CS486 Final Project Proposal
Comparison of Different Machine Learning Algorithms for Crop Fields Classification

Mary Wang(20816042)
Gloria Song(20818442)
University of Waterloo

1. Introduction

Global climate change refers to the long-term changes in temperature, precipitation, wind patterns, and other measures of climate that occur over several decades or longer. The Earth's climate is changing, largely due to human activities, such as the burning of fossil fuels (coal, oil, and natural gas), deforestation, and agriculture. These activities are releasing large amounts of greenhouse gasses into the atmosphere, causing the Earth's temperature to rise and leading to many other impacts on the climate system.

The effect of global climate change is widespread including rising temperatures, melting of glaciers and polar ice caps, sea level rise, and more frequent and intense extreme weather events, such as hurricanes, droughts, and floods. These impacts, in particular, have a direct effect on the crop yields as well as the overall productivity of agriculture.

Specifically speaking, rice is a staple food for more than four billion people and the livelihood for a fifth of the world’s population. But both the amount of land available for rice and the yield growth rates are in decline with irrigated rice yields in developing countries forecast to decrease by 15% due to climate change. As one of the world’s leading rice producers, Vietnam is losing an estimated US$10b in 2020 to climate change according to the World Bank.

Although world hunger cannot be solved in one solution, we can start by examining one single crop in one particular area. Therefore, in this project, we will use radar and optical satellite data from Microsoft’s Planetary Computer to build a tool to identify rice crops in the An Giang province of Vietnam.

2. Related works

Singh et al.(2022) classified the land cover and the crops type using Sentinel-2A open data. The study focused on a region in India where Rabi post-monsoon crops such as wheat and gram are primarily grown in. They derived the Soil Adjusted Vegetation Index(SAVI) product from the time-series Sentinel-2A images of this region. The SAVI data available are used to derive its component using principal component analysis(PCA). Then the non-correlated and higher variable PC layers can be used for RF machine learning training.They assess the accuracy of the model by generating stratified random data points and compare the predicted results with corresponding Google Earth Pro. In the end, they concluded that RF machine learning algorithm and time-series images of finer spatial resolution together do help create more precise and accurate mapping of cropland, though the data preprocessing has become particularly important.

Ndikumana et al.(2018) evaluated the potential of high spatial and temporal resolution Sentinel-1 remote sensing data to map different agricultural land covers in Camargue, France. In order to do this, they started by considering supervised machine learning algorithms that are commonly used in agricultural land cover classification, such as K nearest neighbor (KNN), random forest (RF) and support vector machine (SVM). Besides, they also proposed that recurrent neural network (RNN) which offers models to explicitly manage temporal dependencies among data can be suitable for the mining of multitemporal SAR Sentinel-1 data. They compared the results from RNN-based classification approaches with those from standard machine learning approaches and concluded that although the results are promising for both classic approaches and advanced deep RNN technique, the latter offered equally great results on all classes, while KNN, RF and SVM showed different behaviors on different classes.

I.M.D. Rosa et al.(2016) implemented six machine learning algorithms including random forests (RF), support vector machine (SVM), artificial neural networks (NN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and decision trees (DT) to classify birds using radar data collected from two locations in Portugal.However, instead of performing a multi-class classification, they trained and tested all possible hierarchical binary classifications. By comparing the area under the receiver operating characteristic (AUC), the overall accuracy of each classifier with its statistical significance, the sensitivity, true positive (TP) rate and specificity and true negative (TN) rate, they concluded that all models had greater difficulty separating among birds species than separating birds from wind turbines.

Since there are too many factors to predict yield, Ansarifar et al.(2021) presents a new predictive model (referred to as the interaction regression model) for crop yield prediction. The core of this model lies in a combinatorial optimization algorithm, which has three salient properties. First, it outperformed state-of-the-art machine learning algorithms with respect to prediction accuracy in a comprehensive case study. Second, it can identify multiple scenarios that combine different factors and develop testable hypotheses. Third, it was able to explain the contributions of weather, soil, management, and their interactions to crop yield. This paper provides a complete experimental procedure, including the method of constructing the model and the interpretation of the results.

Khaki et al.(2020) presents a deep learning framework using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to build a CNN-RNN model to crop yield prediction based on environmental data and management practices. "The CNN part of the model was designed to capture the internal temporal dependencies of weather data and the spatial dependencies of soil data measured at different depths underground. The RNN part of the model was designed to capture the increasing trend of crop yield over years due to continuous improvement in plant breeding and management practices. The performance of the model was relatively sensitive to many variables, including weather, soil, and management." This article builds a model based on CNN-RNN, conducts experiments, and compares the result with random forest (RF), deep fully connected neural networks (DFNN), and LASSO. This article provides a large number of data sources and cites many related articles for reference.

