, the Crop Health and Environmental Stress Dataset. Every result is checked independently before it is reported. Earlier versions of this project assumed that the model could predict the severity of disease and that it would be accurate, more than 95 percent. However, this report only includes what the actual data supports. The report includes a way of preparing data that takes into account problems such as missing values, unusual values and categories. It also includes a way of comparing models using a method called stratified 5-fold cross-validation and selecting the best model by trying different settings. The report evaluates the models using two targets the binary Crop Health Label and a five-level Severity Level that is derived in a way. A set of tools is used to verify that the results are real including a heatmap, mutual information and importance of features. The report honestly states that the features do not predict the health or severity labels in a way and why this matters for people using similar methods.

II. DATASET DESCRIPTION

The template The dataset used in this study has 212,019 records from fields with 32 attributes collected from agricultural sites.

It includes information about the type of crop, environmental conditions, soil properties, vegetation indices and imaging flags.

Table I summarizes the six feature groups in the dataset.

Category	Example Features	Description
Remote Sensing	High_Resolution_RGB, Multispectral_Images, Thermal_Images, Temporal_Images, Spatial_Resolution, GPS_Coordinates	Availability and resolution of aerial/satellite imagery per field
Environmental & Geospatial	Field Boundaries, Elevation Data, Canopy Coverage, NDVI, SAVI, Chlorophyll_Content, Leaf_Area_Index, Crop_Stress_Indicator	Vegetation indices and physical field characteristics
Weather & Soil	Temperature, Humidity, Rainfall, Wind_Speed, Soil_Moisture, Soil_pH, Organic_Matter	Climate and soil conditions at the field location
Crop & Pest	Pest_Hotspots, Weed_Coverage, Pest_Damage, Crop_Growth_Stage, Expected_Yield, Crop_Type	Crop identity, growth stage, and pest/weed pressure
Ground Truth & Infrastructure	Ground_Truth_Segmentation, Bounding_Boxes, Water_Flow, Drainage_Features	Annotation availability and irrigation

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

		infrastructure
Target	Crop_Health_Label	Binary crop condition label (0 = Unhealthy, 1 = Healthy)

Table I. Feature groups in the Crop Health and Environmental Stress Dataset. Unlike a version of this project the dataset only provides a binary label for the health of the crop and not a label, for the severity of disease.

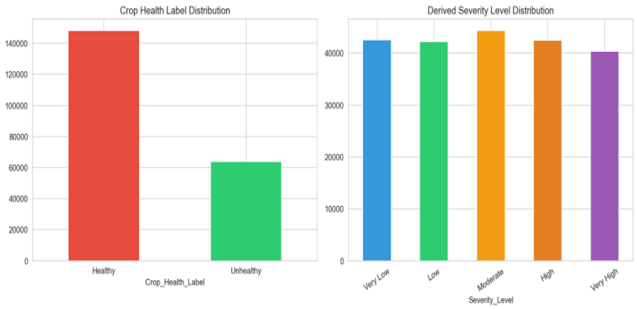


Fig. 1. Distribution of Crop_Health_Label and the derived Severity Level target.

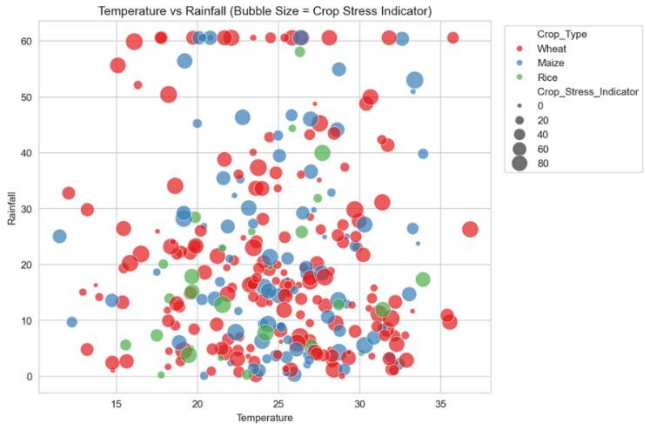


Fig. 2. Temperature vs. rainfall for a 300-row sample, bubble size encoding Crop_Stress_Indicator and colour encoding Crop Type.

