Crop Yield Prediction Model

Naggita Ethel
Department of Networks
Makerere University
2100707022
21/U/07022
ethelnaggita@gmail.com

***Abstract*—This paper discusses my research, implementation and results of a crop yield prediction model I created to solve the problem of food insecurity using artificial intelligence. This model uses a dataset obtained from Kaggle containing numeric data of rainfall(measured in mm per year), pesticides(measured in tonnes), temperature of different countries across the world alongside the yields obtained from the year 1990 to 2013 for the ten most consumed crops in the world measured in hectograms per hectares. These crops include maize, rice, wheat, soy bean, potatoes, sorghum among others.**

The paper includes an introduction, background of the project, existing works, methodology, results and discussions, conclusion as well as references.

I am to discuss the research as well as my term paper contributions under existing works. In my methodology, I am to describe my dataset, data preparation and explanatory data analysis, model selection and optimization as well as model accountability. This paper also includes links to my dataset, python source code, machine learning interpretation technique as well as hypothesis for my research questions.

An agent is an entity that interacts with its environment by perceiving its surrounding via sensors and acting through effectors or actuators while an environment is the surrounding of an agent. The agent for this project is a computer which contains the algorithm developed using python programming language. This uses a keyboard, mouse or even touchpad as the sensors for inputting percepts (rainfall, temperature, pesticide values, year, country, crop) from the environment which is the garden. The agent has actuators like the monitor, laptop screen for displaying the likely to be produced crop yields. The environment for the agent is fully observable (The agent's sensors give it access to the complete state of the environment), stochastic (The next state of the environment is not determined by the current state) and also dynamic because it changes over time.

I split my dataset into training data(70%), testing data(10%) and evaluation data(20%) to be used during modeling. To increase my model accuracy and avoid over and under fitting, I performed hyperparameter tunning. During model evaluation, I evaluated my model using validation dataset, calculated evaluation metrics like mean absolute and mean squared error, precision, recall and f1-score plus a scatter plot to show the relationship between predicted and actual values.

My research questions included how do algorithms make crop yield prediction possible, how do pesticides help in crop prediction, Is there a reasonable improvement in food security due to crop yield prediction models, Is it possible to make fairly accurate predictions with a small dataset?

My research objectives include to come to realization of my project through development of a problem solving agent, to create co operation between artificial intelligence(ai) and humans. Most people are insecure around ai and I want to bridge this gap, to create an agent that represents knowledge through prediction of yields, to be able to create a learning agent through the supervised learning algorithms, to create an interpretable and well accounted for model. Also, to look for hypothesis for my research questions.

***Index Terms*—Artificial intelligence, Crop yield, Machine learning, Prediction, Food security**

I. Introduction

My keywords are artificial intelligence, crop yields, machine learning, prediction and food security. This is because I developed a machine learning model as a way of using artificial intelligence to encourage food security through prediction of crop yields. Artificial intelligence is the development of intelligent agents that can perceive, interpret and act in an environment. Machine learning is a subset of artificial intelligence which deals with development of computer systems that can learn and adopt without out direct instruction using algorithms. Crop yield is the amount of harvested crop. Food security is the state of having sufficient food. Prediction is a forecast of the future.

The problem to be solved is food insecurity. This matters to the community because zero hunger is the second sustainable development goal put in place therefore very important to come up with solutions to make sure it is attained. I therefore decided to come up with a crop yield prediction model using random forest regression to solve the problem of food insecurity using machine learning. It also matters to everybody in the world because we all need food for survival and to prevent malnutrition.

A crop yield prediction model will help reduce food insecurity through enabling farmers plant more earlier and storage of abundances as the quantity of harvest is known at an early stage.

