CROP YIELD PREDICTION USING MACHINE LEARNING

PAVAN KUMAR CHEERLAVANCHA

ABSTRACT:

Agriculture is the backbone of the Indian economy, with more than half of the country's people depend on agriculture. Prediction of crop yield includes forecasting factors like temperature, humidity, rainfall, etc., and crop yield based on soil moisture includes few measures like NPK (Nitrogen, Phosphorous and potassium) and pH values. Machine learning (ML) is a useful decision-making model for estimating crop yields, and also for deciding what crops to plant and what to do during the crop's growing season. Here we use different type of machine learning (ML) models such as like Decision tree, Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest etc.

The main purpose of the report is to provide an interface to the farmers which help farmers to choose crop that is best suitable for cultivation based on parameters like nitrogen, phosphorous, humidity and temperature rather than choosing crops based on trends in the market.

1)PROBLEM STATEMENT:

The lack of knowledge among farmers regarding the selection of appropriate crops for cultivation due to insufficient understanding into soil characteristics, temperature, humidity, seasonal change etc. farmers face difficulties in making good decisions regarding crop selection techniques. The other reason is farmers being stuck on to their instincts and usage of same crop pattern rotation which in turn reduces crop fertility leading to an overall decrease in crop production.

Predicting the crop yield well ahead of its plantation would help farmers and market contractors strategize befitting actions to market and store their produce. These kinds of predictions will also help farmers minimize losses due to crop failure.

2) CUSTOMER NEED ASSESSMENT:

a) **Soil Analysis and Preparation:** Farmers require soil testing to know its fertility, pH levels, nutrient content, and soil structure. This analysis informs decisions regarding soil modification, such as fertilizers or lime, to optimize soil health and crop growth.

b) **Crop Selection Guidance:** Farmers need assistance in selecting appropriate crops based on factors like soil type, climate conditions, market demand, and rotation considerations. Guidance on crop which is suitable for the environment is essential for maximizing yields and profitability.

c) **Weather and Climate Information:** Access to accurate weather forecasts and climate data is critical for farmers to plan planting schedules, irrigation strategies, and pest management practice.

d) **Planning Techniques and Timing:** Farmers require guidance on optimal planting techniques, including seed depth, spacing, and planting density, tailored to each crop type. Timely planting decisions based on factors like soil temperature, moisture levels, and expected weather patterns are crucial for achieving maximizing yields.

e) **Weather Management Strategies:** Required water availability is essential for crop growth, making water management a priority for farmers, especially in regions prone to drought or water scarcity. They need support in implementing efficient irrigation systems, scheduling irrigation events, and conserving water resources through techniques like mulching or rainwater harvesting.

f) **Pest and Disease Management:** Farmers face threats from pests, diseases, and weeds that can significantly impact crop yields and quality. They require information on integrated pest management (IPM) practices, including pest monitoring, biological control methods, and judicious use of pesticides to minimize risks while safeguarding environmental and human health.

g) **Financial Planning and Budgeting:** Before planting, farmers need assistance in developing comprehensive financial plans and budgets covering input costs, labour expenses, equipment investments, and potential revenues. Access to financing options, government subsidies, and insurance schemes can help mitigate financial risks associated with crop production.

2.1) MARKET AND BUSINESS NEED ASSESSMENT:

a) **Market Analysis:** Identifying the demand for the chosen crop in our market. Analysing market trends, including pricing fluctuations and consumer preferences.

b) **Cost Analysis:** Determine the cost of cultivation, including seeds, fertilizers, pesticides, labour, machinery. Estimate operational expenses such as irrigation, utilities, and transportation.

c) **Revenue Forecasting:** Estimate the expected yield per acre/hectare based on soil testing results, climate conditions, and farming practices. Calculate the potential revenue based on market prices and projected yield. Consider seasonal variations in pricing and demand.

d) **Risk Assessment:** Identifying potential risks such as pests, diseases, adverse weather conditions, and market fluctuations. Develop strategies to mitigate these risks, such as crop insurance, pest control measures.

e) **Sustainability Assessment:** Evaluate the environmental impact of crop cultivation, including soil health, water usage, and pesticide usage. Explore sustainable farming practices to minimize negative environmental effect.

