A coherent algorithm for crop prediction using machine learning techniques with historical data.

Abstract:

Agriculture is the backbone of the Indian economy, with more than half of the country's people relying on it for subsistence. Crop production is predicted using machine learning techniques based on parameters such as temperature, humidity, ph., rainfall, and crop name. The most popular and powerful machine learning algorithm, Random Forest, can do both classification and regression tasks. They are used in crop selection to reduce crop yield output losses, regardless of the distracting environment. Weather, climate, and other related environmental elements have posed a significant danger to agriculture's long-term viability. Machine learning (ML) is significant since it offers a decision-support tool for Crop Yield Prediction (CYP), which may help with decisions like which crops to cultivate and what to do during the crop's growing season.

Introduction:

Agriculture is the backbone of India’s economy since its plays a vital role in the survival of every human and animal in India. The worldwide population was estimated at 1.8 billion in 2009 and is predicted to increase to 4.9 billion by 2030, leading to an extreme increase in demand for agricultural products. To produce in mass quantity, people are using technology in the wrong way. New kinds of hybrid varieties are produced day by day. However, these varieties do not provide the essential contents of the naturally produced crop. These unnatural techniques spoil the soil. It all leads to further environmental harm. Most of these Unnatural techniques are used to avoid losses.

The core objective of crop yield prediction is to achieve higher agricultural crop production and many established models are exploited to increase the yield of crop production. When the producers of these crops know the accurate information on the crop yield it minimizes the loss. To achieve this We can use the past information on weather, temperature and several other factors the information is given.

Literature Survey:

Devika and Ananthi et. al. utilized data mining techniques to predict the annual yield of major crops. Farmers were opposed to harvesting the yield because of insufficient availability of water sources and unpredictable weather variations but these issues were overcome by developing a data mining method. The developed model was gathering crop growing documents that used to be stored and analyzed for valuable crop yield prediction. In some of the data mining actions, the training data can be collected from the previous documents and the gathered documents were used in the phase of training which has to exploit. An advantage of the developed model was that the highest level of crop yield prediction was obtained only in sugarcane, cotton, and turmeric. However, the range was low for other crops such as wheat, rice, etc. [1]

D. Jayanarayana Reddy et. al. collected agricultural environment data and applied various machine learning algorithms such as Support Vector Machine (SVM), CNN, ANN, and DNN. Multiple Linear Regression and compared them based on various metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared Error.[2]

S. P. Raja et. al. collected a dataset called felin dataset and applied pre-processing tasks such as Missing value handling and balancing the imbalanced data followed by feature selection techniques like MRFE, RFE, and Boruta on which they have applied different classification algorithms such as Random Forest, Bagging, KNN, SVM, Decision Tree, Naive Bayes upon which various evaluation metrics are applied and compared for the best. [3]

Tiwari et. al. developed a model for crop yield Prediction by using CNN and Geographical Index. The existing model faced a problem during a continuous breakdown in agricultural drifts for crop cultivation which were not suitable with environmental factors like temperature, weather and soil condition. The developed CNN model which used spatial features as input were trained by BPNN for error prediction. An advantage of the developed model was that it was implemented on a real-time dataset that was taken from authentic geospatial

resources. However, the developed model reduced the relative error but decreased the efficiency of crop yield prediction.

