**CROP RECOMMENDATION**

**Rishika.ch1, G.sanjana2**

**ABSTRACT**

Healthy and sufficient production of food is necessary as the population is increasing. Agriculture is playing a major part in the Indian economy. In recent times, many developments have been made, initially from the selection of crops to harvesting the crop. This paper presents the development and evaluation of a crop recommendation system designed to help farmers make better crop choices based on relevant environmental, soil, and market factors. For treatment, we used a machine learning-based approach that analyzed historical climate data, soil properties, and market data and recommended suitable crop choices for individual farmers. The design of the model is to collect data regarding features like P, N, K, Humidity, PH, Temperature, and Rainfall and predict suitable crops. After collecting the data, we do some preprocessing of the data. Then, we perform EDA (Exploratory Data Analysis) and find the relation between features and the label. Then we apply various ML models to foresee the outcome. The results showed that the yield and profitability of the experiment increased significantly. The implications of this study highlight the potential of data-driven crop recommendation systems to increase agricultural productivity, especially for smallholder farmers. The findings highlight the importance of the benefits of technology and information analysis in agriculture for food security and sustainable rural livelihoods. The implications extend to the widespread use of precision agriculture and the integration of digital tools into agricultural practices to address global food challenges.

**INTRODUCTION**

Precision Agriculture allows us to precisely use taken from user such as water, nutrient supplements, seeds, insecticides, etc. to increase the productivity of crops and maximize the yield or quality. Usually, farmers use traditional patterns for cultivating a crop but the farmer is dependent on soil characteristics and climate conditions. Predicting the right crop is a difficult task as soil and climatic conditions need support for a specific region where these crops grow. There is a significant demand worldwide for crop production as there is an increase in population. The farmer needs to know what crop should be grown in which region with suitable parameters for better yield. For this, we are going to use various ML techniques to predict the suitable crop grown with different features [1].

Agriculture depends highly on conditions of weather which means seasonal agriculture is mainly dependent on weather conditions naturally. Agriculture is reliant heavily on rainfall, farmers may not harvest as expected due to rainfall abundance. Climate changes affect production leading to food insecurity and hunger. To overcome these problems many solutions were provided in different studies [2]. To determine the whole economy of a country agriculture plays a significant role so it is necessary to increase production rate. For this many technological changes needed to be made. To have a high crop yield factors such as properties of soil, conditions of weather, water availability, soil, sunlight, pollution level, and wind should be taken into consideration [3].

Progress in measuring features like temperature, humidity, pH, and analytic parameters like Nitrogen, phosphorous, and potassium is shown by sensors [4]. During the coronavirus pandemic, there was a serious threat to food security and economic growth, to overcome these issues we need to develop solutions that are more reliable with very few human resources. These days farmers are also using modern technologies that provide quality information about crops. Even companies like Google and Amazon are attracted to these technologies of agriculture [5].

New innovative technologies such as hybrid products leads to unhealthy life. This is due to not having awareness about the cultivation of crops at the right place and time. This also may lead to food insecurity. Algorithms of supervised learning form a model that contains input and output values [6]. For agriculture, more than 60 percent of the country’s land is used. So, new technologies usage in the agriculture field is also necessary. The country’s profit depends on the profit of farmers. Farmers usually follow traditional methods for growing suitable crops in particular soil but they don’t have enough knowledge of soil nutrients like nitrogen, phosphorous, potassium, etc. So, they don’t get sufficient profit and good yield. For this, we have designed a system that will predict or recommend the crop suitable based on nitrogen, phosphorous, and potassium levels [7].

Accurate prediction of yield is a problem and could facilitate farmers to take precautionary measures to increase production. Crops that are either for food (rice, wheat, maize, etc.,) or commercial purposes (cotton, cashew, sugarcane, etc.) both are important [8]. Smart farming helps to work with drones, sensors, and satellites. Sensors access climate forecasts, topography, temperature, and acidity of soil. The detailed location and forecasts of climate can be accessed easily by sensors [9].

Precision agriculture helps farmers to get more profit with less labor. By applying proper calculations and estimations precision agriculture can help farmers. But the predictions made must be accurate or else it may lead to major loss. Many people are researching to make accurate predictions. Ensembling can also be used here [10].