3) TARGET SPECIFICATION:

a) **Accuracy:** The primary target is to develop models with high accuracy in predicting crop yields. Accuracy can be measured using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) etc.

b) **Data Quality:** Check whether high-quality data is used for model training. This includes accurate and reliable historical data on factors influencing crop yield, such as weather data, soil quality, crop type, agricultural practices, and any other relevant variables.

c) **Feature Selection:** Feature selection techniques such as correlation analysis, feature importance ranking, or domain knowledge expertise can be employed to select the most important features.

d) **Model Interpretability:** Aim for models that provide insights into which factors are driving the predictions. This is particularly important in agricultural applications, where interpretability can help farmers understand the underlying reasons for yield predictions and make informed decisions.

e) **Scalability:** Ensure that the developed models are scalable and can handle large datasets efficiently. Scalability is crucial for practical deployment in real-world agricultural systems where large-scale data processing may be required.

f) **Real-time Prediction:** Depending on the application, there might be a requirement for real-time or near-real-time prediction of crop yields. In such cases, model inference speed and efficiency become important targets.

g) **Integration:** Ensure that the ML models can be integrated into existing agricultural systems or decision-support tools used by farmers or agricultural experts.

h) **Cost-effectiveness:** Develop cost-effective solutions that provide significant value relative to the resources invested. This includes considerations such as computational resources required for model training and deployment, as well as the overall economic feasibility of implementing the solution in agricultural practices.

4) EXTERNAL SEARCH

- www.google.com
- www.https//ieeexplore.ieee.org
- www.researchgate.net/publication/371553138_Applications_of_Machine_Learning_in_Agriculture
- www.machine_learning_models.com
- www.sciencedirect.com/science/article/pii/S0168169920302301

5) BENCHMARKING:

SERVICES	TRADITIONAL METHODS	USING ML MODELS
Prediction Accuracy	Traditional methods of crop yield prediction often relied on simplistic statistical models or historical averages, which may not capture the complexities of the underlying data.	ML models can leverage advanced algorithms to analyse complex relationships between various factors affecting crop yield, resulting in more accurate predictions.
Feature Selection	Traditional methods may have limited capabilities in selecting relevant features or variables that significantly influence crop yield.	ML algorithms can automatically identify and incorporate a wide range of features, including weather data, soil characteristics, agricultural practices, and crop phenology, leading to more comprehensive and informative models.
Model Flexibility	Traditional models may lack flexibility and scalability, making it challenging to adapt to different crops, regions, or changing environmental conditions.	ML techniques offer greater flexibility, allowing for the development of adaptable models that can be tailored to specific crops, regions, or even individual farms.
Scalability	Traditional methods may face challenges in handling large-scale datasets or real-time data streams, limiting their applicability in modern agricultural systems.	ML algorithms are highly scalable and can efficiently process large volumes of data, making them well-suited for real-time crop yield prediction and decision support in precision agriculture applications.
Decision Support	Traditional methods may provide limited decision support capabilities, primarily focusing on forecasting crop yields without offering actionable insights for farmers.	ML models can not only predict crop yields but also provide actionable recommendations and decision support tools based on the predicted outcomes.

6) APPLICABLE PATENTS:

- Checking for patents related to machine learning algorithms for crop yield prediction.
- Ensuring compliance with existing patents or seeking licensing agreements if necessary.

7) APPLICABLE REGULATIONS:

- Compliance with agricultural and environmental regulations regarding data usage and farming practices.
- Stick to data privacy laws when handling sensitive information.

8) APPLICABLE CONSTRAINTS:

- Limited budget for research and development.
- Need for expertise in machine learning, agriculture, and data analysis.

9) BUSINESS OPPOURTUNITY:

- License the crop yield prediction technology to agricultural companies, government agencies, or research institutions.
- Negotiate licensing fees based on the scale of usage, geographic coverage, and value-added features.
- Partner with agricultural equipment manufacturers, seed companies, or agribusinesses to bundle crop yield prediction services with their products.
- Share revenues through partnership agreements or revenue-sharing models.

