Analysis of Yield Drivers in Crop Production using Machine Learning and Exploratory Data Analysis

**VINAY KUMAR REEDY KADAPANA(W9610384)**

***MACHINE LEARNING-DATA SCIENCE***

# ABSTRACT

**The problems produced by climate change, a growing world population, decreasing water resources, and increasing soil erosion, made it crucial to predict crop production to ensure food security. The growth and final grain production of crops are impacted by several environmental factors like temperature, and water availability in a non-linear way. Machine learning methods can be applied to these non-linear relationships between production and its covariates. Considering the fundamental components that lead to predicting losses or gains is pivotal for production forecasting. To deal with this problem, we investigated how to maintain interpretability while benefiting from the improved predictive performance of Machine Learning methods. Four machine learning models are applied and then analyses of their learned features are done. The random forest model proved superior to the other models tested (ridge regression, decision tree, and artificial neural network) and facilitated the interpretation of attributes and processes contributing to production variability.**

***Keywords—****Crop production in India, Ridge Regression, Decision Tree, Random Forest, Artificial Neural Network.*

# INTRODUCTION

In precision agriculture, estimating crop yield is a challenging task, and numerous machine-learning models have been suggested and verified. Since crop production is influenced by a wide range of variables, including climate, fertiliser use, and seed variety, this problem calls for multiple datasets [1]. This implies that predicting agricultural yields is a difficult undertaking that involves a number of intricate steps. While current crop production prediction models can reasonably estimate actual production, there is still room for improvement in the production prediction performance [2].

A variety of variables [3] that vary for each square meter of land affect crop productivity. These factors impact the production and growth of crops. The authors [4] worked with agronomists to increase crop production and reduce soil degradation in cultivated grasslands.

## *About the Dataset*

The Dataset of India Agriculture is an extensive collection of information documenting crop production over a period of several years. The dataset consists of 246,091 observations with 7 attributes including state names, district names, crop years, seasons, crop types, cropped areas, and crop production figures. The data provides useful insights related to trends in crop production across all of India. The major aim of the dataset is to apply different machine-learning techniques in order to predict crop production based on the given attributes. This dataset is quite a significant resource for researchers, stakeholders, and policymakers interested in the Indian agricultural sector.

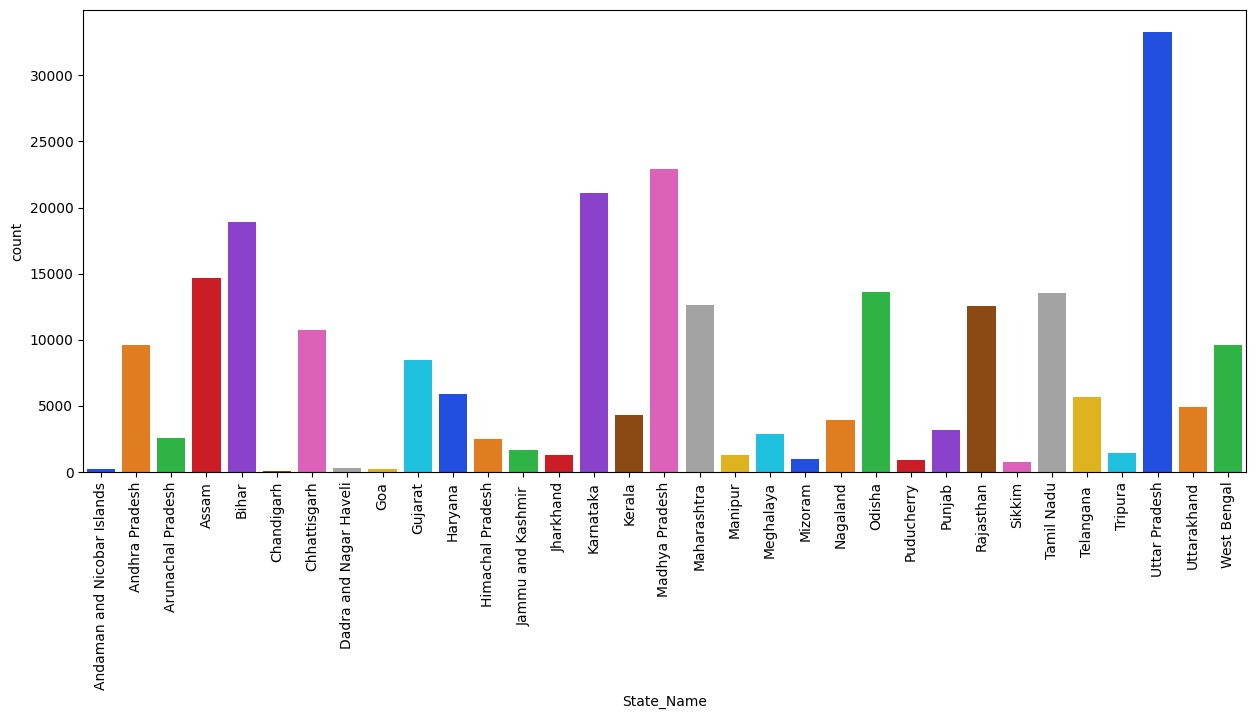
The data used in this study can be accessible via <https://www.kaggle.com/datasets/abhinand05/crop-production-in-india>