To satisfy a wide range of human needs we need to use modern technological advancements. Data science and machine learning are showing their versatility to satisfy the needs of humans and the needs of agriculture [11]. Precision agriculture is helping farmers to predict disease in advance so that farmers do not get much loss. Spraying more fertilizers than required can lead to crop damage or even the nutrients of the soil can be reduced [12].

Wrong crop selection will lead to less yield. AI and deep learning tools can be used to provide more production in the agriculture sector. Artificial neural networks are used to select crops with the best yield [13].

It is complicated to say which plant to grow in a specific season. The wrong decision leads to an impact on the economy of agriculture. A combination of ML models can be used to select the most suitable crop to be sown [14]. While using the IOT system the prediction framework is fed with collected data from sensors for obtaining suggestions. It is very important to meet the goals of sustainable agriculture in our country. [15].

**LITERATURE REVIEW**

TABLE

Table 1 – Literature Survey

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author Name** | **Model Used** | **Parameters Used** | **Merits** | **Limitations & Drawbacks** |
| Gopi et al. (2022) | EO, KELM,  Random Forest | Accuracy,  Precision, Recall, Specificity, PR-Score, ROC-Score, F1-Score,  MCC. | ->The highest performance was given as 97.91%. | ->It is a very long process. |
| Kuradusenge et al. (2023) | Random Forest, Regression, Support Vector Regression | R^2,  MAE, RMSE | ->The RF model has shown superior performance on the data as shown by its R2, MAE, and RMSE values for both crops. | ->Early information is not provided to prevent the crops. |
| Senapathy et al. (2023) | IoTSNA-CR, MVSM-DAG-FFO,  SVM, RF,  Naïve Bayes, ANN,  FFO,  ADT,  Ada, AdaBoost | Accuracy,  Precision,  Recall,  F-Score | ->It provided a very good improvement in increasing crop yield. | ->Data collected is very little. |
| Bakthavatchalam et al. (2022) | IOT,  Data Mining,  SVM | Accuracy,  Weighted Avg ROC | ->It is user-friendly for the illiterate also. | -> IoT-based precision farming is not available. |
| Bouguettaya et al. (2022) | UAV,  CNN,  R-CNN, SVM,  ANN | Precision,  Accuracy,  SLIC,  FCN | -> UAV-based data helped to predict easily. | -> Classifying based on UAV imagery is difficult. |
| Bondre et al. (2019) | SVM, Random Forest. | Accuracy,  Error | ->High yield is obtained based on suitable location for crops. | ->There is no application or website to check the image and predict the disease of crops. |
| Nischintha et al. (2020) | SVM | Euclidean distance,  Accuracy. | ->It helps by providing information about the nutrients required. | ->Automatic location prediction cannot be done. |
| Palanivel et al. (2019) | Linear Regression,  ANN,  SVM | Mean Squared Error,  Root Mean Squared Error, Mean Absolute Error | ->It improved the technology by using big data techniques. | ->Preparing data using big data techniques is not easy. |
| Palanivel et al. (2019) | DCNN, KNN, SNN, ANN, LDA,  SVM | RMSE, MAE,  RRSE,  R, Mean forecast error, Average cycle Timer, MAPE, RRSE,  Precision, recall, F-measure | ->Different types of models are used combinedly to predict accurately. | ->The images which are not in proper condition may degrade the model. |
| Rajak et al. (2017) | SVM, Naïve Bayes, ANN, Random Forest | Precision,  Principal Component analysis | ->It helps farmers to reduce chemical use in crop production. | ->There are no large number of attributes in the dataset. |
| Sani et al. (2023) | KNN, Decision Trees, SVM, Random Forest | Accuracy,  Mean square error | ->A large dataset model is trained and performance is measured using accuracy  Scores. | ->It leads to data diversity. |
| Dinge et al. (2018) | KNN,  SVM,  Decision Tree, Random Forest | Accuracy,  MSE,  F-Score | ->It helps to increase the sustainability of crops and gives good growth. | ->It requires a user feedback mechanism. |
| Nandwani et al. (2023) | Naïve Bayes,  Decision Tree, | Accuracy,  Precision,  F1-Score,  Recall,  cv\_scores | ->It increased the profit values for the farmers. | ->It cannot provide a more comprehensive recommendation. |
| Desai et al. (2022) | Random Forest | Precision,  Accuracy,  Recall,  F1-Score,  Support | ->GUI provides an intuitive dashboard and easy navigation. | ->It requires frequent data  Updates. |
| Anguraj et al. (2021) | Random Forest,  Naïve Bayes | MSE,  Precision | ->GUI provides suitable crops to grow for the desired output | ->We need to perform gradient descent to minimize the error. |