The main purpose of this article written by Iniyan et al.(2023) is to let machine learning models make predictions to help farmers decide which crops to plant and improve yields. The paper investigates a variety of methods for predicting crop yields using a variety of soil and environmental variables. They mainly focus on gathering an agricultural csv dataset for Crop yield prediction, data pre-processing, training models, testing, and comparing results. Training models include Multiple Linear regression, Decision tree Regression, Gradient Boosting, Elastic Net, Lasso, Ridge, Partial Least Squares Regression and feature engineering-based LSTM models. LSTM is the most efficient model based on the comparison and validation of training data and test data.

3. Methodology

3.1 Data Set

We are going to use a Crop Location data set in which there is information about the latitude and longitude of 600 crop fields and the class of each field(Rice or Non rice).

3.2 Data Processing

We will randomly divide the data set into two subsets. One for training and one for testing. The training set will contain 80% of the data while the rest 20% are in the testing set. There is no overlap between the training set and testing set.

3.3 Algorithms

We are going to try six different machine learning algorithms for the classification of crop fields.

- **K Nearest Neighbors(KNN):** We are going to use the scikit-learn implementation to fit a KNN. Use cross-validation to find the best hyperparameter k. Push KNN to its maximum performance using bagging
- **Logistic Regression:** We are going to use the scikit-learn implementation to fit a multivariate logistic regression with regulation. We will use cross-validation to find the best hyperparameter λ.
- **Random Forest(RF):** We are going to use the scikit-learn implementation to fit a RF. We will start by manually setting a grid of parameters. Then use GridSearchCV or RandomizedSearchCV to obtain the best parameters for the model.
- **Decision Tree(DF):** We are going to use the scikit-learn implementation to fit a DF. We will start by manually setting a grid of parameters. Then use GridSearchCV or RandomizedSearchCV to obtain the best parameters for the model.
- **Support Vector Machine(SVM):** We are going to use the scikit-learn implementation to fit a SVM. We will start by manually setting a grid of parameters. Then use GridSearchCV or RandomizedSearchCV to obtain the best parameters for the model.
- **Neural Network(NN):**
    - Step 1: Define the neural network structure (# of input units, # of hidden units, etc).
    - Step 2: Initialize the model's parameters
    - Step 3: Loop: Implement forward propagation, Compute loss, Implement backward propagation to get the gradients, Update parameters (gradient descent)

- We will merge step 1-3 into one function called nn_model(). Once we've built nn_model() and learnt the right parameters, we will fit the model and make predictions on testing data.

3.4 Train and Test the Model

Since we have limited data, then we can use **cross validation** (CV) to test the effectiveness of each model. The main idea is to use the K-Folds Cross Validation approach. Split the training set randomly into K folds. Then fit the model using the K-1 folds, validate the model using the remaining fold and record the values/errors. Repeat this process until every fold serves as the test set. Last, take the average of all the recorded values to see the performance of this model. We will start by choosing K=5.

3.5 Model Assessment

We will use the following indexes to evaluate the performance of a model.

Index	Definition
Accuracy	Percentage of the number of correct predictions out of the number of total predictions.
Precision	Percentage of positive instances out of the total predicted positive instances.
Sensitivity	Percentage of positive instances out of the total actual positive instances.
Specificity	Percentage of negative instances out of the total actual negative instances.
F1 score	$\frac{2*Precision*Recall}{Precision+Recall}$ It is the harmonic mean of precision and recall. The higher the F1 score, the better.

Mean-squared error	The average of squared differences between the predicted output and the true output. (used for regression model)
ROC curve	It is a graph where x-axis is the 1-specificity and y-axis is the sensitivity. AUC is the area under the curve, the higher its numerical value the better.

3.6 Conclusion/Discussion

Our goal is to find the most accurate model. Using the result from the model assessment part to choose the best performing/accurate model. The best models can accurately predict whether a crop will be produced at a given latitude and longitude. We can also compare other models with the best model to discuss model performance when the amount of data increases or the number of variables increases, etc.

References

Ansarifar, J., Wang, L., & Archontoulis, S. (2021). An interaction regression model for crop yield prediction. Scientific Reports, 11(1). doi: 10.1038/s41598-021-97221-7 https://www.nature.com/articles/s41598-021-97221-7

A study on crop yield forecasting using classification techniques. (2023). Retrieved 16 February 2023, from https://ieeexplore.ieee.org/document/7725357

Evaluating a Machine Learning Model. (2019). Retrieved 16 February 2023, from https://medium.com/@skyl/evaluating-a-machine-learning-model-7cab1f597046

Iniyan, S., Akhil Varma, V., & Teja Naidu, C. (2023). Crop yield prediction using machine learning techniques. Advances In Engineering Software, 175, 103326. doi: 10.1016/j.advengsoft.2022.103326 https://www.sciencedirect.com/science/article/pii/S0965997822002277

Khaki, S., Wang, L., & Archontoulis, S. (2020). A CNN-RNN Framework for Crop Yield Prediction. Frontiers In Plant Science, 10. doi: 10.3389/fpls.2019.01750 https://www.frontiersin.org/articles/10.3389/fpls.2019.01750/full

Various ways to evaluate a machine learning models performance. (2019). Retrieved 16 February 2023, from https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Why and how to Cross Validate a Model?. (2020). Retrieved 16 February 2023, from https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f