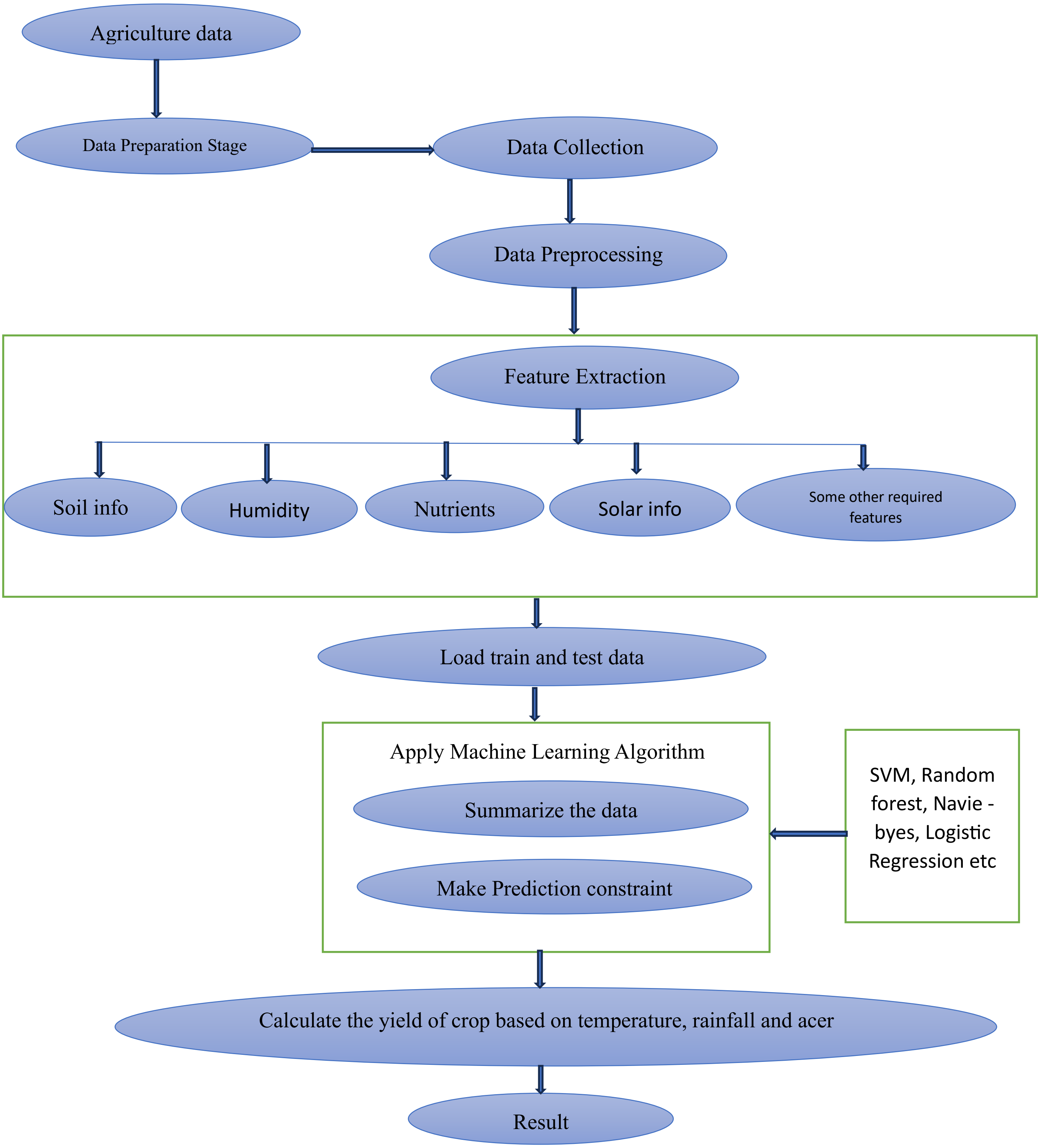
10) CONCEPT GENERATION:

Concept is simple that we want to enhance crop productivity and its quality by using of machine learning.

10.1) BLOCK DIAGRAM:

The steps that are involved in crop yield prediction using machine learning methodology are stated as follows. Firstly, the agriculture Data is utilized for the crop yield prediction, Next, the data is undergone for pre-processing to remove the noisy data. The pre-processed data is undergone for feature extraction process that includes features such as soil information, nutrients, field management etc. which are used to perform the classification using ML algorithms. The results obtained by the existing models using ML algorithms are effectively described in the following section.

Figure 1: The flow diagram of the crop yield prediction using ML algorithms.



CODE IMPLEMENTATION:

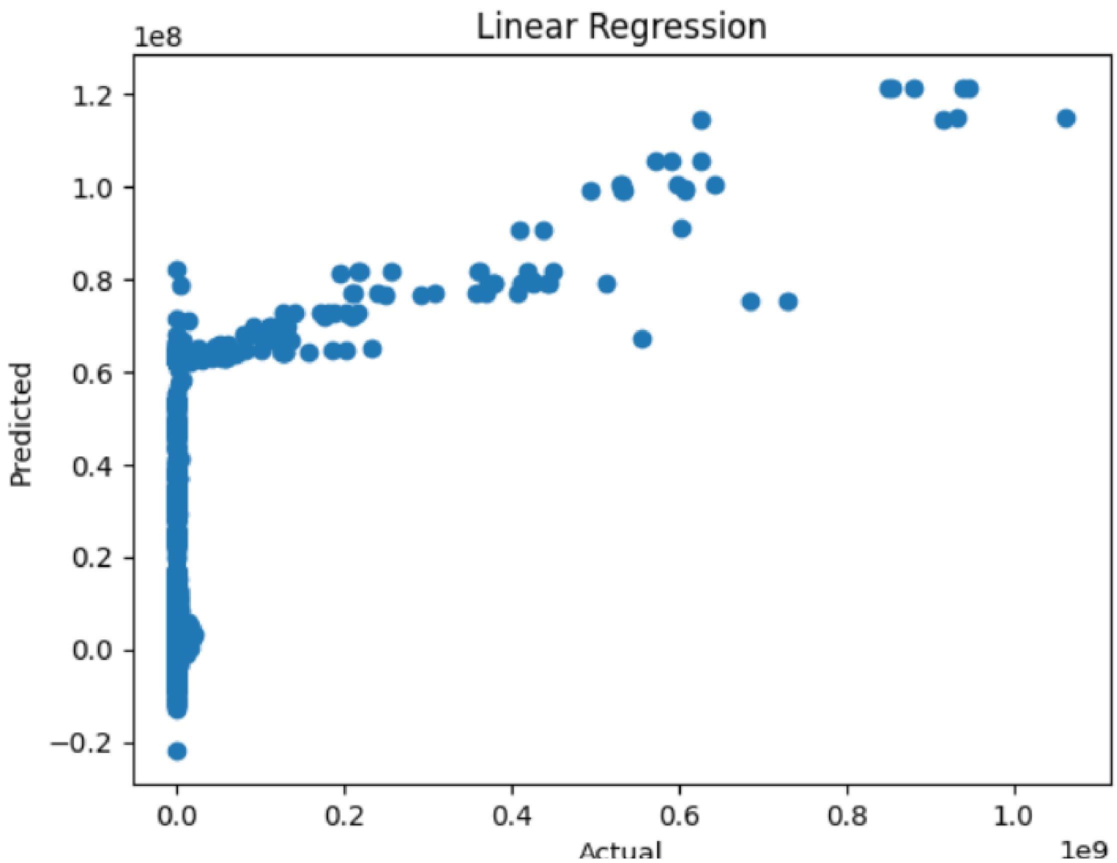
a) For Crop Production:

```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[ ] df=pd.read_csv("/content/crop_production.csv")
```

```
[ ] df
```

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0



The difference between actual and predicted values.

Random forest Regressor

```
[ ] from sklearn.ensemble import RandomForestRegressor
    model = RandomForestRegressor(n_estimators = 11)
    model.fit(x_train,y_train)
    rf_predict = model.predict(x_test)
    rf_predict
```

```
array([ 4533.55454545,   670.20909091,  2394.27272727, ...,
         257.54545455, 13055.54545455,   109.09090909])
```

```
[ ] model.score(x_test,y_test)
```

```
0.9720809047329542
```

```
[ ] # Calculating R2 score

    from sklearn.metrics import r2_score
    r1 = r2_score(y_test,rf_predict)
    print("R2 score : ",r1)
```

```
R2 score :  0.9720809047329542
```

Here the Random Forest Regressor accuracy is more we can use this algorithm for yield prediction. After developing the website using Django we can deploy the model into it then we can use.