SUMMARY

Gopi et al. (2022) [1] have developed a model that combines equilibrium optimization and kernel extreme learning machine for crop recommendation, followed by random forest for yield prediction. Simulations demonstrate its superior performance with a maximum accuracy of 97.91%, outperforming other methods. This approach addresses the complexities of variable factors like soil type and climate in effective crop management. It combines EO and KELM for recommendation, and RF for yield prediction. EO improves KELM's performance. Kuradusenge et al. (2023) [2] highlight the importance of early information sharing for planning and reducing food insecurity, particularly in countries like Rwanda. The future direction involves integrating IoT for soil moisture and weather monitoring along with the RF model for better yield prediction. Senapathy et al. (2023) [3] introduced the IoT-based soil nutrient classification and crop recommendation model to suggest crops. It also aids in minimizing the use of fertilizers/pesticides in the soil to maximize the yield.

Bakthavatchalam et al. (2022) [4] have developed a model based on attributes such as Nitrogen, Phosphorus, potassium, pH, and humidity, rainfall and temperature to recommend the crop. The main objective of this case study is to find a model that foresee the high-productive yeild. Bouguettaya et al. (2022) [5] have shown a technology such as UAV-based remote sensing is used widely to gather different types of data that is used to obtain many precision agriculture applications such as classification of crops. Bondre et al. (2019) [6] have introduced and executed a system to foretell the yield of the crop from before data. It can be attained by using ML models such as Support Vector Machines (SVM) and Random Forest on agriculture dataset. Nischintha et al. (2020) [7] have spread the information about the crop that suits the specific land/area based on weather conditions. It provides the info about acquires the quality and quantity of nutrient suppliments, wanted seeds for yield, etc. Palanivel et al. (2019) [8] have proposed for foresee the crop production with the help of ML models in huge data solving paradigm.

Palanivel et al. (2019) [9] have developed techniques or methods using convolutional deep learning methods which gives an accuracy of 92.51%. Rajak et al. (2017) [10] have shown that using precision agriculture is characterized by collecting soil from farms by experts and testing in the lab. The data that is taken from the soil testing lab given to the suggested system is used to collect data and do an ensemble model with a majority voting technique using SVM and ANN to recommend a crop with high accuracy and efficiency. Sani et al. (2023) [11] have developed a model that is trained on a large dataset, and its performance is evaluated using accuracy scores. The proposed system offers valuable insights for farmers, researchers, and policymakers, aiding them in ineffective crop management and decision-making for land productivity improvement.

Dinge et al. (2018) [12] have developed a crop-suggested system using ML to bolster sustainable agriculture and yield optimization. By factoring in elements like nitrogen, phosphorus, potassium, and humidity, the system advises optimal crops for specific sites. Nandwani et al. (2023) [13] have shown that the research underscores the perspective of ML algorithms in farming and suggests augmenting the system with market data, fertilizer availability, and pest insights for more robust recommendations. The successful implementation highlights the project's viability. Desai et al. (2022) [14] have shown that this study employs Random Forest and other machine learning methods for accurate crop prediction. The recommendation system promotes modern tools, precision farming, and educated decisions for increased profits, especially benefiting novice farmers. Anguraj et al. (2021) [15] have shown that This IoT and ML integrated system proves invaluable to farmers, aiding in informed decision-making for enhanced productivity.