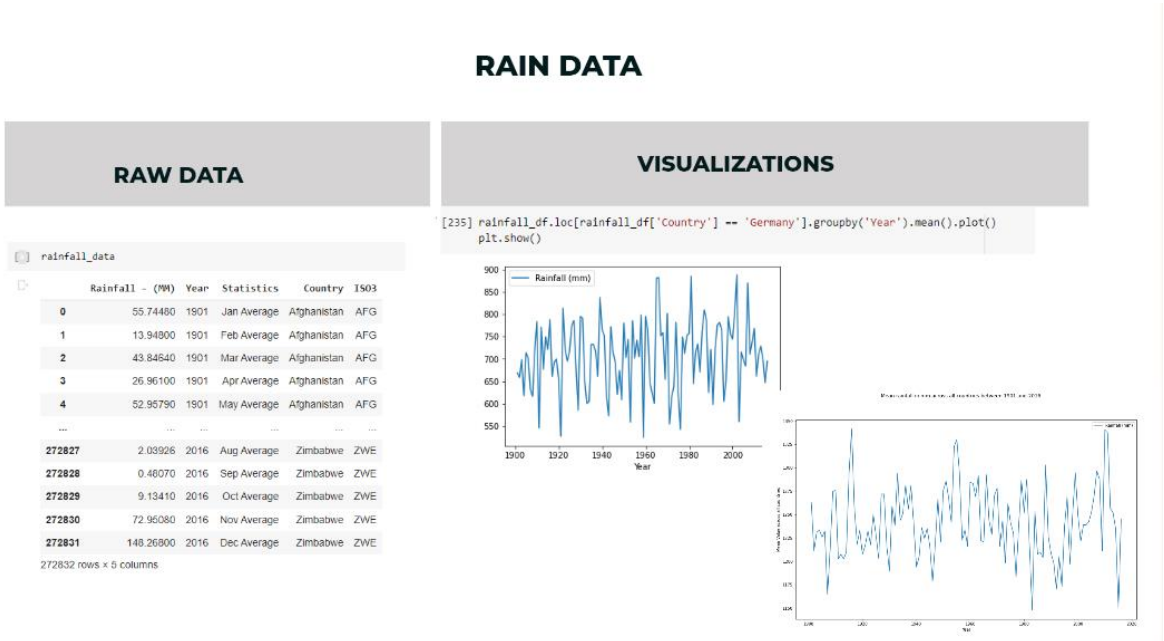
[4].Dr Y. Jeevan Nagendra Kumar et. al. proposed an approach to predict crop yield using a supervised machine learning algorithm called random forest on a dataset containing several factors such as temperature, humidity, ph., rainfall, and crop name. [5]

benefits to the society:

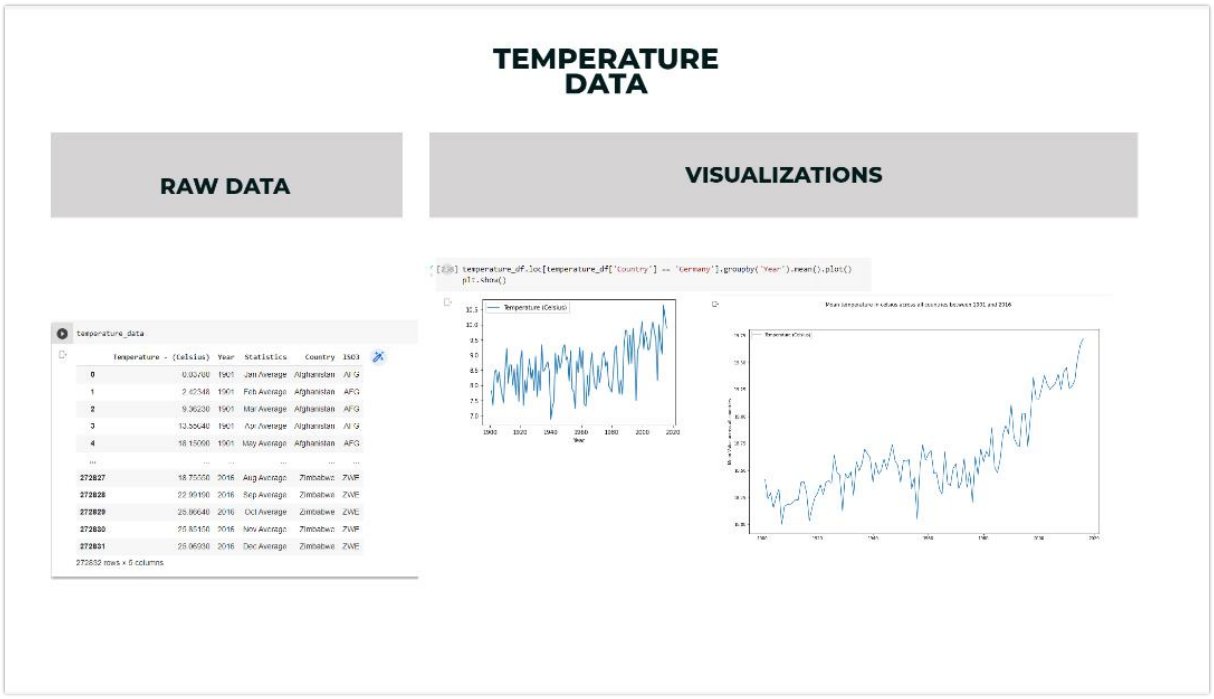
- Increased Production
- Water Conservation
- Real-Time Data and Production Insight
- Lowered Operation Costs
- Increased Quality of Production
- Accurate Farm and Field Evaluation

