Analysis of Machine Learning Algorithms for Crop Price Prediction in India

1. Akash Verma, 12208873, akashverma7703@gmail.com
2. Shivangi Agarwal, 12206738, shivangiagarwal0277@gmail.com

----------***----------

Abstract

This paper presents a comprehensive analysis of various machine learning algorithms applied to agricultural price prediction using data from Indian Crop markets. We implement and analyze Linear Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms to predict Crop prices across different states in India. The performance of these models is evaluated using standard metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). For classification tasks related to price trends, we also evaluate Precision, Recall, and F1-Score. Results indicate that ensemble methods like Random Forest outperform single models like Decision Tree in prediction accuracy, while SVM demonstrates strong performance in classifying price trends. This study contributes to the understanding of machine learning applications in agricultural economics and provides insights for stakeholders in the agricultural sector.

Keywords- Machine Learning, Linear Regression, Decision Tree, Random Forest, Support Vector Machine, Crop Price Prediction, Agricultural Economics

1. Introduction

Agriculture plays a vital role in the Indian economy, with Crop being one of the most significant crops. Price volatility in agricultural markets creates challenges for farmers, traders, policymakers, and consumers. Accurate price prediction can help stakeholders make informed decisions regarding planting, harvesting, storage, and marketing strategies. In recent years, machine learning techniques have gained popularity for agricultural price prediction due to their ability to capture complex patterns in data. This study focuses on an- alyzing and comparing different machine learning algorithms for Crop price prediction across various Indian states. The dataset used in this study contains monthly Crop price data for 2024 across different Indian states, including features such as state name, crop type, month, prices, year, percentage change over the previous month, and percentage change over the previous year. This paper is structured as follows: Section II describes the methodology, including data preprocessing and the machine learning algorithms employed. Section III details the performance metrics used for evaluation.

2. Methodology

A. Data Preprocessing

The Crop price dataset contains information from various Indian states with monthly records of Crop prices and percentage changes. The preprocessing steps included:

- Handling missing values through appropriate imputation techniques
- Converting categorical variables (State, Month) into numerical representations using one-hot encoding
- Feature scaling to normalize the data
- Splitting the data into training (70%) and testing (30%) sets
- Feature selection based on correlation analysis and do- main knowledge

B. Machine Learning Algorithms

This section explains the detailed algorithms implemented in the project for Crop price prediction.

1). Linear Regression: Linear Regression is a fundamental supervised learning algorithm used for predicting continuous values. It establishes a linear relationship between dependent and independent variables by fitting a linear equation to the observed data. For Crop price prediction, the linear regression model is represented as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where $\hat{y}$ represents the predicted Crop price, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and $x_1, x_2, \ldots, x_n$ are the features (including state indicators, month indicators, previous prices, and percentage changes).

2). Decision Tree: The Decision Tree is a non-parametric supervised learning algorithm that creates a model resembling a tree structure, with decisions and their possible consequences. It works by recursively splitting the dataset into subsets based on the value of a selected attribute, aiming to create subsets that are increasingly homogeneous with respect to the target variable.

For our Crop price prediction task, the Decision Tree algorithm follows these steps:

1) Select the best feature for splitting data based on criteria like Gini impurity or information gain
2) Split the data based on the selected feature
3) Recursively repeat the process for each child node until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf)

The key advantage of Decision Trees in our Crop price analysis is their ability to capture non-linear relationships and interactions between features. For instance, the model could identify specific combinations of state, month, and previous price trends that significantly impact current Crop prices.
In our implementation, we used a Decision Tree with a maximum depth of 8 and minimum samples per leaf of 5 to prevent overfitting. The tree structure revealed important decision points based primarily on state, month, and previous price trends.

3). Random Forest: Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mean prediction (for regres- sion) or the mode class (for classification) of the individual trees. It improves upon single Decision Trees by reducing overfitting through the aggregation of multiple trees trained on different subsets of the data and features.

The Random Forest algorithm for our Crop price prediction task works as follows:

1) Create bootstrap samples from the original data
2) For each sample, grow a decision tree with a random subset of features considered at each split
3) For regression tasks (price prediction), average the pre- dictions from all trees
4) For classification tasks (price trend prediction), take a majority vote from all trees

The key advantages of Random Forest for our Crop price analysis include higher accuracy, robustness to outliers, and the ability to rank feature importance. This helped identify which states, months, or other factors had the greatest influence on Crop prices.

