BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI HYDERABAD CAMPUS (April 2025)

BITS F464: Machine Learning

A Project Report
On

IMPROVING AGRICULTURAL PRODUCTIVITY BY APPLYING MACHINE LEARNING ON FARMING DATA

BY

Varun Ravichandran
Abhishek Verma
Divyansh Rungta

Under the supervision of
Prof. Paresh Saxena

ABSTRACT

The research paper we choose demonstrates the procedure to apply machine learning to forecast crop yield in South India. The purpose is to strengthen farm planning and optimize the use of resources. The authors used an integrated dataset of weather, soil, and farm variables from five states in India. They compared and evaluated different regression models, i.e., Linear Regression, Random Forest, Gradient Boosting, and Extra Trees, based on measures such as R-squared, MAE, and RMSE. The Extra Trees Regressor outperformed all the others with an R-squared of 0.9615. The results point out how machine learning can be used to make informed agricultural decisions.

INTRODUCTION

India is an agricultural nation. The agricultural sector is a major contributor to the nation's economy. The authors recognize the significance of accurately forecasting crop yields and therefore they would like to apply machine learning (ML) models to improve the understanding of the key drivers of crop productivity. Their overall objective is to produce models that not only accurately forecast crop yields but also identify the optimal factors such as rainfall, soil composition that significantly influence agricultural outcomes. These parameters have complex relations and patterns and so they train models to identify these patterns.

We chose this topic because accurate crop yield prediction is vital for ensuring food security and supporting farmers in making informed decisions. By leveraging ML, we aim to enhance agricultural planning and productivity in a data-driven and sustainable way.

METHODOLOGY

1. Data collection : Once the required data is identified, we proceed to identify the data sources to collect the data. The data is from government open access data sources.

   Data sources links: http://data.icrisat.org/dld/src/crops.html

2. Data pre-processing : The data set that we are using contains information from all over India. We split the data into south India and the rest of India. We have dropped the state column because it is redundant, all the information necessary can be retrieved from the district column. One of k coding formats has been followed. We have applied standard scalar on the irrigation column to restrict values between 0 and 1. Redundant data (empty and negative values) have been replaced by 0.

3. Model Training and model evaluation : The model training phase involved implementing a variety of regression algorithms to predict crop yield based on a curated set of features including soil nutrient composition, annual rainfall, and fertilizer usage. For each of the selected crops—**RICE, RABI SORGHUM, SORGHUM, SUGARCANE, and COTTON**—the corresponding yield value was extracted as the target variable. We have trained our models only on the south Indian data set. This is the performance matrix that we achieved.

Overall Preformance Metrics (SOUTH INDIA)

	MAE	MSE	RMSE	R-squared Score
Linear Regressor	0.435927	0.515096	0.698755	0.540075
Gradient Boosting Regressor	0.371607	0.348491	0.580349	0.682879
Random Forest Regressor	0.278123	0.245925	0.486717	0.774794
K-nearest Neighbours Regressor	0.301710	0.265821	0.508456	0.753226
Descision Tree Regressor	0.350497	0.403429	0.617897	0.639083
Bagging Regressor	0.292862	0.272960	0.511172	0.751750
Extra Tree Regressor	0.353389	0.410153	0.630779	0.623151
Bayesian Ridge Regressor	0.435328	0.516677	0.699552	0.538967
Ridge Regressor	0.435919	0.515101	0.698756	0.540073

To broaden our scope we considered the entire Indian data set in hope of improving our models. Here is the performance matrix we achieve on doing so.

Overall Preformance Metrics (ENTIRE INDIA)

	MAE	MSE	RMSE	R-squared Score
Linear Regressor	0.381482	0.447235	0.660523	0.563874
Gradient Boosting Regressor	0.414967	0.436679	0.656303	0.575270
Random Forest Regressor	0.240972	0.260467	0.502998	0.745736
K-nearest Neighbours Regressor	0.261050	0.290240	0.530161	0.715999
Descision Tree Regressor	0.294217	0.407256	0.631435	0.603550
Bagging Regressor	0.250598	0.277456	0.519412	0.729174
Extra Tree Regressor	0.294472	0.388565	0.616785	0.622140
Bayesian Ridge Regressor	0.380257	0.446537	0.660024	0.564575
Ridge Regressor	0.381459	0.447212	0.660505	0.563896

Comparing the models based on MAE

Random Forest Regressor < Bagging Regressor < K-nearest Neighbours Regressor < Decision Tree Regressor < Extra Tree Regressor < Bayesian Ridge Regressor < Ridge Regressor < Linear Regressor < Gradient Boosting Regressor

Comparing various models based on RMSE

Random Forest Regressor < Bagging Regressor < K-Nearest Neighbours Regressor < Extra Tree Regressor < Decision Tree Regressor < Gradient Boosting Regressor < Bayesian Ridge Regressor < Ridge Regressor < Linear Regressor

By looking at MAE and RMSE we conclude that Random Forest Regressor provides the best prediction.