**Fig. 1:** Crop production dataset with all features

# EXPERIMENTAL SETUP

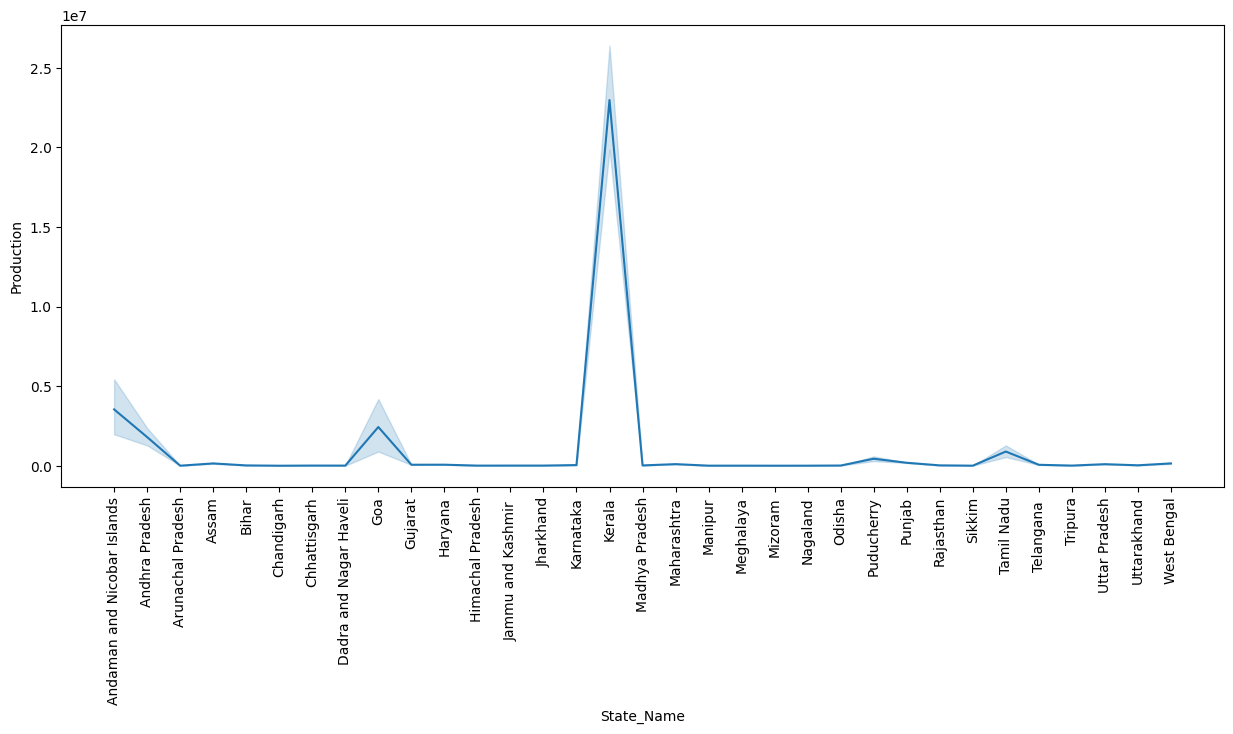
## *Exploratory Data Analysis and Visualisation*



**Fig. 2:** Countplot for State name feature

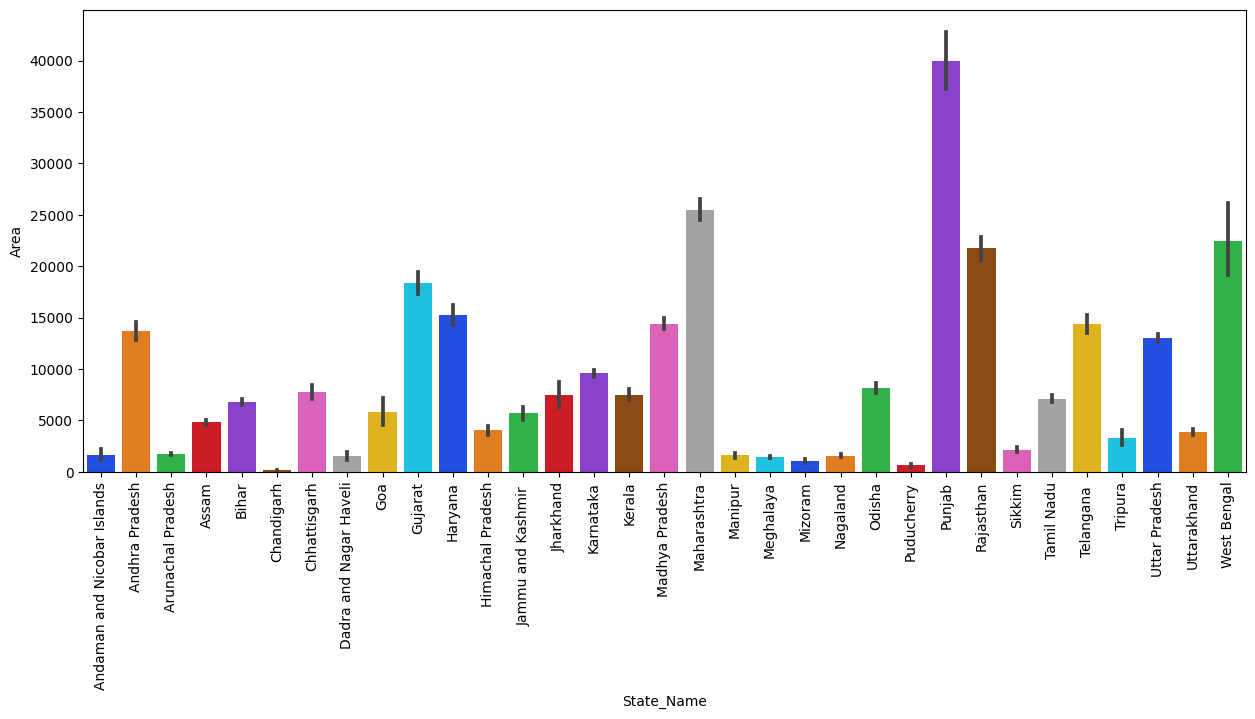
Figure shows the count of each class present in the state name feature. There are a total of 33 classes in this feature.