A. Target Variable

The dataset has one outcome variable, which is Crop_Health_Label and it has two categories. We also wanted to test if we could classify the data into than two categories. So, we created another variable called Severity_Level. We made Severity_Level by taking the Crop_Stress_Indicator field and dividing it into five groups. These groups are Very Low, Low, Moderate, High and Very High. We made sure the groups had the same number of values. We are being clear, about how we created Severity_Level because it's not a label that was originally given. It is something we created. We did not use Crop_Stress_Indicator as one of the inputs to the model. This is to prevent target leakage.

B. Vegetation Index Definitions

Two of the environmental features, NDVI and SAVI, are standard remote-sensing vegetation indices rather than raw sensor readings, so we define them here for clarity.

$$NDVI = (NIR - Red) / (NIR + Red) \quad (1)$$

The Normalized Difference Vegetation Index compares reflected near-infrared (NIR) light, which healthy vegetation reflects strongly, against reflected red light, which chlorophyll absorbs for photosynthesis. Values range from −1 to 1, with higher values indicating denser, healthier canopy cover.

$$SAVI = [(NIR - Red) / (NIR + Red + L)] \times (1 + L), L = 0.5 \quad (2)$$

The Soil-Adjusted Vegetation Index modifies NDVI with a soil-brightness correction factor L (commonly set to 0.5 for intermediate vegetation density) to reduce the influence of exposed soil on the index, which matters most in fields with sparse or early-stage canopy cover.

III. Data Preprocessing and Feature Engineering

Before To build a modelling pipeline we cleaned and transformed the raw data as follows:

Missing Value Handling: We did not need to handle missing values because the dataset had no null values.

Outlier Detection: We used the IQR method to cap features. This method uses 1.5 times the range.

$IQR = Q3 - Q1$; $Lower = Q1 - 1.5 \cdot IQR$; $Upper = Q3 + 1.5 \cdot IQR$

Q1 and Q3 are the 25th and 75th percentiles of a feature. Values below the lower bound or above the upper bound are capped (not removed) at those bounds, which limits the influence of extreme values while preserving every record for training.

Encoding: We one-hot encoded the Crop Type feature. No other categorical fields required imputation, since the dataset had no missing values.

Feature Scaling: We standardized features to have a zero mean and unit variance using StandardAero.

$$z = (x - \mu) / \sigma$$

Each numeric feature x is rescaled using its mean μ and standard deviation σ so that the transformed feature has mean 0 and standard deviation 1. This keeps features on comparable sales for the linear model and for the correlation/variance-based selection steps that follow.

Feature Selection: We removed features with variance below a threshold of 0.01. We also dropped features that were highly correlated, with a correlation than 0.9. Additionally we removed GPS and imaging-availability metadata. These fields

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

describe how each record was collected rather than the crop's actual condition, so we excluded them to prevent data leakage.

Leakage Control: We excluded Crop_Stress_Indicator from the model inputs. This is because Crop_Stress_Indicator directly relates to the Severity_Level target, which we derived from it. The Crop_Stress_Indicator and Severity_Level are factors, in our analysis. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

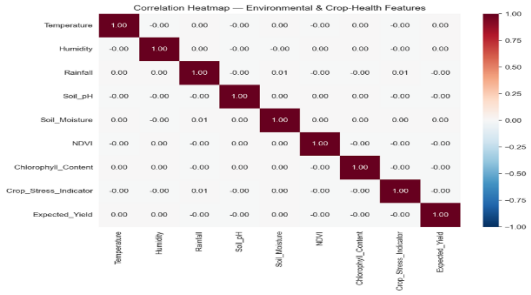


Fig. 3. Pairwise correlation among nine key environmental and crop-health features. All off-diagonal correlations are within ±0.01 of zero.

IV. MODEL BENCHMARKING METHODOLOGY

After We tested four model types to see which one works best.