PROBLEM STATEMENT

* Agriculture is a fundamental pillar of global food production. The increasing global population, and changing climate patterns have intensified the pressure on farmers to maximize crop productivity while minimizing resource wastage. However, traditional farming practices often lack the data-driven insights necessary to adapt effectively to these challenges, resulting in suboptimal crop choices, reduced yields, and economic instability for agricultural communities.
* The main objective of our research is crop recommendation i.e., to predict the suitable crop.
* Various parameters affect the production of the crop such as N, P, K, temperature, humidity, pH, rainfall, etc. So, based on all these factors selecting a suitable crop is most efficient.
* The wrong selection of the crop may lead to food insecurity and create a big loss. So, selecting the right crop is a major and very crucial task.
* The purpose of our research:
* The main purpose is to increase the crop production.
* Providing food for such a large population is a big and challenging task. For this, we need to have a large production and this can be achieved only when a huge amount of production is done.
* The main problem is that crops require many parameters for having a better yield such as water requirement, soil quality, proper fertilizers, seeds, climatic conditions, etc.

**PROPOSED WORK AND METHODOLOGY**

**ARCHITECTURE**

Data Collection

Form Dataset

Data Preprocessing

**Perform EDA (Exploratory Data Analysis)**

**PERFORMANCE EVALUATION**

**PREDICTION**

Precision, recall, f1-score, support

ML MODELS

Fig-1 - Architecture

**Steps:**

1. Firstly, we need to collect data.
2. Form a dataset from the collected data.
3. Then we do data preprocessing on the dataset formed.
4. Then we perform EDA (Exploratory Data Analysis) on the dataset.
5. For the prediction process, we use different Machine Learning (ML) models.
6. Lastly, to evaluate the performance we use precision, Recall, F-Score, and Support.

FIG 1 – ARCHITECTURE

1. Data Collection and Preprocessing:

We must obtain the dataset by survey process conducted by various trusted sources and handle missing values in the data then we normalize the data or scale the numerical features. To evaluate the model, divide the dataset into training and test sets.

2. Exploratory Data Analysis (EDA):

To learn more about each consumer category, do EDA. Analyze exploratory data to learn more about the dataset.

3. Hypothesis Testing:/Statistical Analysis:

To find out if there are significant changes in the replies depending on the segment type or other parameters, use statistical tests (such the chi-squared test).

4.Model Selection:

We must use various regression algorithms. Some regression algorithms are CatBoost, Random Forest , Gradient Boosting (e.g., XGBoost) SV regression & Neural Networks.

5. Hyperparameter Tuning:

Using methods like grid search or random search, we optimize the hyperparameters of the selected model(s). To make sure the model is generalizable, use cross-validation.

6. Training The Model:

On the training dataset, run the chosen model or models.

7. Evaluation Metrics:

We choose evaluation metrics based on the problem (e.g., MSE, RMSE, r2 SCORE, MAPE).

8. Model Interpretation:

Based on the model selected, consider interpreting the results. Techniques like feature importance analysis, SHAP values, or LIME can help understand what features contribute most to a segment's responsiveness.

Algorithms used:

1.Random Forest Regression:

An approach to collective learning is Random Forest Regression. During training, several decision trees are constructed. The average (for regression) or mode (for classification) of each tree's contribution to the final prediction is used. It is renowned for its capacity in handling different types of data and reliability against overfitting.

2.Support Vector Regression (SVR):

SVR is a regression method that extends Support Vector Machines to issues related to regression. It locates a hyperplane that minimizes the error margin and most accurately matches the data. SVR works well at identifying complex patterns in data and is especially helpful when working with high-dimensional data.

3.Gradient Boosting Regression:

Regression is an ensemble method that repeatedly constructs several decision trees. Each tree fixes the errors of the before ones, creating a model that is precise. This is renowned for its flexibility and predictive capacity and can handle both regression and classification jobs.

4.XGBoost Regression:

A advanced use of gradient boosting is called XGBoost (Extreme Gradient Boosting).It is appropriate for huge datasets because of its great efficiency and flexibility. To avoid overfitting, XGBoost employs normalized trees and performs well on regression problems.