Data Used:

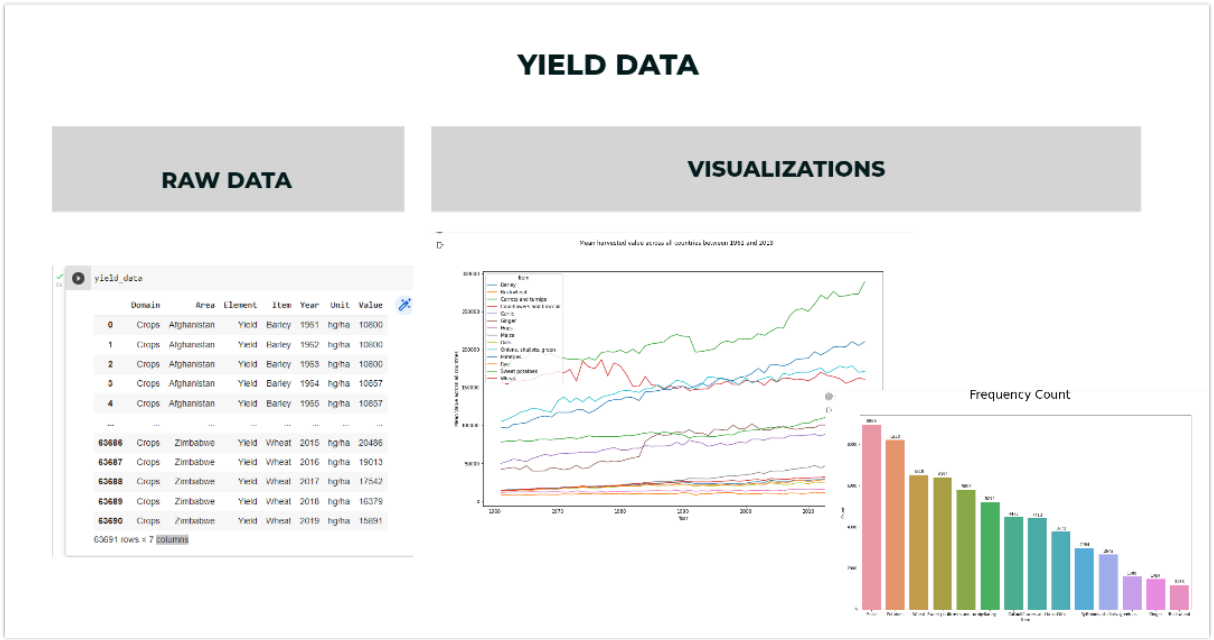
Rain Fall Data:



Temperature Data:



Yield Data:



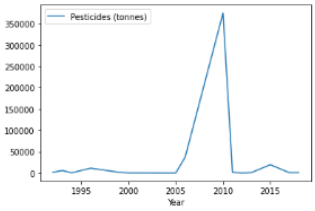
Pesticides Data:

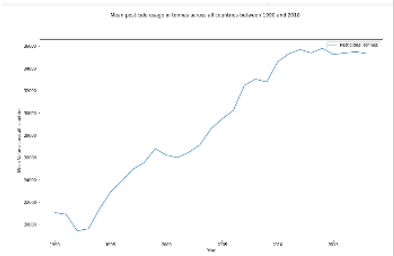
PESTICIDES DATA

RAW DATA

VISUALIZATIONS

```
[244] pesticides_df.loc[temperature_df['Country'] == 'Germany'].groupby('Year').mean().plot()
      plt.show()
```