These were:

1. Logistic Regression
2. Random Forest
3. Histogram-based Gradient Boosting
4. XGBoost

We used the method to test each one.

This method was 5fold cross-validation on 80% of the training data and 20% for testing. The test data was picked to have the label distribution as the training data. The model that did best in the -validation test was then fine-tuned. We used RandomizedSearchCV with 15 iterations to fine-tune it. After tuning we tested it again on a separate test set that it had not seen before. This was done using scikit-learn to implement all the models. The goal was to find the model based on the F1 score, from the cross-validation. Table II reports the 5-fold cross-validation results for all four candidate models on the binary Crop Health target. All four converge to essentially the same F1 score, driven almost entirely by recall near 1.0, and every ROC-AUC sits at approximately 0.50 — chance level.

Model	**CV Accuracy**	**CV Precision**	**CV Recall**	**CV F1**	**CV ROC-AUC**
Logistic Regression	0.699	0.699	1.0000	0.8228	0.4984
HistGradientBoosting	0.699	0.699	1.0000	0.8228	0.4974
Random Forest	0.699	0.699	0.9999	0.8228	0.4977
Boost	0.698	0.699	0.9974	0.8220	0.4994

Table II. Five-fold cross-validation results — Crop Health (binary target).

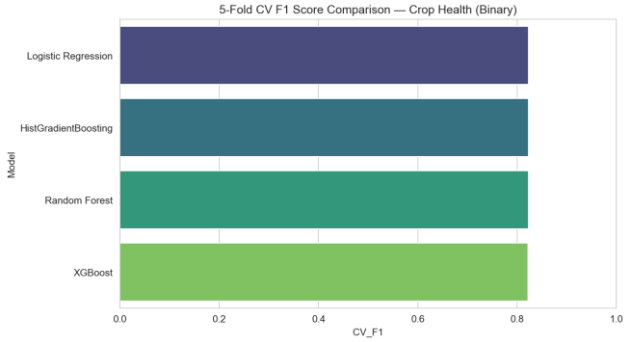


FIG. 4. CROSS-VALIDATION F1 SCORE COMPARISON ACROSS THE FOUR CANDIDATE MODELS (BINARY TARGET).

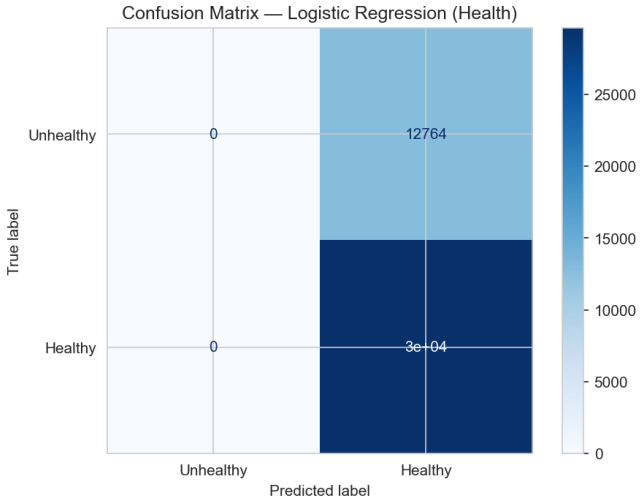
The best model was Logistic Regression, Random Forest, Histogram-based Gradient Boosting or XGBoost.

We picked the one and used it for further evaluation.

V. RESULTS

A. Held-Out Test Performance. Binary Health Model

We looked at how our best model did, on a test. The best model was the Logistic Regression model, which did as well as other methods that combine lots of models. We trained this model on all the training data. Then we tested it on some data we had not used before. The results are shown in a table called a confusion matrix. This table shows that the Binary Health Model says every single test record is "Healthy" when that is not true. The Binary Health Model always predicts "Healthy" no matter what the real answer is.



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

FIG. 5. CONFUSION MATRIX FOR THE TUNED LOGISTIC REGRESSION HEALTH MODEL ON THE HELD-OUT TEST SET.