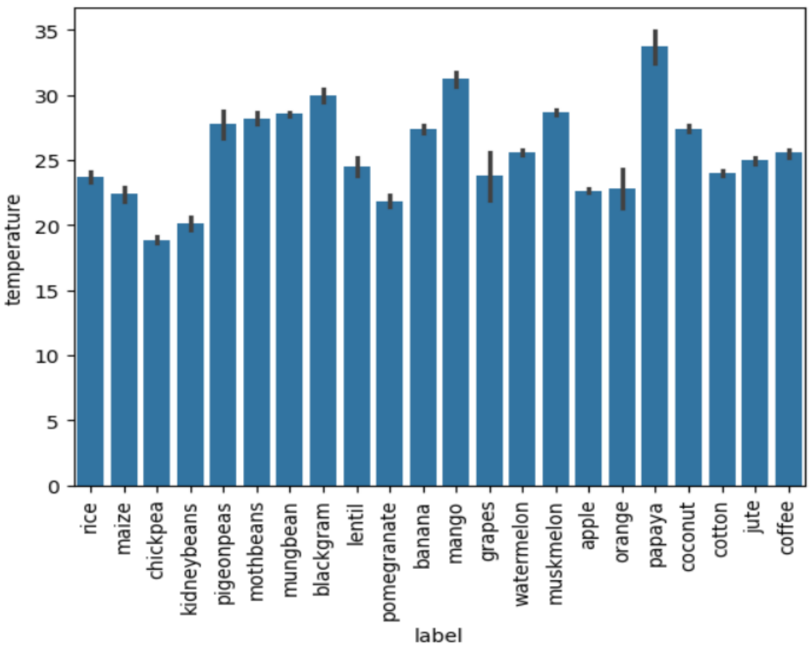
b) **For Crop Recommendation:**

```
# importing necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

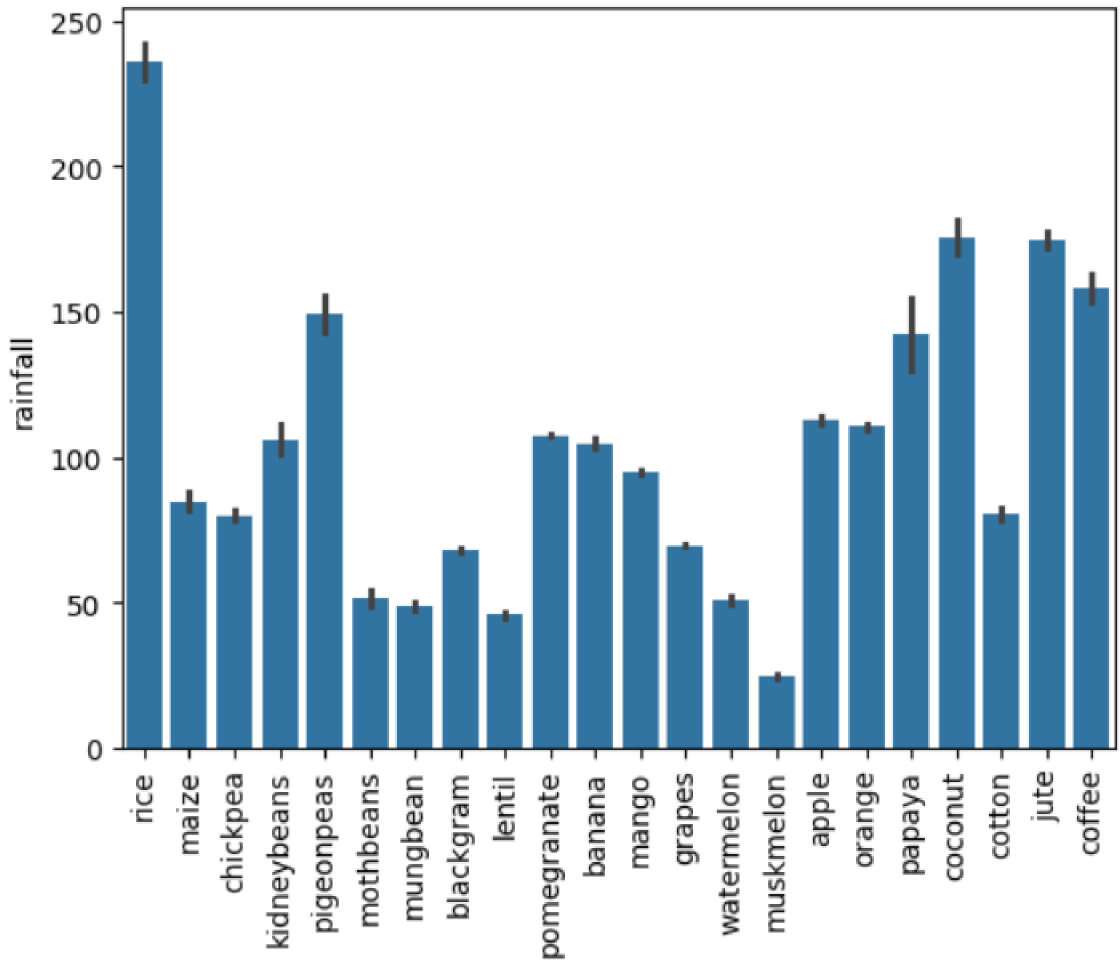
```
[ ]  # loading the dataset
     df=pd.read_csv("/content/Crop_recommendation.csv")
     df
```

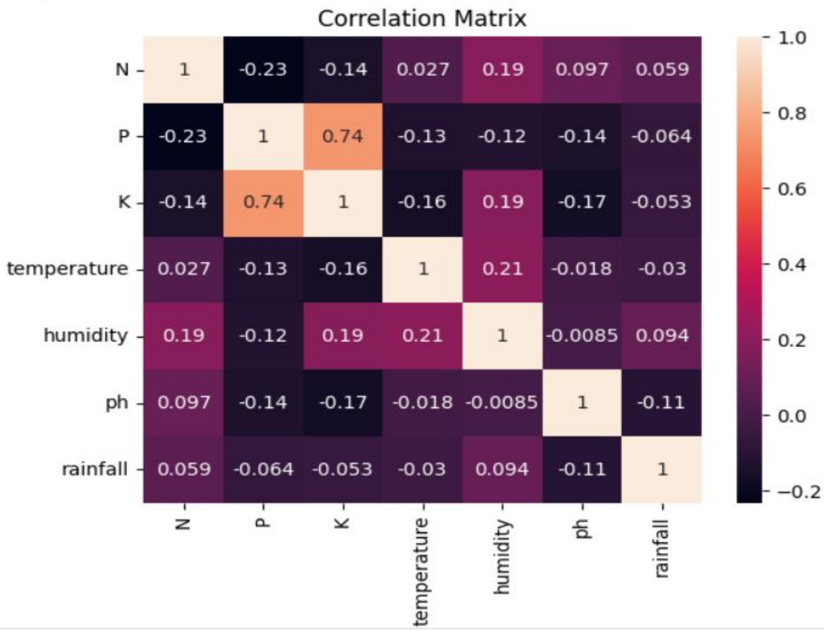
	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