The count of Uttar Pradesh is 33306 (highest), Madhya Pradesh is 22943, Karnataka is 21122, Bihar is 18885, Assam is 14628, Odisha is 13575, Tamil Nadu is 13547, Maharashtra is 12628, etc. The lowest count state is Chandigarh with a value of 60.



**Fig. 3:** Line plot for State name and Production feature

Figure presents a line plot between state name and production feature. In Kerala state, the production of the crop is maximum around 22,000,000. While in other states there is not that much production, the maximum production for other states is not more than 3,500,000.



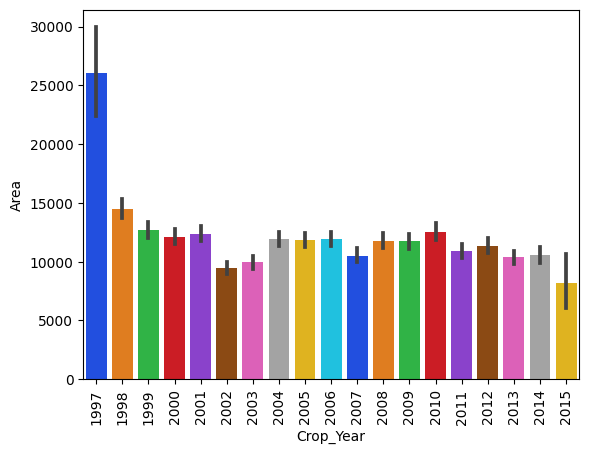
**Fig. 4:** Barplot for State name and Area feature

Figure displays the barplot between state name and area feature. It shows that Punjab state covers the maximum area of around 40000 units. Chandigarh state covers the minimum area which is almost 100 units.



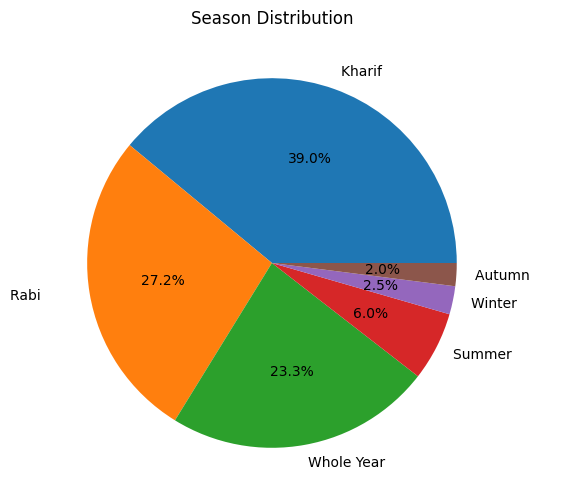
**Fig. 5:** Line plot for crop year and Production feature

Figure shows the line plot between crop year and production feature. Between 2010 and 2012 there is the highest production of crops around 1,000,000, while the minimum production is in the year 2015 around 20,000.



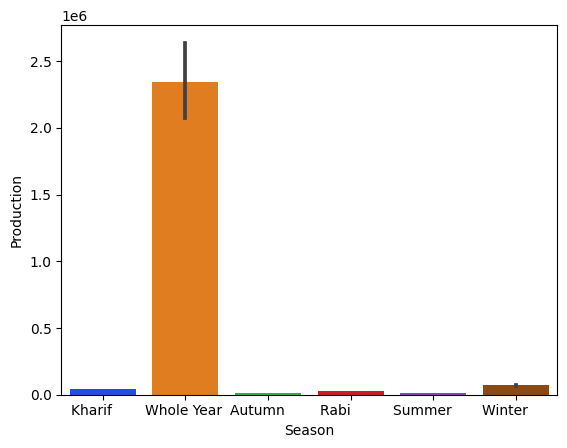
**Fig. 6:** Barplot for Crop Year and Area Feature

The figure shows the barplot between crop year and area feature. In 1997 the maximum area for crop production was around 25000 units while in 2015 it had a minimum area of production of around 10000 units.