II. BACKGROUND AND MOTIVATION

I needed to use artificial intelligence for my model because it is an artificial intelligence project. I needed to use machine learning because I was to develop machine learning models under artificial intelligence that was to predict yields through learning and adopting. I focused on a crop yield prediction model in order to solve the problem of food insecurity and also work towards achieving the second most important sustainable development goal. This is because I believe no one in the world deserves to starve even with the increased poverty and changing climatic conditions especially in Africa. I decided to implement my project using random forest regression technique because as a supervised learning algorithm, it ensures better accuracy through seeking better predictive performance by combining predictions from multiple models unlike decision trees. Furthermore, I am working with continuous data (my model is outputting yields that have more than two values) therefore required to work with regression. I also implemented my model using an unsupervised learning algorithm which is k-means clustering to understand relationships between features.

III. Literature Review (Existing Works)

There has been an increased development and use of crop yield prediction models. From tradition prediction models like the static regression approach to modern techniques like satellite imagery, deep neural networks among others. Many crop yield prediction models have already been created using unsupervised techniques like k-Means clustering, fuzzy clustering and supervised techniques like decision trees, linear regression, among others. Supervised learning algorithms are those that are trained using labeled data(input data is provided along with the output). Decision trees are tree-like visual models that illustrate every possible outcome of a decision, while linear regression is a linear approach for modeling the relationship between the dependent variable(in this case production/yield) on the Y-axis and one or more explanatory or independent variables(temperature, rainfall etc.) on the X-axis using a line of best fit known as regression line. These are both supervised learning techniques.

Unsupervised algorithms are those trained using unlabeled data(Only input data is provided). K-means clustering is an unsupervised learning algorithm where input data is put into clusters and each data point belongs to only one cluster. The 'k' stands for the number of clusters to be formed. Fuzzy clustering is a soft clustering algorithm which allows an object to belong to more than one cluster using membership degrees.

A. Research gaps in the literature

Linear regression is prone to over fitting. This is where a model has a high performance when tested using testing data but has a low performance when presented with validation data. It also works on an assumption that there is linearity between dependent and independent features or variables. Decision trees have high variance(model changes quickly with a change in training data) and give low prediction accuracy.

Unsupervised techniques like k-clustering are more depedent on initial values like choosing number of clusters and clustering outliers. Some people fail to carry out hyper parameter tunning using elbow or silhouette score and plot to determine the correct number of clusters. Fuzzy clustering performs poorly on datasets tat contain clusters with unequal densities or sizes. It is also sensitive to outliers.

One of the problems still facing crop yield prediction models is that a crop's response to factors like soil type, management practices, pests and diseases and climate patterns during the season is highly non linear and frequently difficult to understand. Also, crop yield prediction depends on multiple factors like crop genotype, environmental factors, management practices which are all hard to capture into one dataset. Another problem still faced is biased, inconsistent and inaccurate data. Data might be biased to a specific region. Changes in data overtime make it difficult to build a model that accurately predicts.

B. Term Paper Contributions

- To tackle some of the above problems, I chose to develop a random forest regression model which is less prone to over fitting, has better predictive performance than decision trees as it combines predictions from different subsets of decision trees.
- Most unsupervised learning algorithms lack transparency and interpretability but with a supervised learning algorithm like random forest, a model can easily be explained.
- I performed hyper parameter tunning which increases model performance for both random forest regression and k-means clustering.
- I used both a supervised and unsupervised learning algorithm to continue understanding relationships between my features.
- I used a dataset that contains data from most countries around the world therefore not regionally biased and has a large quantity therefore more examples for the model to learn.

IV. Methodology

The problem being investigated is food insecurity. We all need food to survive and avoid malnutrition. Zero hunger is also the second sustainable development goal therefore very important to implement. My model can be used for prediction of crop yields in the world for the ten most consumed crops. Artificial intelligence is addressing the problem through creation of a crop yield prediction model.

- Data Collection: This is the process of obtaining a proper dataset to be used by the model. I obtained my data from a dataset on Kaggle. Kaggle is an online website that contains free datasets and teaches machine learning.
- Data Pre processing: This is the manipulation or dropping of data before it is used to enhance performance. It involves data cleaning and data exploration.
- Model training: This is the process of teaching the model through examples using training data which is 70% of the original dataset for the random forest regression model.
- Model testing: This involves inputting new data based on the function inorder to receive expected outcomes. This was done using testing data which is 10% of my original dataset for the random forest regression model.
- Model Evaluation: This is determining overall performance of the model as well as its strengths and weaknesses. This was first done using validation dataset which is 20% of the original dataset. Before model evaluation, I performed hyper parameter tunning.