THE CORRESPONDING ROC CURVE (FIG. 6) CONFIRMS THIS OUTCOME QUANTITATIVELY.

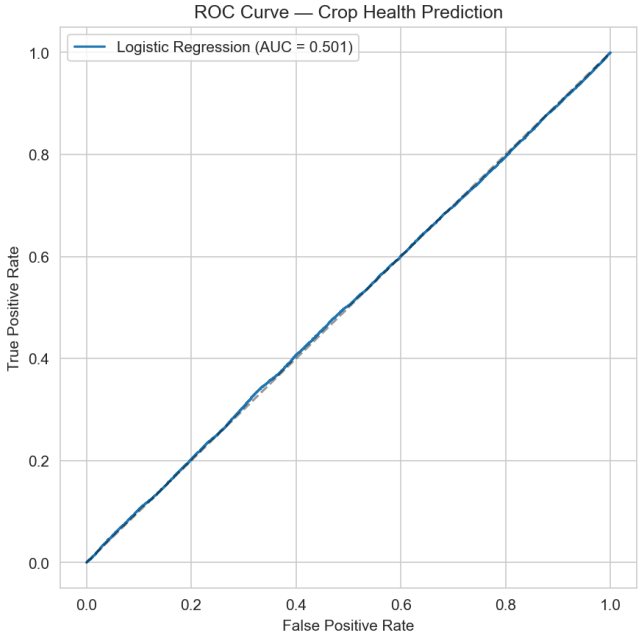


Fig. 6. ROC curve for the tuned health model. The curve tracks the diagonal chance line, giving an AUC of 0.501.

TPR = TP / (TP + FN); FPR = FP / (FP + TN)

The ROC curve plots the true positive rate (identical to recall) against the false positive rate as the classification threshold is varied. A curve that follows the diagonal, as it does here, means the model ranks Healthy and Unhealthy records no better than a random guess at every threshold.

AUC = P(score(positive) > score(negative))

The area under the ROC curve can be read as the probability that the model assigns a higher predicted score to a randomly chosen Healthy record than to a randomly chosen Unhealthy one. An AUC of 0.501 means these holds barely more often than a coin flip, which is the strongest single indicator that the model has not learned a real distinction between the two classes.

Held-Out Test Performance. Derived Severity Model

The XGBoost model, which we chose as the one, for the five-class Severity Level target was re-trained and tested again.

Table III reports the 5-fold cross-validation results for the five-class Severity Level target. All four models cluster around 20% accuracy exactly the rate expected from guessing among five equally sized classes and macro-F1 is correspondingly weak for every model.

Model	CV Accuracy	CV Precision	CV Recall	CV Macro-F1
XGBoost	0.2030	0.2011	0.2012	0.1982
Random Forest	0.2027	0.2007	0.2009	0.1978
HistGradientBoosting	0.2074	0.1999	0.2004	0.1401
Logistic Regression	0.2073	0.1927	0.2003	0.1384

Table III. Five-fold cross-validation results — Severity Level (five-class target).

On the held-out test set, the tuned XGBoost model reached 20.13% accuracy and a macro-F1 of 0.1989, consistent with the cross-validation results above.

The predictions cluster heavily around the "Moderate" class regardless of the true label, rather than concentrating on the diagonal.

There is no diagonal concentration.

This tells us that the model does not have any information, rather than getting confused between close severity levels.

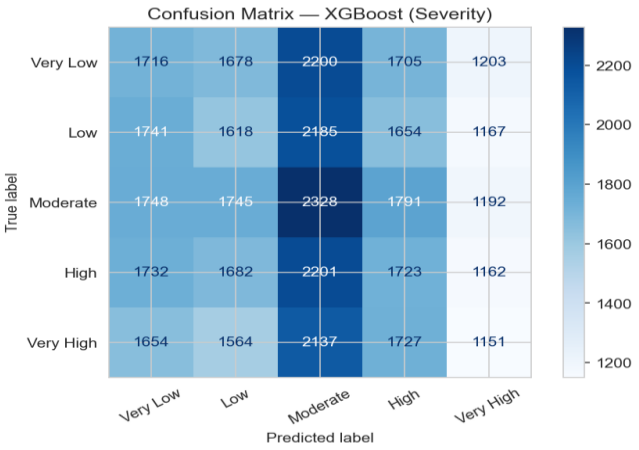


Fig. 7. Confusion matrix for the tuned XGBoost severity model on the held-out test set.