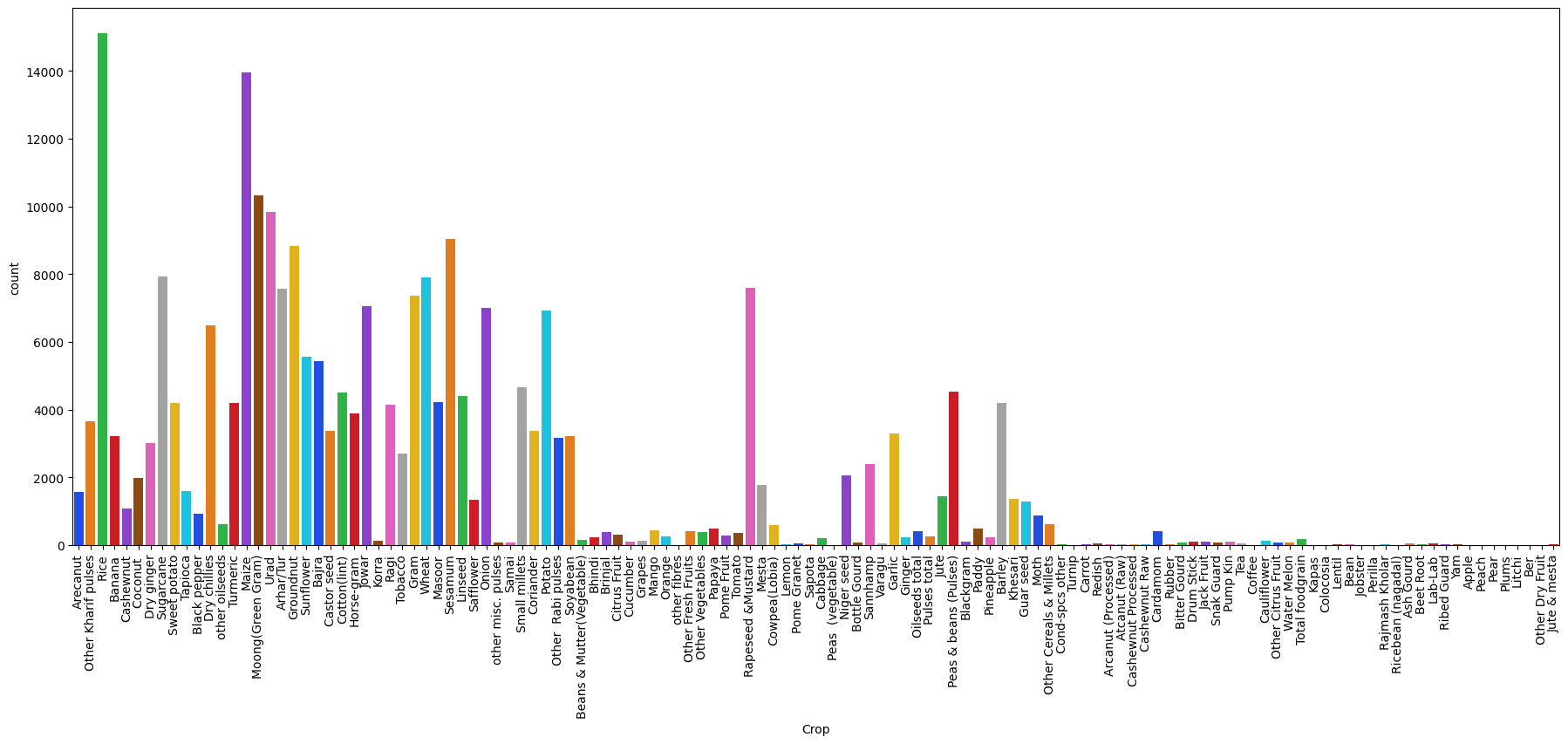
**Fig. 7:** Pie Chart for season feature

Figure displays the pie chart for the season feature showing the distribution of this feature. There are six classes in season and the distribution of particular classes is as follows: Kharif is 39.0%, Rabi is 27.2%, Whole year is 23.3%, Summer is 6.0%, Winter is 2.5%, and Autumn is 2.0%.