Validation data set contains data values that are new to the model therefore good for model evaluation. During evaluation using validation dataset, the accuracy of my random forest regression model turned out to be 98.586%. This shows that my model is neither over nor under fitted as it can present almost the same accuracy as with testing data after hyper parameter tunning which was 98.694% I therefore ended up creating a generalized crop yield prediction model using random forest regression.

I also plotted a scatter diagram to show the relationship between predicted and actual values and represent accuracy, the closer the scatter plots to the regression line the more the accuracy.

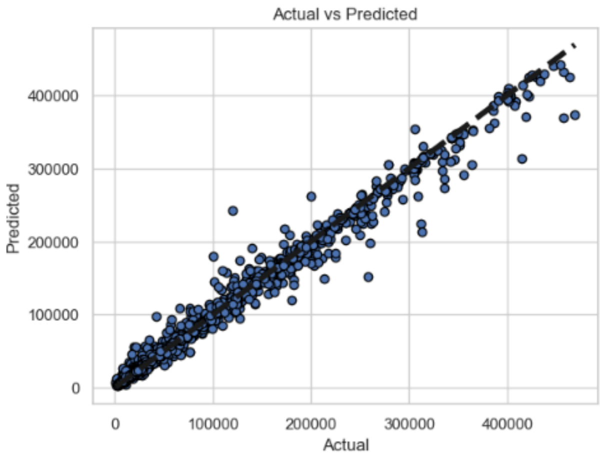


Fig. 1. Scatter diagram to show accuracy for the random forest regression model

A. Dataset Description

The dataset contains rainfall, temperature, pesticides, crop, year and country as independent features and yield as the dependent feature. Independent features determine dependent features. Average rainfall in mm, average temperature and quantity of pesticides measured in tonnes used in a year for a certain crop in a certain country is found in each column. It contains values for factors that affect crop yields plus corresponding yields. Through these examples, a model can be able to predict yields when presented with new input data. Through predicting yields, the problem of food insecurity can be solved. I chose the above dataset because it was easy to work with because it was a .csv file, it had relevant data required for the type of project I was trying to come up with, easy to understand, it had consistent and uniform data which hardly had any errors.

B. Data Preparation and Explanatory Data Analysis

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting data, data cleaning, data exploration and labeling raw data into a form suitable for machine learning algorithms. During data exploration, I calculated the mean, standard deviation , minimum and maximum of each column among others. I also found out the datatype of each column. Crop and country had object data type, year, yield and average rainfall had int datatype while pesticides and temperature had float datatype. Data cleaning is part of data preparation which involves spotting errors in our selected data set. Errors corrected during data cleaning include removing or replacing blanks (null cells), removing duplicates, removing excess space and standardizing text for clearer understanding.

During data cleaning, I renamed my columns for standardization, deleted the first column which was useless and looked for null values which were not found in my dataset. Explanatory Data Analysis is the process of performing initial investigations on data so as to discover patterns, to spot anomalies and check assumptions with the help of summary statistics and graphical presentations. This involves data cleaning, data visualisations, independence plots among others. For data visualization, I created a correlation heat map to show relationship between features. The darker the color, the lower the correlation and vice versa.

In the next step of data visualization, I created a boxplot to show the quantities of crops produced. Potatoes is the most produced crop.

I then used scatter plots to find outliers using crops. Outliers are data points that are significantly different from the rest of the dataset.

I created a bar graph to show ten countries with the highest yields. India had the highest number of yields.

For feature indepedence plots,I plotted bar graphs to show frequency of rainfall, temperature among others. The amount of rainfall generally decreased with time while the temperature increased with time.

