2. CROP YIELD PREDICTION MODEL

The crop yield prediction model was developed using a dataset from Kaggle, consisting of 19698 rows and 10 columns. The dataset contains agricultural data for various crops cultivated in multiple states in India during the years 1997-2020. It is focused on predicting crop yields based on several agronomic factors, such as weather conditions, fertilizer and pesticide usage, and other relevant variables.

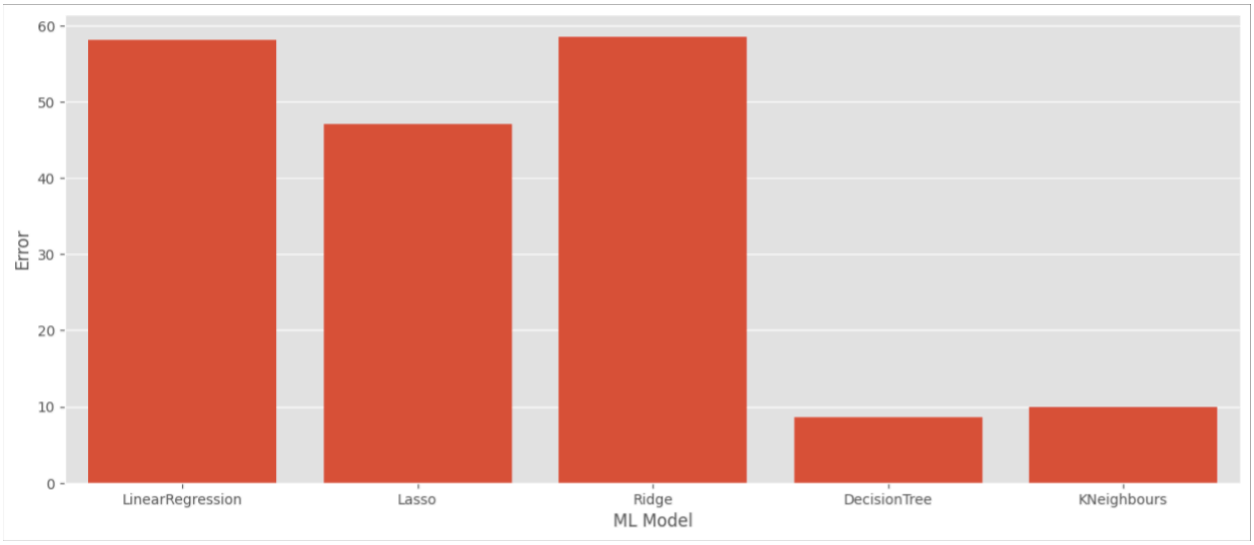
Crop	Crop_Year	Season	State	Area	Production	Annual_Ra...	Fertilizer	Pesticide	Yield
Arecanut	1997	Whole Year	Assam	73814	56708	2051.4	7024878.38	22882.34	0.796086957
Arhar/Tur	1997	Kharif	Assam	6637	4685	2051.4	631643.29	2057.47	0.710434783
Castor seed	1997	Kharif	Assam	796	22	2051.4	75755.32	246.76	0.238333333
Coconut	1997	Whole Year	Assam	19656	126905000	2051.4	1870661.52	6093.36	5238.051739
Cotton(lint)	1997	Kharif	Assam	1739	794	2051.4	165500.63	539.09	0.420909091
Dry chillies	1997	Whole Year	Assam	13587	9073	2051.4	1293074.79	4211.97	0.643636364
Gram	1997	Rabi	Assam	2979	1507	2051.4	283511.43	923.49	0.465454545
Jute	1997	Kharif	Assam	94520	904095	2051.4	8995468.4	29301.2	9.919565217
Linseed	1997	Rabi	Assam	10098	5158	2051.4	961026.66	3130.38	0.461363636
Maize	1997	Kharif	Assam	19216	14721	2051.4	1828786.72	5956.96	0.615652174

Based on a comprehensive study of the dataset, the constraints showed a linear pattern and hence five different linear machine learning models, Linear Regression, Lasso, Ridge, Decision Tree and K-Nearest Neighbours.
The above mentioned machine learning models were selected for the following reasons:

- **Linear Regression** is computationally efficient, requires minimal hyperparameter tuning, and serves as a strong baseline for more complex models. It is particularly suited for continuous variables like yield and small datasets.
- **Lasso** handles multicollinearity effectively and is suitable for datasets with many features, even when some are irrelevant or highly correlated.
- **Ridge** is particularly suitable for datasets with numerous correlated features but retains all features instead of performing feature selection. It performs well for linear relationships.
- **Decision Trees** are robust to outliers and missing data, making them versatile for agricultural datasets.
- **K-NN** is highly interpretable and works well for small datasets but can be computationally expensive for large datasets. Additionally, its performance depends heavily on the choice of k, distance metric, and feature scaling.

To evaluate model performance, the dataset was split into training and testing sets, with 80% used for training and 20% used for testing. The above mentioned machine learning models were applied, and their effectiveness was evaluated based on mean absolute error and R2 score to ensure reliability.

1. **Linear regression**:
   The linear regression model exhibited a mean absolute error (MAE) of **58.19** and an R^2 score of **0.8468**. Despite its relatively high error, the model demonstrated robustness and efficiently produced results within a few seconds, making it a reliable but less accurate choice compared to other models.
2. **Lasso**:
   Lasso regression achieved a lower MAE of **47.03** compared to linear regression but yielded a slightly reduced R^2 score of **0.8431**. While it demonstrated improved accuracy in terms of error minimization, the trade-off in R^2 score makes it a moderate contender for this dataset.
3. **Ridge**:
   Ridge regression recorded an MAE of **58.58** and an R^2 score of **0.8450**. This model showed no significant improvement in R^2 score while also incurring a slightly higher error than linear regression, positioning it as the least preferred model for this dataset.
4. **Decision Tree**:
   The decision tree model outperformed all others, with an MAE of **8.63** and an R^2 score of **0.9660**. Its drastic reduction in error and near-perfect R^2 score make it the most effective and preferred model for the given dataset.
5. **K-NN**:
   The k-NN model demonstrated strong performance with an MAE of **9.91** and an R^2 score of **0.9347**. Although it significantly reduced the error and achieved a high R^2 score, its slightly higher error value compared to the decision tree model places it as the second-best choice for this dataset.



The most efficient model (Decision Tree in this case) was utilized out of the five models for the chosen dataset and was used to predict the crop yield.
A function was defined where the user can provide inputs (weather conditions, fertilizer and pesticide usage, and other relevant variables in this case) and can find the predicted yield (per unit area).The results of the predicted yields were compared with the actual yield values from the dataset that showed almost negligible error.