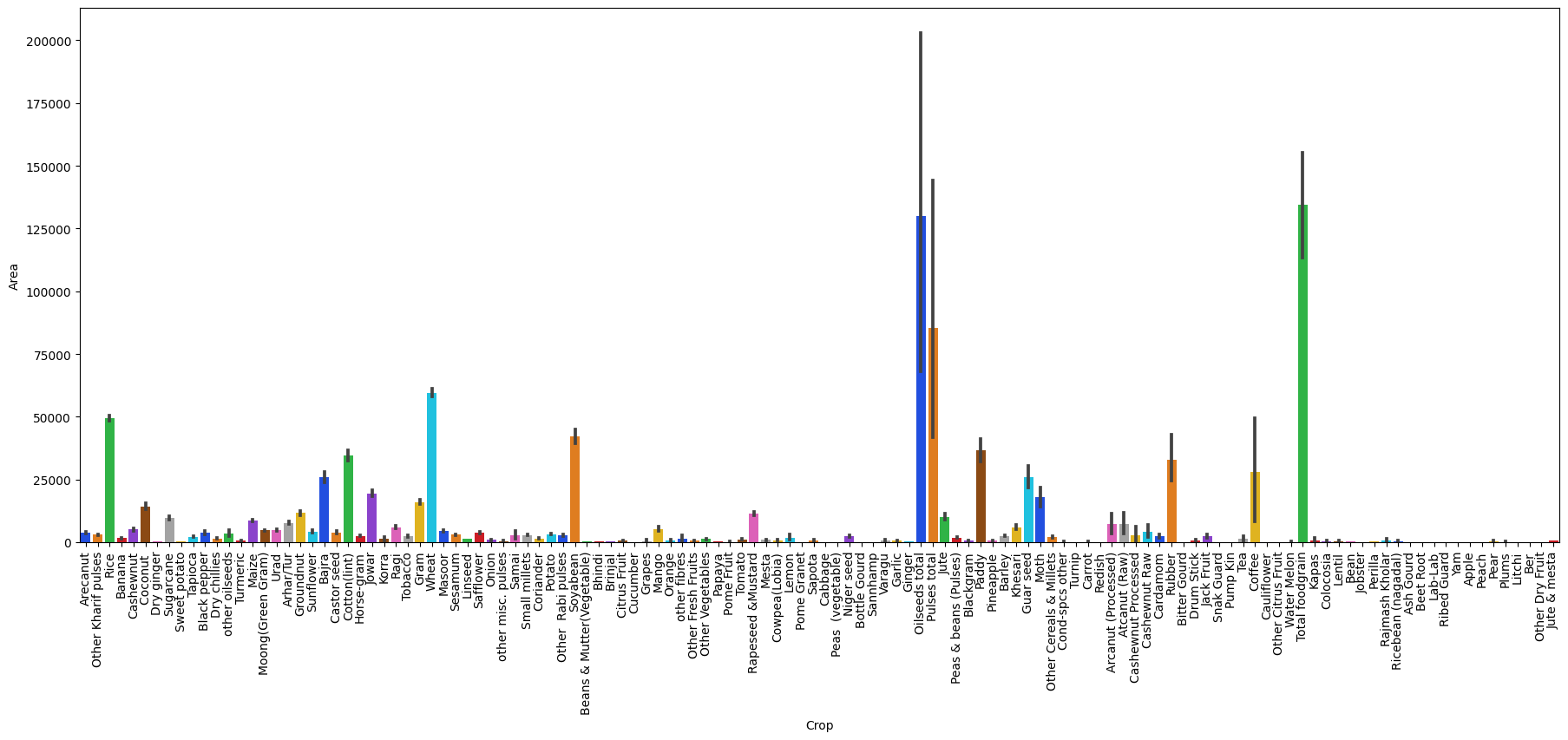
**Fig. 8:** Barplot for season and production Feature

Figure displays the barplot between season and production feature. The whole year there is the highest production of the crop around 22,000,000 while other seasons i.e., kharif, autumn, rabi, summer, and winter have very less production of the crop.



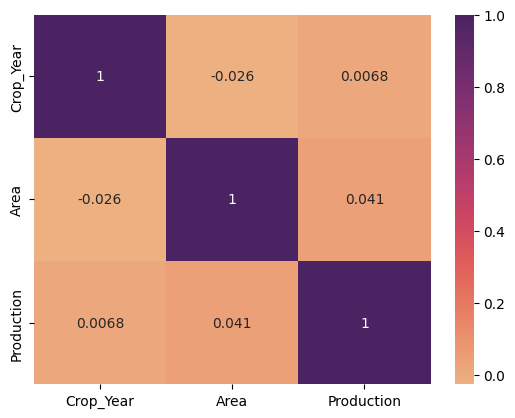
**Fig. 9:** Countplot for crop feature

Figure presents the count plot for crop features. This shows the count of each crop present in this feature. There are a total of 124 crops and the count of few crops are: Rice is 15104, Maize is 13947, Moong(Green Gram) is 10318, Urad is 9850, Sesamum is 9046, Litchi is 6, Coffee is 6, Apple is 4, Peach is 4, Other Dry Fruit is 1, etc.



**Fig. 10:** Barplot for Crop and Area Feature

Figure shows the barplot between crop and area features. Oilseeds total and total foodgrain crops have a maximum area for the production of around 125000 units. While all other crops have less area for the production of crops as compared to these crops.



**Fig. 11:** Heatmap for correlation between features

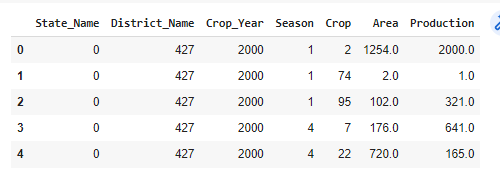
The figure is showing the heatmap that is displaying the correlations between features. A high correlation is found between the following features. Production and crop year, and Production and area.