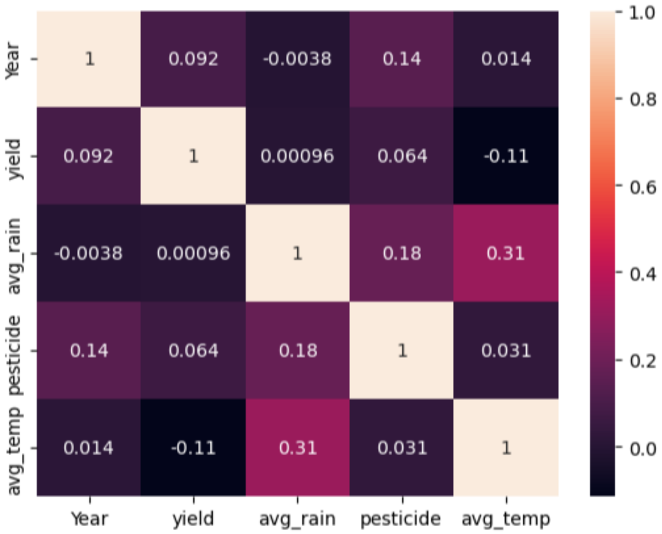
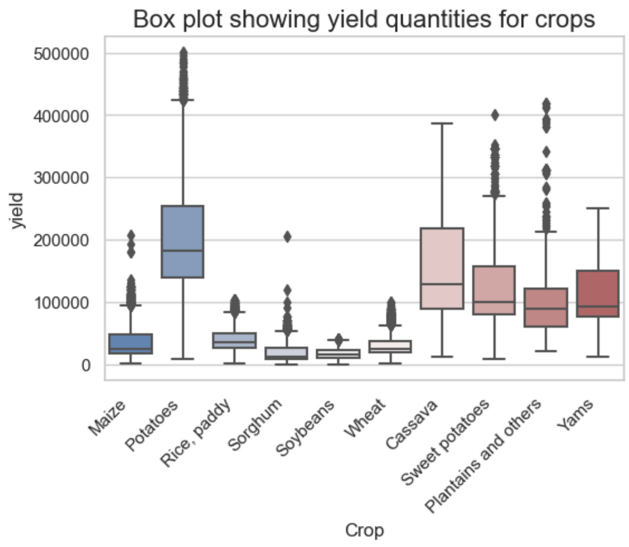


Fig. 2. Correlation heatmap

I converted the crop and country non numeric independent



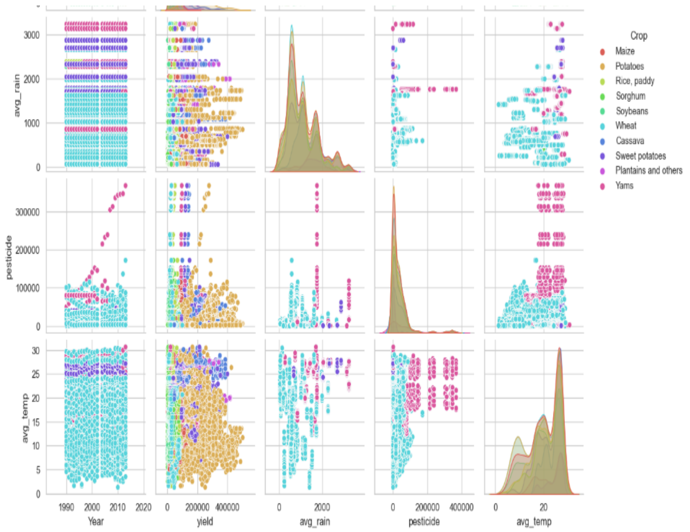
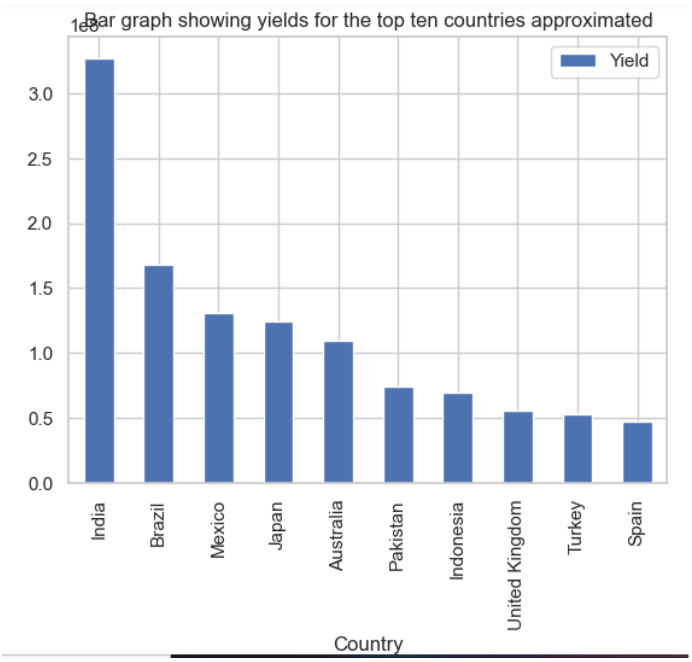