Merged Data:

```
[226] rain_temp_df = pd.merge(rainfall_df, temperature_df, on=['Country', 'Year', 'ISO3'])
      rain_temp_yield_df = pd.merge(rain_temp_df, yield_df, on=['Country', 'Year'])
      rain_temp_yield_pest_df = pd.merge(rain_temp_yield_df, pesticides_df, on=['Country', 'Year'])
      rain_temp_yield_pest_df.drop('ISO3', axis=1, inplace=True)
      data = rain_temp_yield_pest_df[['Year', 'Country', 'Item', 'Rainfall (mm)', 'Temperature (Celsius)', 'Pesticides (tonnes)', 'Yield (hg/ha)']]
      data
```

	Year	Country	Item	Rainfall (mm)	Temperature (Celsius)	Pesticides (tonnes)	Yield (hg/ha)
0	1990	Albania	Barley	812.23450	12.051221	121	10000
1	1990	Albania	Carrots and turnips	812.23450	12.051221	121	150000
2	1990	Albania	Cauliflowers and broccoli	812.23450	12.051221	121	171429
3	1990	Albania	Garlic	812.23450	12.051221	121	65000
4	1990	Albania	Maize	812.23450	12.051221	121	36613
...	...	...	...	...	...	...	...
25224	2016	Zimbabwe	Maize	455.13295	22.310167	2185	4405
25225	2016	Zimbabwe	Oats	455.13295	22.310167	2185	20505
25226	2016	Zimbabwe	Potatoes	455.13295	22.310167	2185	51792
25227	2016	Zimbabwe	Sweet potatoes	455.13295	22.310167	2185	27283
25228	2016	Zimbabwe	Wheat	455.13295	22.310167	2185	19013

25229 rows × 7 columns

Theoretical Analysis:

Applying one hot encoding on data:

- One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Say suppose the dataset is as follows: The categorical value represents the numerical value of the entry in the dataset

```
[ ] data
```

	Year	Rainfall (mm)	Temperature (Celsius)	Pesticides (tonnes)	fertilizers	Country_Albania	Country_Algeria	Country_Angola	Country_Antigua and Barbuda	Country_Argentina	...	Item_Ginger	Item_Hops	Item_Maize	Item_Oats	Item_Onions, shallots, green	Item_Potatoes	Item_Rye	Item_Sweet potatoes	Item_Wheat	Yield (hg/ha)
0	2009	1270.37230	12.435142	411	35697.4698	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	26471
1	2009	1270.37230	12.435142	411	35697.4698	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	333737
2	2009	1270.37230	12.435142	411	35697.4698	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	233033
3	2009	1270.37230	12.435142	411	35697.4698	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	80319
4	2009	1270.37230	12.435142	411	35697.4698	1	0	0	0	0	...	0	1	0	0	0	0	0	0	0	6818
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
615	2009	352.48230	18.643265	13697	346243.9590	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	21398
616	2009	897.40706	24.319350	915	7314.1868	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	10000
617	2009	897.40706	24.319350	915	7314.1868	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	145000
618	2009	897.40706	24.319350	915	7314.1868	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	104287
619	2009	897.40706	24.319350	915	7314.1868	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	12500

Regression Analysis:

Regression analysis is used to analyze and determine the relationship between response variable and explanatory variable. The variables considered for analysis in this research work are pesticides , rainfall , temperature , yield , fertilizers .Crop yield is a dependent variable which depends on all these ecological factors.

```
[ ] y = data['Yield (hg/ha)']
    X = data.drop('Yield (hg/ha)', axis=1)
```

```
[ ] from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[ ] def mean_absolute_percentage_error(y_true, y_pred):
        y_true, y_pred = np.array(y_true), np.array(y_pred)
        return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
```