## *Data Preprocessing*

### *Label Encoding*

By applying the technique of Label Encoding in a machine learning project, categorical features are converted into numerical and making it possible to fit them into machine learning models that take only numeric input.

When the Label Encoding technique is used for categorical features, it converts them to numerical values between zero and the total number of classes minus one. For example, if a categorical feature has six classes, it assigns them with the numerical values 0, 1, 2, 3, 4, and 5 to represent them numerically.

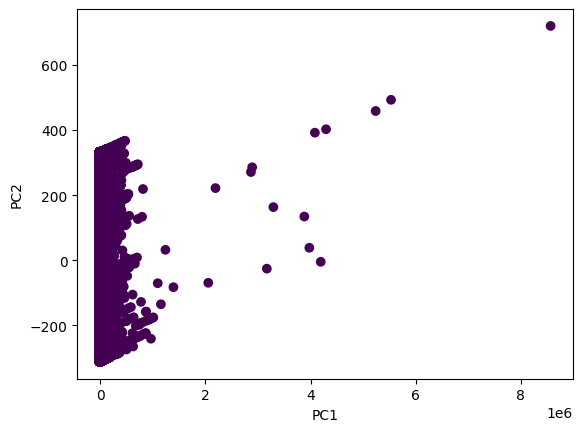


**Fig. 12:** Dataset after label encoding

There are four features in the dataset that are in categorical form. These features are converted into numerical form by applying label encoding. The four features are State name, District name, Season, and Crop.

### *Principal Component Analysis (PCA)*

High-dimensional data can be made simpler using the PCA technique, which also preserves data trends and patterns. It can be accomplished by transforming the data into a smaller number of dimensions that provides compressed summaries of the original features. PCA [5] can handle two main problems that occur in the data: the competitive cost and the increased error rate that results from improving various testing when looking for relations between each feature and a result.



**Fig. 13:** PC1 and PC2 points

The scatter plot shown in Figure presents the distribution of the data in a reduced feature space, with the first principal component (PC1) on the horizontal axis and the second principal component (PC2) on the vertical axis. This plot shows the analysis of insights into the distribution and degree of separation between classes.

The plot also tells how effectively the classes are separated in this reduced feature space.

## *Model Building*

### *Ridge Regression*

GridSearchCV hyperparameter tuning technique is applied to the Ridge Regression model [6]. Three parameters namely: ‘alpha’, ‘fit\_intercept’, and ‘solver’ are specified with different values. Then the ridge regression model, parameters, and cross-validation value is passed as parameters in GridSearchCV. The best parameters are obtained after training data on GridSearchCV.

Then ridge regression is trained by using those parameters' values and the best test score is obtained by them. Ridge Regression best parameters: ‘alpha' is 1.0, 'fit\_intercept’ is True, and 'solver' is 'svd', and test score: 0.00339.

### *Decision Tree Regressor*

GridSearchCV hyperparameter tuning technique is applied to the Decision Tree model. Three parameters namely: max\_depth, ‘min\_samples\_split’, and ‘min\_samples\_leaf’ are specified with different values. The instance of the decision tree model, parameters, and cross-validation value are passed as parameters in GridSearchCV. The best parameters are obtained after training data on GridSearchCV. Then the decision tree [7] is trained by using those parameters' values and the best test score is obtained by them.

Decision Tree best parameters: 'max\_depth' is 5, 'min\_samples\_leaf’ is 3, 'min\_samples\_split' is 2, and test score: 0.69656.

### *Random Forest Regressor*

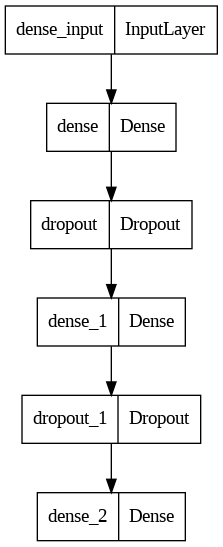
GridSearchCV hyperparameter tuning technique is applied to the Random Forest model. Three parameters namely: ‘max\_depth’, ‘min\_samples\_leaf’, and ‘n\_estimators’ are specified with different values. The instance of the random forest model, parameters, and cross-validation value are passed as parameters in GridSearchCV. The best parameters are obtained after training data on GridSearchCV.

Then random forest [8] is trained by using those parameters' values and the best test score is obtained by them. Random Forest best parameters: 'max\_depth' is 5, 'min\_samples\_leaf' is 2, 'n\_estimators' is 10, and test score: 0.72078.

### *Neural Network (NN)*

The deep learning library Keras was used to construct the ANN [9] architecture in this study. The ANN model contains five layers of neurons. The number of neurons in the input layer corresponds to the number of features in the input dataset. The ReLU activation function is utilised in the hidden layers, and a dropout layer with a value of 0.2 is added after the first hidden layer, which has 64 neurons. The third layer, which has 32 neurons, is followed by a dropout layer with a value of 0.2.

One neuron with a linear activation function, which defines the regression job, is present in the output layer. The architecture of the ANN model is shown in Figure, which has already been explained.