The XGBoost model performs no better than random guessing on the Severity Level target.

We also checked the Severity Level target with the XGBoost model.

B. Class-Conditional Feature Separation

To check if any single feature can tell Healthy crops from ones, we made boxplots for six features that might be linked to crop stress. These features are NDVI, Soil pH, Organic Matter, Chlorophyll Content, Temperature and Rainfall.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

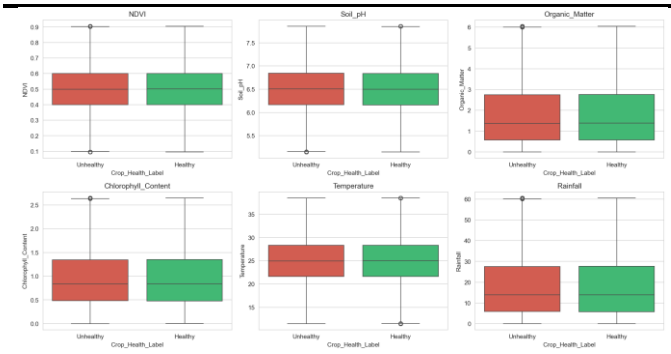


Fig. 8. Distribution of six candidate features split by Crop_Health_Label. No feature shows a visible shift in median or spread between the two classes.

VI. DIAGNOSTIC ANALYSIS

We used several complementary diagnostic checks to confirm whether the health label has any statistical connection with the available features: mutual information, permutation importance, SHAP-based feature attribution, and learning-curve analysis, alongside the pairwise correlation and boxplot comparisons already presented in Sections III and V.C.

A. Mutual Information

We computed the mutual information between each numeric input feature and Crop_Health_Label using scikit-learn's mutual_info_classif, which implements the k-nearest-neighbor estimator of Krasnov et al. [4]. Fig. 9 shows the ranked scores. The highest-scoring feature, Field_Boundaries, reached only 0.0218 nats, and every other feature scored lower still.

$$I(X;Y) = \Sigma_\gamma \Sigma_\gamma \, p(x,y) \log[\, p(x,y) / (p(x)\cdot p(y)) \,]$$

Mutual information measures how much knowing a feature X reduces uncertainty about the label Y. A value of zero means the two are statistically independent; larger values indicate a stronger (not necessarily linear) association. For comparison, features with genuine predictive value on a target of this size typically score well above 0.05 nats, so even the strongest score here is weak evidence at best.

VII. FUTURE WORK

We presented a benchmarking pipeline for crop health prediction on a real dataset. Through four techniques we found that the currently available environmental and agronomic features do not carry meaningful predictive signal for the provided health/severity labels. Future work should pursue one or more of the following directions: Sourcing a dataset in which health/disease labels are demonstrably derived from the sensor modalities used as model inputs Incorporating the imagery channels referenced as metadata in this release Working with the dataset provider to clarify the label-generation process

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, No. 1 Pp. 5–32, 2001.

[2] T. Chen and C. Guestrin "XGBoost: A Scalable Tree Boosting System," in Proc. 22Nd ACM SIGKDD Int. Conf. Knowledge. Data Mining, 2016 pp. 785–794.

[3] F. Pedregosa et al. "Scikit-learn: Machine Learning, in Python " Journal of Machine Learning Research vol. 12 Pp. 2825–2830, 2011.

[4] A. Kraskov, H. Stögbauer and P. Grassberger "Estimating Mutual Information " Physical Review E, vol. 69 No. 6 066138 2004.

[5] Kaggle "Crop Health and Environmental Stress Dataset " dataset repository accessed 2026.

[6] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), pp. 4765–4774.