5.CatBoost Regression:

It is a is a gradient boosting method designed specifically to deal with categorical information. It uses techniques like ordered boosting and automatically encodes categorical data. CatBoost is known for its performance, speed, and applicability to a variety of problems related to regression.

6.Neural Networks Regression:

Artificial neural networks are used in neural network regression to represent complicated connections in data. It is made up of layers of linked neurons that can recognize patterns in input data. Although very adaptable and capable of collecting complex patterns, neural networks may need careful tuning and considerable data.

**EXPERIMENTAL WORK**

Here, we usually follow the architecture and perform the steps according to the architecture.

We collected data using various parameters and formed a dataset.

**DATASET DESCRIPTION:**

The dataset contains 8 columns and 2199 rows.

The dataset contains 7 parameters:

N=Nitrogen, P=Phosphorous, K=Potassium, Temperature, Humidity, Ph, Rainfall.

All these parameters are numeric.

The last column represents the label (target) and shows the type of crop suitable to grow in the given parameters.

The label values are categorical.

These label values may be rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mung bean, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, coffee.

Dataset link: <https://www.kaggle.com/datasets/altafk/real-time-crop-recommendation-system>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset Name** | **Number of rows** | **Numbe**  **of columns** | **Features** | **Target** |
| Real\_Time\_Crop\_  Recommendation system | 2199 | 8 (Features + Target) | N=Nitrogen,  P=Phosphorous,  K=Potassium,  temperature,  humidity,  pH,  rainfall | Label  Values:  Rice, maize, chickpea, kidney beans,  pigeon peas, moth beans, mung beans,  black gram, lentil, pomegranate, banana, mango, grapes, watermelon,  muskmelon, apple, orange, papaya,  coconut, cotton, jute, coffee |

Table 2 – Dataset Description

Models used for Prediction Purpose:

1. SVM (Support Vector Machine).

2. Random Forest.

3. KNN (K-Nearest Neighbours.

4. Decision Tree.

5. Logistic Regression.

6. Gradient Boosting.

7. light gbm

8.neural network

1. SVM:

SVM is a Supervised machine learning algorithm.

It can be used for both classification and regression.

It is a parametric method.

Supporting vectors are used to find the decision plane (hyperplane).

The mathematical formula for SVM is: b+∑ αjXTXj

2. Random Forest:

Random forest is a supervised machine-learning algorithm.

It can be used for both regression and classification.

Random Forest is an ensemble learning algorithm.

Random Forest combines more than one algorithm of the same or different type for classification.

It builds several decision trees and takes the best one for classification purposes and the average one for regression purposes.

Random forest increases the overall result.

Mathematically formula for the random forest is:

Random forest uses bagging.

^ B

F =1/B (∑ fb(x’))

b=1

sigmoid = √∑Bb=1 (fb(x’)-f^)2/(B-1)

3. KNN:

It is a supervised machine-learning algorithm.

It can be used for both regression and classification.

It is a non-parametric method.

We do not need to form an equation. It directly learns from data.

Mathematically to find nearest neighbours we use the distance formula. The mathematical formula of KNN is: √(x2-x1)2+(y2-y1)2

4. Decision Tree:

It is a supervised machine-learning algorithm.

It can be used for both regression and classification.

It is a non-parametric method.

We do not need to form an equation. It directly learns from data.

Decision Trees are like tree-like structures.

They make decisions by recursively splitting data based on features to create leaf nodes with class

labels.

Entropy: E(T)= -Plog2P

Information Gini: E(T) – E (T, X)

5. Logistic Regression:

It is a supervised machine-learning algorithm.

It is a parametric method.

It is a binary classifier.

It can be used for classification purposes only.

When Features are continuous values and the target is of categorical values then we use LR-logistic regression.

Logistic regression is a classification algorithm used for binary and multiclass categorization purposes.

It models the likelihood of an case belongs to a selected class.

Sigmoid = 1/1+e-Z

Cost function error = [-y log(yp) + (1-y) log (1-yp)

6. Gradient Boosting:

Gradient boosting which is popularly known as Gboost is a type of boosting method.

Gradient boosting comes under ensembled learning.

In boosting we usually train or execute more than one machine learning model one after the other.