**Fig. 14:** Architecture of ANN model

# RESULT & ANALYSIS

### *Performance Metrics*

Four performance metrics—mean squared error (mse), mean absolute error (mae), root mean square error (rmse), and r2-score—are used to compare the performance of the machine learning models.

The mse the regression model is calculated as,

The mae of the regression model is calculated as,

The rmse of the regression model is calculated as,

The r2 score of the regression model is calculated as,

Here, SSres: The sum of squared residuals, SStot: The total sum of squares.

### *Results Obtained*

On test data, the effectiveness of four trained machine learning models is assessed using the metrics mse, mae, rmse, and r2-score.

The results for each model are shown in Table 1, and it can be seen that the random forest model outperforms all other machine learning methods (ridge regression, decision tree, and artificial neural network).

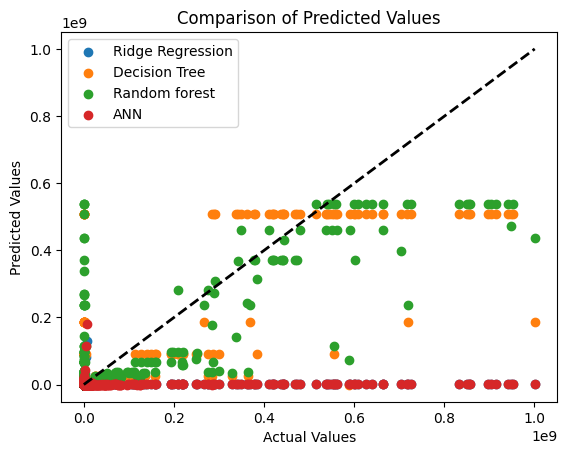
### Artificial neural network is not performing better, it is giving the lowest value of r2-score and high value for MSE, MAE, and RMSE.

**Table 1:** Models evaluation testing score

| **Model** | **MSE** | **MAE** | **RMSE** | **R2-score** |
| --- | --- | --- | --- | --- |
| **Ridge Regression** | 301751664631947.06 | 1292388.16230 | 17371000.68021 | 0.00339 |
| **Decision Tree** | 91873232846345.45 | 419388.46614 | 9585052.57400 | 0.69657 |
| **Random Forest** | 84541927682372.67 | 405148.45029 | 9194668.43787 | 0.72078 |
| **Artificial Neural Network** | 302703152514999.43 | 762530.96925 | 17398366.37489 | 0.00025 |

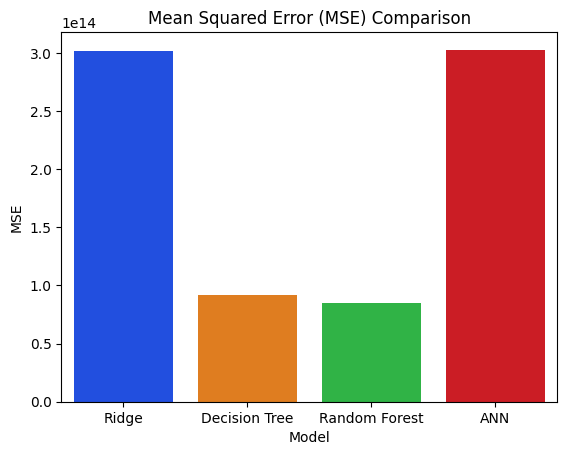
### That’s why Artificial neural networks should not be considered for this type of problem. Similarly, Ridge regression is not performing well having the highest MAE value. Bur random forest and decision tree both are performing well there is a slight variation between metrics values.

### *Comparison of Results*



**Fig. 15:** Scatter plot for comparing predicted values

Figure shows the scatter plot for all four models comparing the predicted values with the actual value. It shows that the ANN model and Ridge regression predicted points are almost overlapped while the Random forest model points are scattered in the plot and close to actual points.