```
[ ]  sns.barplot(x=df["label"],y=df["temperature"])
     plt.xticks(rotation = 90)
```

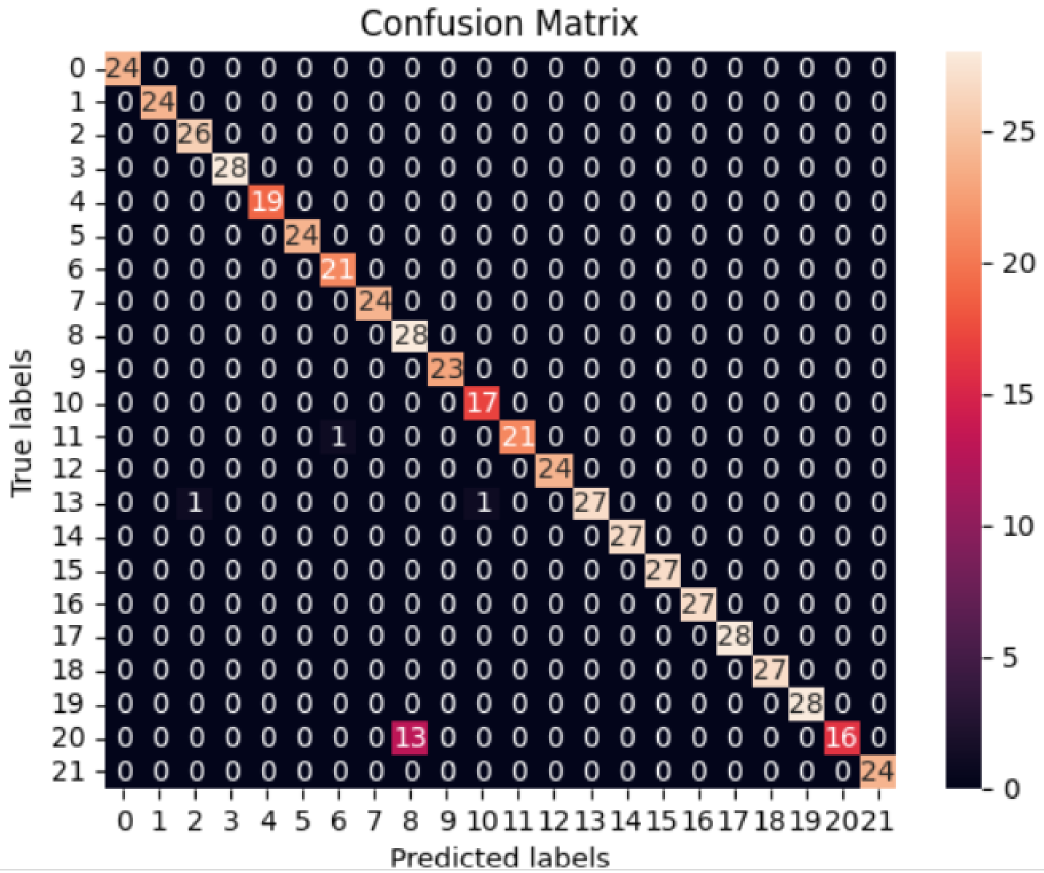


```
[ ] sns.barplot(x=df["label"], y=df["rainfall"])
    plt.xticks(rotation = 90)
```





```python
# creating a confusion matrix
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test.values.argmax(axis=1), gnb_pred.argmax(axis=1))
#cm = confusion_matrix(y_test, gnb_pred)
ax= plt.subplot()
sns.heatmap(cm, annot=True, fmt='g', ax=ax);
# labels, title and ticks
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
```



Decision Tree Algorithm