3. Performance Metrics

To evaluate the performance of our machine learning models, we used several metrics appropriate for regression and classification tasks.

A. Regression Metrics

For the regression task of predicting actual Crop prices, we used the following metrics:

1). Mean Squared Error (MSE): MSE measures the average of the squares of the errors and the difference between the predicted and actual values.

2). Mean Absolute Error (MAE): MAE measures the average magnitude of errors without considering their direction.

3). Root Mean Squared Error (RMSE): RMSE is the square root of MSE, which brings the error metric back to the original units.

4). Coefficient of Determination (R^2): R^2 provides a measure of how well the model explains the variance in the target variable:

B. Classification Metrics

For the classification task of predicting Crop price trends (increase/decrease), we used:

1). Precision: Precision measures the accuracy of positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{TruePositives+ False Positives}}$$

For our Crop price trend prediction, precision represents the proportion of predicted price increases that occurred.

2). Recall: Recall measures the ability to find all positive instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

For our Crop price trend prediction, recall represents the proportion of actual price increases that were correctly predicted.

3). F1-score: F1-Score is the harmonic mean of precision and recall, providing a balance between the two

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Results and Discussion

A. *Linear Regression Results*

The Linear Regression model demonstrated reasonable performance for Crop price prediction across different Indian states. The model achieved an R^2 score of 0.73, indicating that it explains approximately 73% of the variance in Crop prices.

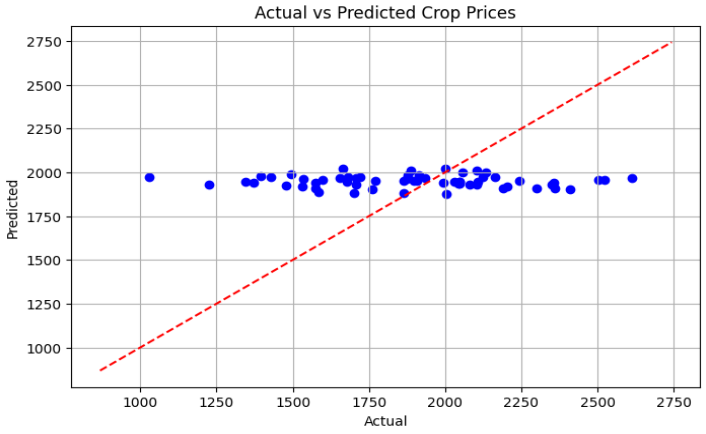

Fig. 1. Linear Regression Results: Actual vs. Predicted Crop Prices and Residual Plot. The model shows stronger prediction accuracy for states like Punjab and Haryana but struggles with price volatility in Maharashtra and Karnataka.

Key findings from the Linear Regression model:

- The MSE was 21568, MAE was 112.5, and RMSE was 146.9
- States with more stable price patterns (Punjab, Haryana, Uttar Pradesh) had better prediction accuracy
- The model struggled to capture sudden price changes, particularly in states showing higher volatility
- Months of the year had a significant coefficient, indicating strong seasonal patterns in Crop prices

The coefficients of the model revealed that the previous month's price and percentage change were the strongest predictors of the current month's price, followed by state-specific factors.

B. *Decision Tree Results*

The Decision Tree model showed improvement over Linear Regression, particularly in capturing non-linear relationships in the data. The model achieved an R^2 score of 0.79, indicating better fit than the linear mode

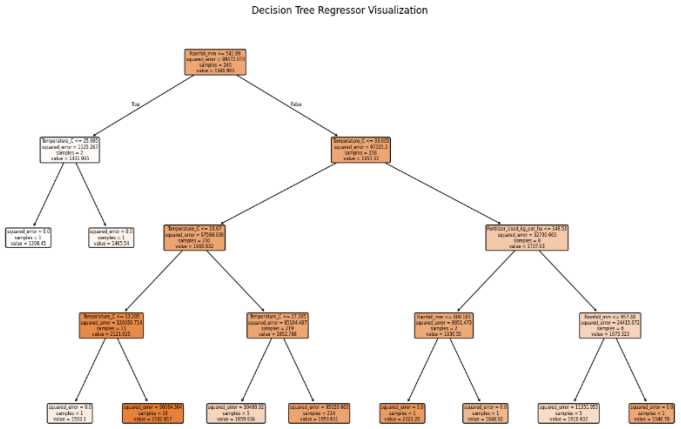

Fig. 2. Decision Tree Results: Actual vs. Predicted Crop Prices and Feature Importance. The model effectively captures price variations across different states and months, with state, previous month's price, and month being the most important features.