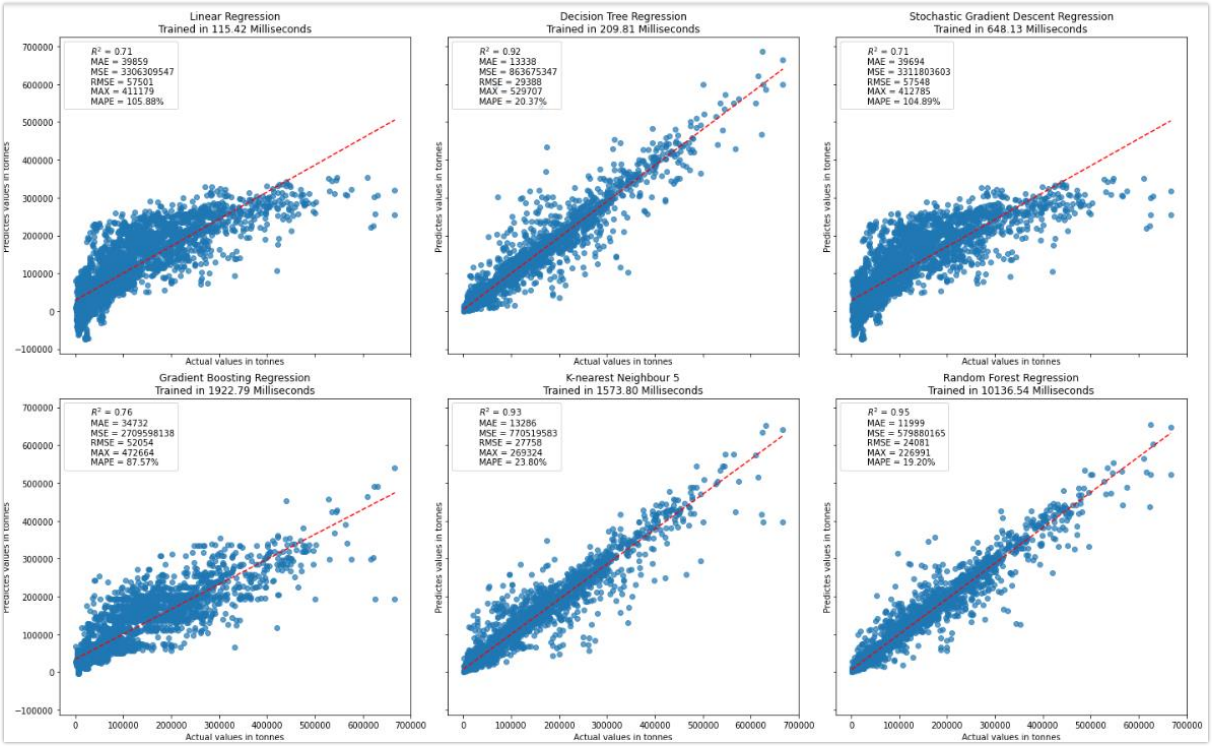
```
[ ] def plot_regression_results(ax, y_test, y_pred, title, estimated_time, scores):

      # linear least-squares
      slope, intercept, rvalue, pvalue, stderr = linregress(y_test, y_pred)
      ax.plot([y_test.min(), y_test.max()], [intercept+y_test.min()*slope, intercept+y_test.max()*slope], '--r')

      ax.scatter(y_test, y_pred, alpha=0.7)

      # Anzeigen der Werte in einer Box
      extra = plt.Rectangle((0, 0), 0, 0, fc="w", fill=False,
                            edgecolor='none', linewidth=0)
      ax.legend([extra], [scores], loc='upper left')

      ax.set_xlabel('Actual values in tonnes')
      ax.set_ylabel('Predictes values in tonnes')
      ax.set_title('{}\nTrained in {:.2f} Milliseconds'.format(name, estimated_time*1000))
```

Experimental Investigations:

Types of Regression Analysis Techniques:

- Linear Regression
- Decision Tree Regression
- Stochastic Gradient Descent Regression
- Gradient Boosting Regression
- K-nearest Neighbour
- Random Forest Regression

Regression Analysis Techniques:



CONCLUSION:

Crop yield expectation has been a difficult issue for ranchers since numerous years. This work for the most part centers around dissecting the creation of harvest yield in India from 1999 to 2014, and to anticipate the yield for the following 5 years utilizing the AI techniques.Every paper examines yield expectation with AI however varies from the elements. The examinations likewise contrast in scale, land position, and harvest. The selection of

elements is subject to the accessibility of the dataset and the point of the exploration. Concentrates on additionally expressed that models with additional elements didn't necessarily in every case give the best exhibition to the yield expectation. To find the best performing model, models with more and less elements ought to be tried. Numerous calculations have been utilized in various examinations. The outcomes demonstrate the way that no particular end can be attracted with respect to what the best model is, yet they plainly show that some AI models are utilized more than the others. The most utilized models are the arbitrary timberland, brain organizations, direct relapse, and inclination helping tree. The vast majority of the examinations utilized an assortment of AI models to test which model had the best forecast.