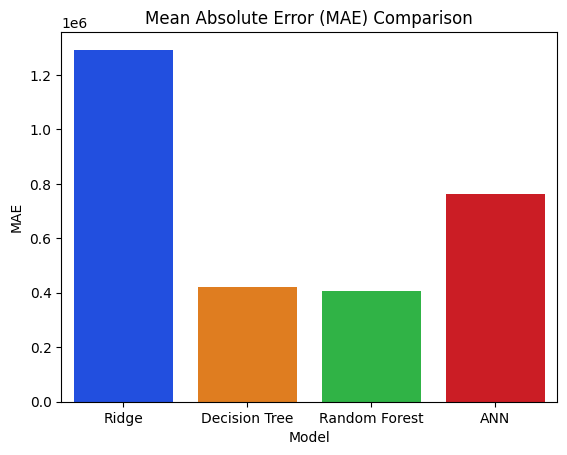


**Fig. 16:** MSE Comparison

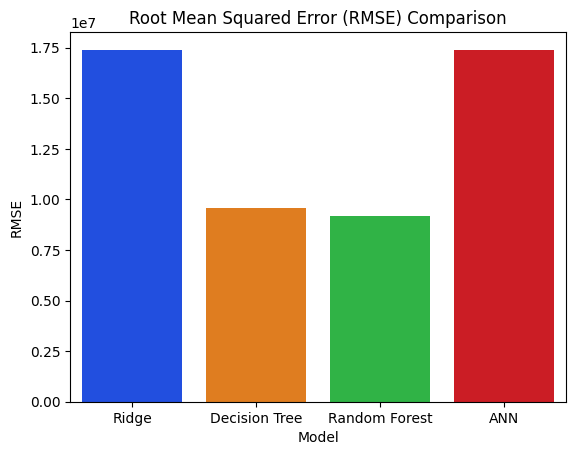
MSE of all four models is compared in the figure showing that the Random forest model has the lowest MSE with a value of 84541927682372.67 while the ANN model has the highest MSE with a value of 302703152514999.43. The decision tree also performs well with an MSE of 91873232846345.45.

MAE of all four models is compared in the figure showing that the Random forest model has the lowest MAE with a value of 9194668.43787 while the Ridge regression model has the highest MAE with a value of 1292388.16230.

The decision tree also performs well with an MAE of 419388.46614 as shown in Fig 17.

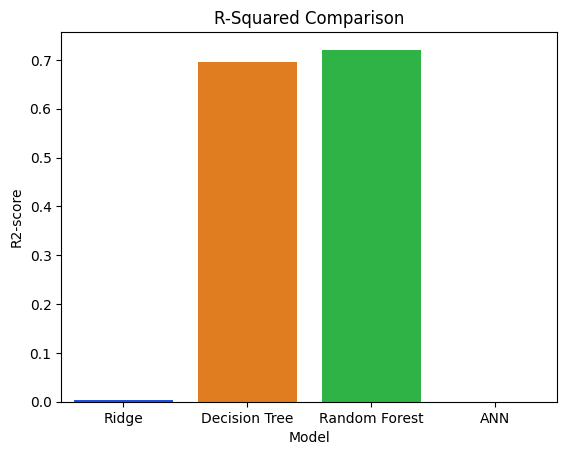


**Fig. 17:** MAE Comparison



**Fig. 18: R**MSE Comparison

RMSE of all four models is compared in the figure showing that the Random forest model has the lowest RMSE with a value of 84541927682372.67 while the ANN model has the highest RMSE with a value of 17398366.37489. The decision tree also performs well with an RMSE of 9585052.57400.



**Fig. 19:** R2-Score Comparison

R2-Score of all four models is compared in the figure showing that the Random forest model has the highest R2-Score with a value of 0.72078 while the ANN model has the highest R2-Score with a value of 0.00025.

The decision tree also performs well with an R2-Score of 0.69657.

# CONCLUSION

The research done in the area of crop production revealed that among the several machine learning models studied, including ANN, decision trees, random forests, and ridge regression. It was found that the random forest model performed the best. The ANN model, on the other hand, performed poorly, as shown by the low r2 score and the high values of MSE, and RMSE. That’s why ANN is not the right choice for this type of problem. The ridge regression model also showed unsatisfactory results and had the highest MAE value and other metrics values are also high. Both the decision tree and random forest models showed promising results, with only small differences in the metric values.

Further research could be done to explore the use of larger data sets and a few more advanced machine-learning algorithms to identify important factors in crop production and develop appropriately adapted treatment plans.

# REFERENCES

[1] Xu, X., Gao, P., Zhu, X., Guo, W., Ding, J., Li, C., Zhu, M. and Wu, X., 2019. Design of an integrated climatic assessment indicator (ICAI) for wheat production: A case study in Jiangsu Province, China. Ecological Indicators, 101, pp.943-953.

[2] Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D., Pozza, L.E., Ugbaje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M. and Bishop, T.F., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precision Agriculture, 20, pp.1015-1029.

[3] Renuka, S.T., 2019. Evaluation of machine learning algorithms for crop yield prediction. Int. J. Eng. Adv. Technol.(IJEAT), 8(6).

[4] MamathaJajur, S., Soumya, N. G., & Raju, G. T. (2019). Crop Recommendation using Machine Learning, Techniques. International Journal of Innovative Technology and Exploring Engineering, 9(25), 658–661.

[5] Pham, H.T., Awange, J., Kuhn, M., Nguyen, B.V. and Bui, L.K., 2022. Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices. Sensors, 22(3), p.719.

[6] Ahmed, A.M., Sharma, E., Jui, S.J.J., Deo, R.C., Nguyen-Huy, T. and Ali, M., 2022. Kernel ridge regression hybrid method for wheat yield prediction with satellite-derived predictors. Remote Sensing, 14(5), p.1136.

[7] Rajeswari, S. and Suthendran, K., 2019. C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. Computers and Electronics in Agriculture, 156, pp.530-539.

[8] Elavarasan, D. and Vincent, P.D.R., 2021. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. Journal of Ambient Intelligence and Humanized Computing, pp.1-14.

[9] Elahi, E., Weijun, C., Zhang, H. and Abid, M., 2019. Use of artificial neural networks to rescue agrochemical-based health hazards: A resource optimisation method for cleaner crop production. Journal of Cleaner Production, 238, p.117900.