```python
# Calculating Accuracy
from sklearn.metrics import accuracy_score
a2 = accuracy_score(y_test.values.argmax(axis=1), decision_pred.argmax(axis=1))
a2
```

```
0.9272727272727272
```

Random Forest Algorithm

```
[ ] # Calculating Accuracy
    from sklearn.metrics import accuracy_score
    a3 = accuracy_score(y_test.values.argmax(axis=1), forest_pred.argmax(axis=1))
    a3
```

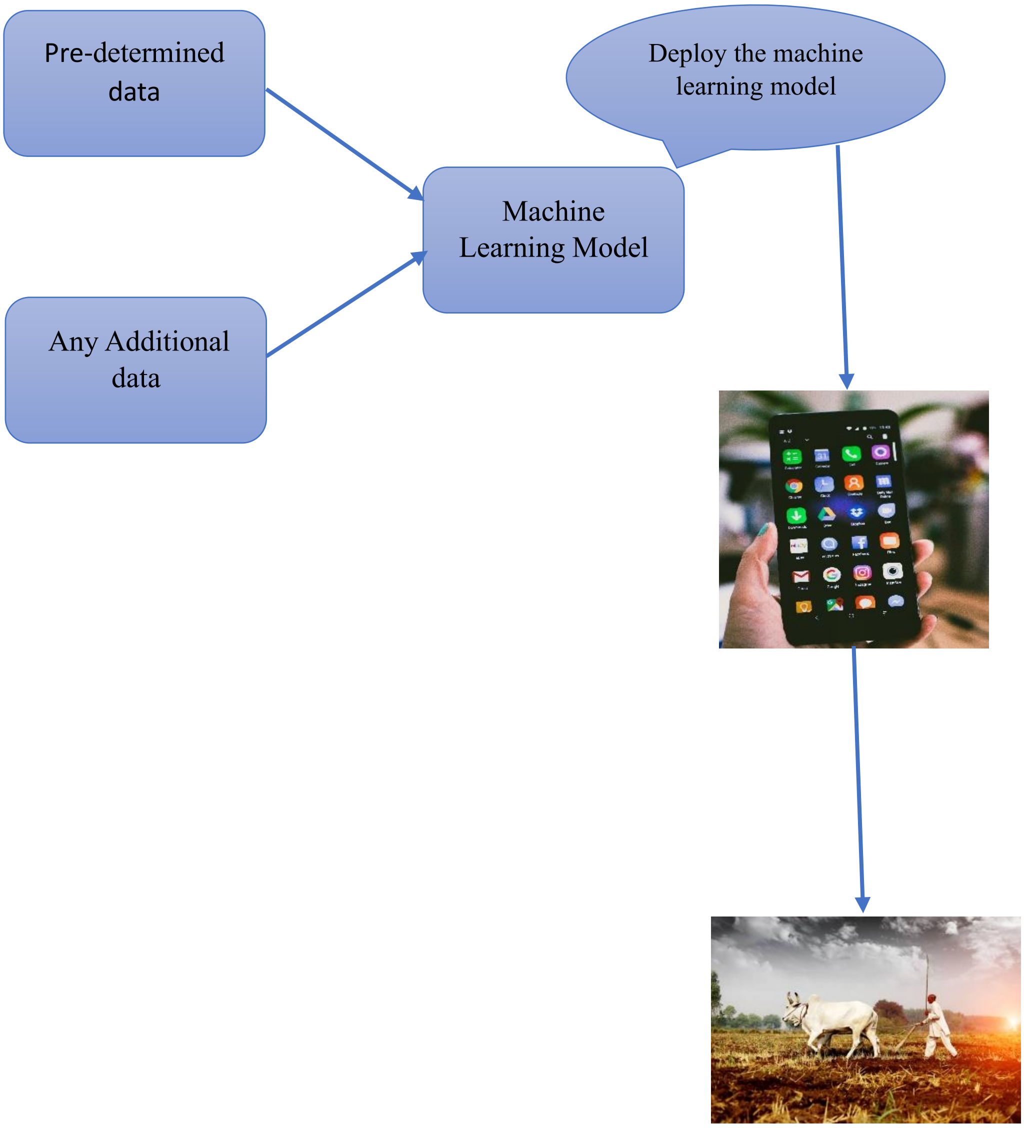
```
0.9709090909090909
```

KNN Algorithm

```
[ ]
    # Calculating Accuracy
    from sklearn.metrics import accuracy_score
    a4 = accuracy_score(y_test.values.argmax(axis=1), knn_pred.argmax(axis=1))
    a4
```