Fig. 3. Scatter plots to find outliers

features into numerals using the get_dummies method in order to be used during modeling. I also used the standard scaler function to rescale year, rainfall, temperature and pesticide values. This is because I wanted the features to be compared on the same scale of the algorithm. For the random forest regression model, I specified my independent features(x) and dependent feature(y).Independent features determine the dependent feature. I split my dataset into training data(70%), validation data(20%) and testing data(10%). For the k-means clustering model, I specified my independent features(x) only.

C. AI model selection and optimization

I implemented my project using my first supervised learning algorithm random forest regression. Random forest regression is a supervised learning method used for prediction in projects



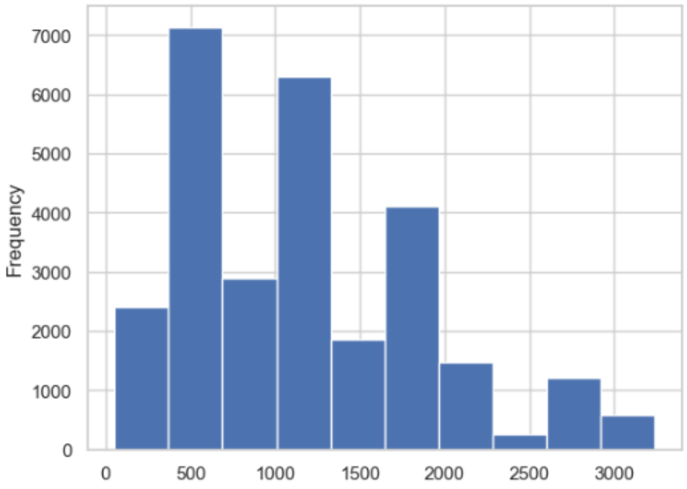


Fig. 4. feature indepedence plot showing frequency of rainfall

containing continuous data. Random forest is a binary tree based machine learning method which can be used for both classification (random forest classifier for data with two labels for example yes or no) and regression (random forest regressor for continuous values for example speed, weight, yields). This algorithm creates decision trees on different data samples and then predict the data from each subset, then by voting, the best solution for the system is found. Random forest uses the bagging method to train the data which increases the accuracy of the result. Bagging is an ensemble algorithm that fits multiple models on different subsets of a training dataset, then combines the predictions from all models. This used parameters like crop, year, country, temperature, rainfall and pesticides. After model testing, this model had an accuracy of 94.9698%. Model optimization is the process of iteratively improving the accuracy of a machine learning model through lowering the degree of error. This was achieved through hyper paramter tunning. This is the process of altering / finding the best combination of hyper parameters values to improve my model performance and avoid over and under fitting. Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up with. They are determined explicitly by the user, set before training of the model and the values do not change during modeling once set.