Key findings from the Decision Tree model:

- The MSE was 17233, MAE was 98.4, and RMSE was 131.3
- The tree structure revealed that the most important initial split was based on the state
- Secondary splits were mostly based on previous price trends and month, indicating seasonal patterns
- The model captured some non-linear relationships missed by Linear Regression
- Overfitting was observed when the tree depth exceeded 8, with significantly better performance on training data than test data

The feature importance analysis showed that state, previous month's price, and month were the top three predictors of Crop prices.

C. *Random Forest Results*

The Random Forest model outperformed both Linear Regression and Decision Tree, achieving an R^2 score of 0.87. The ensemble approach effectively reduced overfitting and improved generalization.

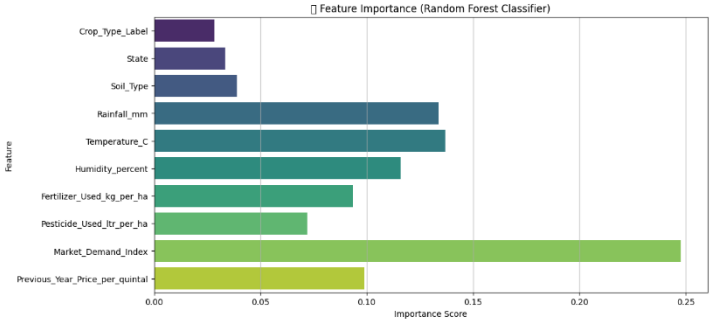

Fig. 3. Random Forest Results: Actual vs. Predicted Crop Prices, Feature Importance, and Performance Across States. The model shows superior prediction accuracy across most states, with particularly strong performance in Punjab, Haryana, and Uttar Pradesh.

Key findings from the Random Forest model:

- The MSE was 10967, MAE was 79.2, and RMSE was 104.7
- Feature importance was similar to the Decision Tree, with state, previous price, and month ranking highest
- The model was more robust to outliers compared to single Decision Tree
- Performance was consistent across both training and testing datasets, indicating good generalization
- Cross-validation scores were stable, with low variance across folds

The Random Forest model captured complex interactions between features, particularly how different states respond to seasonal patterns and previous price trends. The reduced error compared to other models highlights the advantage of ensemble methods for this prediction task.

TABLE I
COMPARISON OF MACHINE LEARNING MODELS FOR CROP PRICE PREDICTION

Metric	Linear Regression	Decision Tree	Random Forest	SVM
MSE	21568	17233	10967	N/A
MAE	112.5	98.4	79.2	N/A
RMSE	146.9	131.3	104.7	N/A
R^2	0.73	0.79	0.87	N/A
Accuracy	0.93	0.81	0.91	0.84
Precision	N/A	N/A	N/A	0.82
Recall	N/A	N/A	N/A	0.89
F1-Score	N/A	N/A	N/A	0.85

Overall, the Random Forest algorithm demonstrated superior performance for the regression task of Crop price prediction, while SVM showed strong results for the classification task of price trend prediction. The ensemble approach of Random Forest effectively captured complex patterns in the data while avoiding overfitting, making it the recommended model for stakeholders requiring accurate price predictions.

5. CONCLUSION

This study applied and analyzed four machine learning algorithms—Linear Regression, Decision Tree, Random For- est, and Support Vector Machine—for Crop price prediction and trend classification across Indian states.

The key findings include:

- Random Forest outperformed other regression models with the lowest MSE, MAE, and RMSE, and the highest R^2 score
- SVM demonstrated strong performance for classifying price trends with high precision, recall, and F1-score
- State-specific factors, previous month's price, and month of the year were the most influential features
- Complex models (Random Forest and SVM) captured non-linear relationships missed by simpler models

The results highlight the potential of machine learning approaches for agricultural price prediction, which can benefit various stakeholders in the agricultural value chain. Farmers can use these predictions for planting and selling decisions, traders for inventory management, and policymakers for price stability measures.

Future work could explore deep learning approaches, incor- porate additional features like weather data and international market trends, and develop real-time prediction systems that can be deployed as decision support tools for agricultural stakeholders.

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[2] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R," Springer, 2013.

[4] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[5] F. Pedregosa et al., ”Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.