```
0.9781818181818182
```

11) FINAL PRODUCT PROTOTYPE:



11.1) PRODUCT DETAILS:

- The developed product/service will be a web-based platform accessible to farmers, agricultural consultants, and environmental scientists.

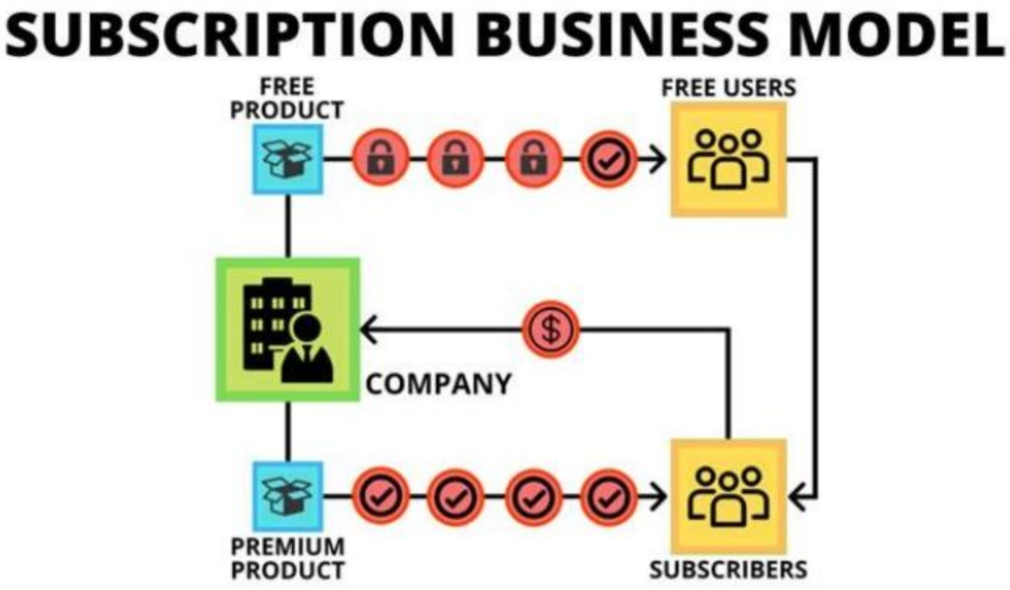
- Users can upload data or connect IoT sensors to the platform for real-time monitoring.

- ML algorithms such as Random Forest, Support Vector Machines, naïve byes, logistic regression can be implemented using libraries like Scikit-learn or Tensor Flow.

- The platform will provide customizable reports and recommendations for optimizing crop yield, fertilization, and soil conservation practices.

- **Back-end:**
- • This involves data collection, pre-processing and integrating the model with the web app.
- • The data entered by the customer should also be collected and stored with the customer's permission.
- **Front-end:**
- • Front-end plays a crucial role as it is the interface with which the customer will be working.
- • The web app could have 2 pages. This first page to login and the second page enters the data.

BUSINESS MODEL



The subscription model is used because of its pricing structure, in which a business will charge customers a recurring fee to access their product or service.

This model should be used by those companies in which the revenue strategy depends on a customer paying multiple and subscription based for prolonged access to a good or service.

OPERATING PLAN:

The important part of our operation is to have ML/DS engineers with a good amount of knowledge about the industry. The product developing team's size should be 3 to 4 where one of the members must be a full stack web developer and the remaining members must be ML engineers. It would be beneficial if almost all of the ML engineers had knowledge about the industry.

The time for developing the product must be decided after a meeting with the client and the team developing the product. Having a clear idea about the deadline is a must and based on that the team can accelerate certain parts of the developing process.

FINANCIAL EQUATION:

Let's assume we want to predict crop yield (Y) based on various market factors (X1, X2, ..., Xn):

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta n*Xn + \varepsilon$

Where:

Y is the predicted crop yield.

X1, X2, ..., Xn are the market factors such as weather conditions, commodity prices, soil quality, etc.

$\beta 0, \beta 1, \beta 2, ..., \beta n$ are the coefficients representing the effect of each market factor on crop yield.

ε is the error term.

12) CONCLUSION:

I would like to conclude that the crop yield is usually affected by different varieties and environmental parameter sets such as soil, climate, and crop management. Soil nutrient analysis is one of the biggest challenges faced by the researchers in producing the yield of the crop. Data cleaning and processing, missing value analysis, exploratory analysis, and model creation and evaluation were all part of the analytical process. Finally, we use a machine learning method to predict the crop, with varying outcomes.

The user-friendly web page built for estimating crop yield can be utilized by any user with their choice of the crop by giving data the farmer will be able to see a list of all possible crops, which will aid them in deciding which crop to cultivate.