Hyper parameters for a random forest model include number of n_estimators, max_features, max_depth among others. Hyper parameter tunning is done through cross validation which is divided into randomized and grid search cross validation. Cross validation is a technique for evaluating machine learning models by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data. For my model, I decided to use randomized search for hyperparameter tunning. In randomized search, the algorithm selects and tests a random combination of hyper parameters. I was able to find the best combination of hyperparameters which I reused in my random forest regressor model and had my accuracy increase. After hyper parameter tunning, my new accuracy turned out to be 98.694%

For the k-means clustering model, I first implemented it with 10 clusters then performed hyper parameter tunning or evaluation to find the right number of clusters using elbow method. This plots a graph of distance of data points from centroid/center against a a certain range of clusters. An elbow point is found where after that point, the distance remains almost the same. This point provides the right number of clusters that the algorithm should form which turned out to be 16. In k-means clustering, the main aim is to reduce the reduce the distance between a datapoint and its center in each cluster. This algorithm I used here has only one hyper parameter n_clusters. It uses parameters like rainfall, temperature, pesticides, country, year and crop.

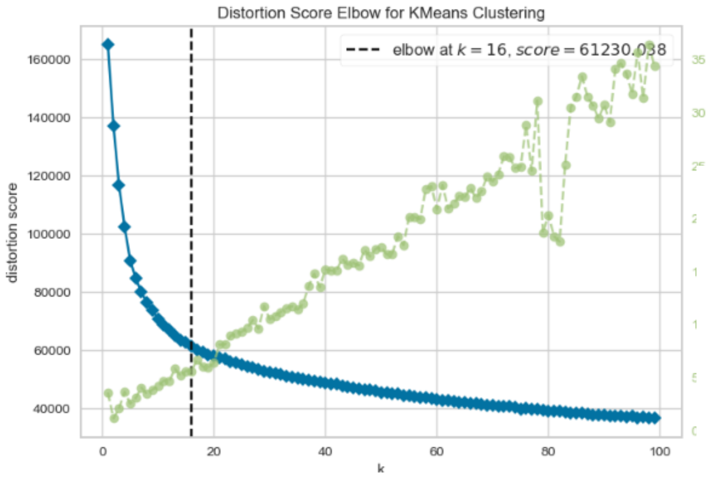


Fig. 5. Elbow plot for k-means clustering

D. AI MODEL ACCOUNTABILITY

This refers to the expectation that organisations or individuals will ensure the proper functioning throughout the lifecycle of the AI systems that they design, develop, operate or deploying accordance with their roles and applicable regulatory frameworks.

I applied accountability in my project through;

- Documentation of key decisions throughout the AI system lifecycle, understanding of the system and conducting auditing where justified.
- Understanding national and international laws, regulations and guidelines that my AI may have to work within.
- Setting clear goals and objectives for my model, putting in place risk management processes and reviewing of monitoring plans.
- Proper understanding of the data used during the model development and when in actual operation. Since data is the life blood of many AI systems, it is important to have a documentation about how it is being used.

  Machine learning explainability/interpretation is the ability of humans to capture relevant information from a model concerning relationships either contained in data or learned by the model. It prevents creation of black boxes. A black box is an artificial intelligence system whose inputs and operations are not visible to the user or any other interested party. A black box is impenetrable meaning a user cannot explain why the system behaves the way it behaves.

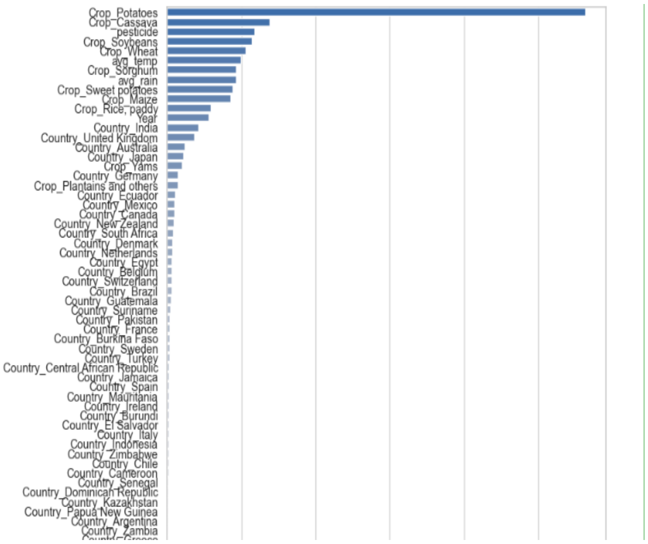


Fig. 6. Feature importance plot for random forest regression model

The machine learning interpretation technique I used was feature importance plot. Feature importance shows the