IMPROVISATIONS

We merged fertilizer and annual rainfall data to the pre existing data. Here is the performance matrix we got on doing so

Overall Preformance Metrics

	MAE	MSE	RMSE	R-squared Score
Linear Regressor	0.357836	0.390599	0.616196	0.599110
Gradient Boosting Regressor	0.383927	0.367564	0.598900	0.619916
Random Forest Regressor	0.246033	0.254973	0.494363	0.738773
K-nearest Neighbours Regressor	0.260473	0.282907	0.520656	0.710154
Descision Tree Regressor	0.321784	0.483752	0.681676	0.503352
Bagging Regressor	0.257596	0.270657	0.510038	0.722108
Extra Tree Regressor	0.324177	0.478021	0.675366	0.510186
Bayesian Ridge Regressor	0.356328	0.390411	0.616001	0.599329
Ridge Regressor	0.357813	0.390586	0.616184	0.599124

RMSE has improved for the following models:-

Linear Regressor, Gradient Boosting Regressor, Random Forest Regressor, K-nearest Neighbours Regressor, Bagging Regressor, Bayesian Ridge Regressor, Ridge Regressor

We want to filter out the models with high bias therefore we perform bias variance decomposition.

	Bias²	Variance	Total Error (MSE)
Linear Regressor	0.392261	0.009951	0.402212
Gradient Boosting Regressor	0.368691	0.013423	0.382114
Random Forest Regressor	0.250401	0.039191	0.289592
K-nearest Neighbours Regressor	0.274048	0.050336	0.324384
Descision Tree Regressor	0.274630	0.218935	0.493565
Bagging Regressor	0.250356	0.059843	0.310199
Extra Tree Regressor	0.256786	0.230883	0.487669
Bayesian Ridge Regressor	0.392020	0.009432	0.401452
Ridge Regressor	0.392250	0.009942	0.402192

Comparison of various models based on Bias

Bagging Regressor < Random Forest Regressor < Extra Tree Regressor < K-nearest Neighbours Regressor < Decision Tree Regressor < Gradient Boosting Regressor < Bayesian Ridge Regressor < Ridge Regressor < Linear Regressor

After looking at the various comparisons of models we can see that **tree based models always perform better**.

IN PURSUIT OF HIGHER ACCURACY WE DECIDED TO TRY STACKING

In stacking we consider the models with low bias (decision tree, random forest, bagging regressor, extra trees regressor and KNN regressor). This is what we got

	MAE	MSE	RMSE	R-squared Score
Stacking Regressor	0.244803	0.250563	0.489778	0.743538

IN ANOTHER ATTEMPT TO REDUCE VARIANCE WE ALSO TRIED XGBOOST

Overall Preformance Metrics when using XGBoost

	MAE	MSE	RMSE	R-squared Score
Ridge Regressor	0.255761	0.276818	0.513394	0.71732

CODE

Link To Code: Machine_Learning_Project.ipynb
Training data.csv: Training data.csv
Annual_rainfall_data.csv: Annual_rainfall_data.csv
Fertilizer_usage_data.csv: Fertilizer.csv

CONCLUSION

This project demonstrated how machine learning can play a vital role in modern agriculture by accurately predicting crop yields using environmental and soil-based parameters. By integrating multiple datasets and experimenting with various regression models, we were able to identify the strengths and limitations of different algorithms in capturing the complex relationships between factors like soil nutrients, rainfall, and fertilizer use.

We have also confirmed that **Tree based models** give the best prediction with **Random Forest Regressor** with the lowest bias

Overall, this project not only improved our understanding of machine learning applications in agriculture but also highlighted how data-driven insights can support farmers and policymakers in making more informed, efficient, and sustainable decisions