Section 1: Project Plan

Project Title:

A Machine Learning Approach to Optimize Pesticide Usage Without Compromising Yield

Research Questions:

1. What machine learning models most effectively predict optimal pesticide levels for maximizing crop yields based on historical data?
2. Are there significant differences in crop yields between countries using an Analysis of Variance (ANOVA)?

Objectives

1. To identify and compare multiple Machine Learning algorithms like Support Vector Machines, Linear Regression, XGBoost, and Decision Trees to compare their performance in predicting optimal pesticide usage for maximizing crop yields.
2. Evaluating the above-mentioned algorithms using multiple performance metrics such as RMSE (Root Mean Squared Error) and R-squared.
3. To develop a Python pre-processing pipeline to clean and pre-process data to enrich the quality of predictive modeling.
4. To test the trained model on a holdout dataset that mimics the real-world data.

Background and Summary

In the present age, the usage of pesticides is very common in agriculture. However, there are many pros and cons to using pesticides on plants and agriculture. While they control a lot of insects and pests thereby increasing productivity, they also pose a risk to the environment. Mainly polluting the water and land, also meddling with other species that don't cause harm to agriculture. To avoid such scenarios ideal usage of pesticides is a major requirement. To address this issue this research can be of great help. This project aims to tackle the prediction challenge of optimizing pesticide usage such that crop yields are maximized. Present methods are old don't take into account multiple factors and can't handle complex conditions. This is where Machine Learning can be used to use its potential to handle complex patterns (*Machine learning: learn, develop, and evolve from data sets*, 2021) in the data with multiple factors affecting a single variable. This project will explore many ML algorithms to funnel the best-performing and most effective algorithm for predicting the optimal usage of pesticides. This project will also delve into statistical analysis to understand the variability in crop yields across various factors using a statistical method called ANOVA (Singh, 2018).

Reference List

- Rajkumar, N. and Mukunthan, M.A., 2023, December. Efficient Crop Yield Analysis Prediction in Modern Agriculture Systems using Machine Learning Algorithm. In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)* (pp. 1-4). IEEE.

- Ranjani, et al., 2021. 'Crop yield prediction using machine learning algorithm', in *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, IEEE, pp. 611-616.

- Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A. and Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE Access*, *9*, pp.63406-63439.

Section 2: Task List and TimeLine

Task	10-Jun	25-Jun	10-Jul	25-Jul	09-Aug	24-Aug	29-Aug
Data Collection and Project Plan							
Literature Review							
Model Development							
Model Tuning							
Statistical Analysis							
Comparison and Results							
FPR							

In the above Gantt Chart, all the major modules responsible for the completion of this research report are given along with their expected time limits.

- Data Collection and Project Plan – Collect the required data and curate a project plan to systematically achieve set goals.
- Literature Review – Try to explore other research in the field and fill in the gap that the previous research is missing.
- Model Development – Using Python programming language to train multiple machine learning algorithms.
- Model Tuning – After training the models, take the best-performing models and tune them further to increase their efficiency.
- Statistical Analysis – Use statistical tests like ANOVA to answer the research questions posed in the above section.
- Comparison and Results – Make sense of the results and create tabular data with performance metrics.
- FPR – Develop a research report encompassing each step in this process with proofs and research results with a critical discussion.

Section 3: Data Management Plan

Dataset Overview

The dataset used in this research provides an overview of agricultural yields along with other factors such as pesticide usage, rainfall, and temperature data over the years for multiple countries. The crops this dataset features are rice, potatoes, wheat, soybeans, and more. This dataset contains almost 29k records spread across multiple years and

countries along with 7 columns that contain geographical, rainfall, and other features. Below is the image that shows the sample records in the data.

	Area	Item	Year	hg/ha_yiel	average_r	pesticides_	avg_temp
0	Albania	Maize	1990	36613	1485	121	16.37
1	Albania	Potatoes	1990	66667	1485	121	16.37
2	Albania	Rice, padd	1990	23333	1485	121	16.37
3	Albania	Sorghum	1990	12500	1485	121	16.37
4	Albania	Soybeans	1990	7000	1485	121	16.37
5	Albania	Wheat	1990	30197	1485	121	16.37
6	Albania	Maize	1991	29068	1485	121	15.36
7	Albania	Potatoes	1991	77818	1485	121	15.36
8	Albania	Rice, padd	1991	28538	1485	121	15.36
9	Albania	Sorghum	1991	6667	1485	121	15.36

The data is taken from an open-source dataset website called Kaggle. All the datasets available on this website are free to use for non-commercial purposes. Hence, ethical approval from any source is not needed.

Data Collection

Source: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=yield_df.csv

Document control

All of the data files will be kept in the GitHub folder. Click on this link to go to the repository: https://github.com/kv22aac/final-project-crop-yield-prediction

Data ethics:

1. Does the data meet GDPR requirements? - Yes
2. Does the project conform to UH ethical policies? – Yes
3. Do you have permission to use the data for your proposed research project? - Yes
4. Are you assured that the data was collected ethically (i.e. by the original people who gathered/collected/ collated/made the data)? – Yes

Security, and Storage

The data will be shared through GitHub.

References

Singh, G. (2018) *What is Analysis of Variance (ANOVA)?*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/ (Accessed: June 12, 2024).

Machine learning: learn, develop, and evolve from data sets (2021) *Brunel*. Available at: https://www.brunel.net/en/management-guide/machine-learning (Accessed: June 12, 2024)