most important independent features whose values make a great impact on the yield values predicted. This depends on feature weights. It is measured by calculating the increase of the model's prediction error after making changes to the feature.
From the plot above, crop potatoes is the most important feature followed by cassava and pesticides among others.
Also, I used k-means clustering unsupervised learning method to understand relationships between my independent features.

V. Results and discussions

I calculated the mean absolute error and the mean squared error as evaluation metrics. Evaluation metrics are used to measure or test the quality of a machine learning model. The mean absolute error is the average measure of how close the predictions are to the final observed outcomes. Mean squared error is the average of the squares of the errors, which means the difference between the observed and predicted values. The mean squared error was bigger than the mean absolute error for the random forest regression model.
I also came up with a confusion matrix for the unique labels of the data. I calculated the precision, recall and f1-score of each of the labels using testing and validation data of the random forest regression model. Accuracy is the measure of how often the algorithm classifies a data point correctly. Recall or sensitivity is the ratio between the number of positive samples correctly classified as positive to all observations in actual class- yes or total number of all positive samples. Precision is the ratio between the number of positive samples correctly classified to the total number of samples classified as positive (either correctly or incorrectly). F1 score is defined as the harmonic mean of precision and recall.

VI. CONCLUSION AND FUTURE WORKS

I therefore came up with a generalized crop yield prediction model which is 98.586% accurate using random forest regression and can be used to make predictions for the ten most consumed crops in the world. I suggest that more independent features should be added in datasets for crop yield prediction models for more accuracy because quantities of crop yields depend on very many factors like soil pH, cultivated land among other. I am to also implement this project using other supervised learning algorithms like support vector regression.
I also came up with hypothesis for my research questions;

- Algorithms help in crop prediction by taking in inputs and sometimes its corresponding output and use mathematics and logic to tell outputs of new inputs.
- Pesticides increase productivity by reducing on losses from weeds, diseases and pests, therefore knowing values for quantity of pesticide used is as equally important as knowing the rainfall values as they both largely influence crop yields.
- There is increased food security due to crop yield prediction models because it encourages food availability.
- The bigger the dataset, the more accuracy is guaranteed due to more examples. Amount of input data(examples) should be ten times more than the dataset parameters(features) for a better accuracy.

A. DATASET AND PYTHON SOURCE CODE

Link to the dataset;
https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset
Link to the notebook;
http://localhost:8888/notebooks/Downloads/Untitled%20Folder/Explanator
Link to the youtube video;
https://youtu.be/gqdADsPvzR8

References

[1] https://assets.kpmg/content/dam/kpmg/xx/pdf/2022/07/modern-risk-management-for-ai-models.pdf

[2] https://www.javatpoint.com/goals-of-artificial-intelligence#: :text=The%20overall%20research%20goal%20of,broken%20down%20int

[3] https://www.elements-magazine.com/8-aims-and-objectives-of-artificial-intelligence/

[4] https://www.sciencedirect.com/science/article/pii/S0168169920302301

[5] https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608

[6] https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

[7] 5 steps to define the scope of a project (rockcontent.com)

[8] https://ieeexplore.ieee.org/document/9579843

[9] https://deepchecks.com/how-to-test-machine-learning-models/

[10] 3 ways to evaluate and improve machine learning models (techtarget.com)

[11] https://www.kaggle.com/general/107836

[12] https://towardsdatascience.com/elbow-method-is-not-sufficient-to-find-best-k-in-k-means-clustering-fc820da0631d