It is a method known for its accuracy and speed of prediction for datasets that are difficult and big.

It helps minimize the error of bias in the model.

The mathematical formulas of gradient boosting are:

n

F0 (x)=arg min ∑ L(yi ,ɣ)

ɣ i=1

n

L = 1/n ∑ (yi - ɣ)2

i=1

7.light gbm:

These are usually comes under decision tree family.

ranking can be done using light gbm and its main focus is ti improve the performance.

8.Neural Networks:

It comes under ML and DL.

It is inspired from the human brain.

For performance evaluation, we use different methods:

1. Precision

2. recall

3. f1-score

4. support

1. precision:

The similarity of 2 or more measurements is shown by precision.

It is not dependent on accuracy.

The mathematical equation for precision is:

Precision = True positives // (True positives + False positives) = TP / (TP + FP)

2. recall:

It calculates the ability of the model to capture all positive instances.

It is the ratio of true positives to the true total actual positives.

The mathematical formula for the recall is:

Recall = TP / (TP + FN)

3. f1-score:

The f1-score is the HM-harmonic mean of recall and precision.

It balances precision and recall and is useful when we want to find a compromise between those two metrics.

The mathematical equation for the f1-score is:

F1 Score = 2 \* (Precision \* Recall) // (Precision + Recall)

4. support:

Support tells us about each class's number of occurrences.

It is the true response of several samples.

Accuracy:

It calculates the quantity of correctly classified instances out of all circumstances.

The mathematical formula for the accuracy is: Accuracy = (TP + TN) // (TP + TN + FP + FN)

EXPERIMENTAL WORK

* The process undertaken this project are as follows:
* Set Up Infrastructure.
* New Python Notebook in Colab.
* Incorporate required libraries Sklearn, Numpy, Pandas, Matplotlib and Seaborn ,shap……etc.
* Dataset Preparation.
* Process the data.
* Exploratory Data Analysis.
* Normalize the data using Sklearn ’sMinMax Function.
* Split the data.
* Hypothesis Testing.
* Perform various ML models.
* Hyperparameter tuning & cross-validation performed.
* Develop Solution Model & improve using GridsearchCV.
* Ensemble top three performing models
* Now for the best model calculate SHap values.
* Mean absolute SHAP values.

**6. RESULTS**

After applying ML models, the results predicted are:

Table 3 - Accuracies

|  |  |
| --- | --- |
| **Model Applied** | **Accuracy Obtained** |
| SVM | 97 |
| Random Forest | 98 |
| KNN | 97 |
| Decision Tree | 98 |
| Logistic Regression | 95 |
| Gradient Boosting | 98 |

After performing the performance evaluation the results obtained are:

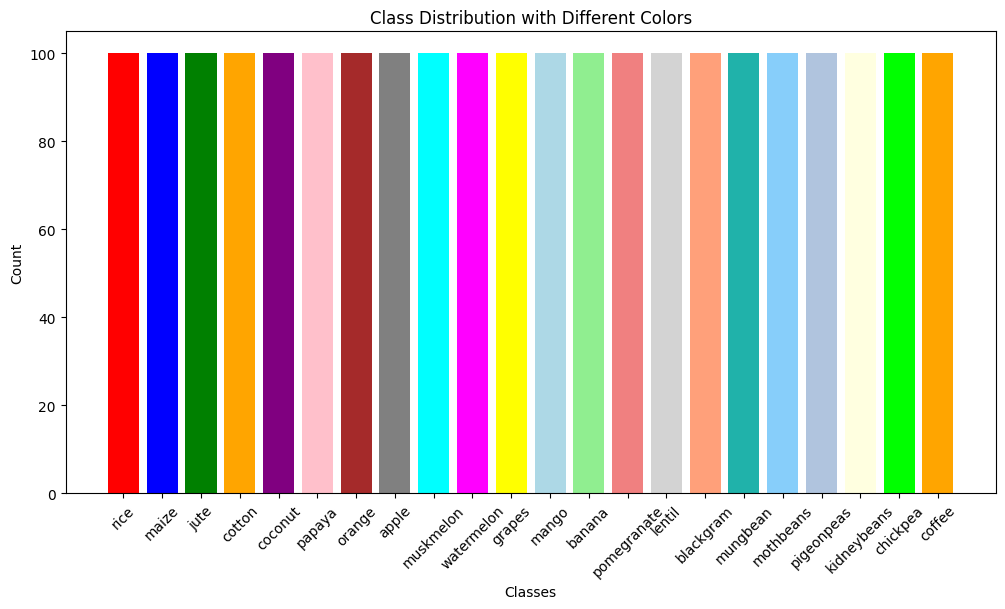
Table 4 – Performance Evaluation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Model used** | **Precision** | **Recall** | **F1-score** | **Support** |
| Accuracy  Macro Avg  Weighted Avg | SVM | 0.98  0.98 | 0.98  0.98 | 0.98  0.98  0.98 | 440  440  440 |
| Accuracy  Macro Avg  Weighted Avg | Random Forest | 0.99  0.99 | 0.99  0.99 | 0.99  0.99  0.99 | 440  440  440 |
| Accuracy  Macro Avg  Weighted Avg | KNN | 0.97  0.97 | 0.97  0.97 | 0.97  0.97  0.97 | 440  440  440 |
| Accuracy  Macro Avg  Weighted Avg | Decision Tree | 0.98  0.98 | 0.98  0.98 | 0.98  0.98  0.98 | 440  440  440 |
| Accuracy  Macro Avg  Weighted Avg | Logistic Regression | 0.95  0.95 | 0.95  0.95 | 0.95  0.95  0.95 | 440  440  440 |
| Accuracy  Macro Avg  Weighted Avg | Gradient Boosting | 0.98  0.98 | 0.98  0.98 | 0.98  0.98  0.98 | 0.98  0.98  0.98 |

* overall accuracy obtained by the SVM model is 97%.
* The overall accuracy obtained by the Random Forest is 99%.
* The overall accuracy obtained by the KNN is 97%.
* The overall accuracy obtained by the Decision tree is 98%.
* The overall accuracy obtained by the Logistic regression is 95%.
* The overall accuracy obtained by the Gradient Boosting is 98%.

Therefore, the best model is the Random Forest.

SUPPORTING GRAPHS:

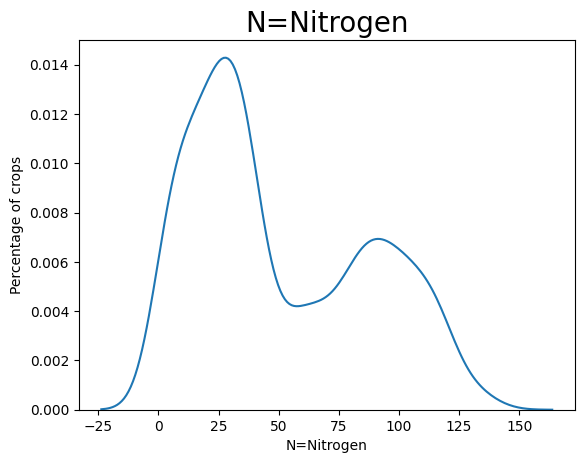


Graph 1 – Label Distribution

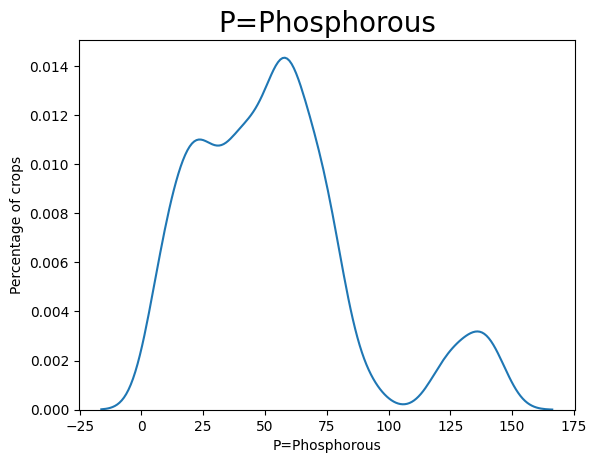
The discussion and analysis of the graph plotted above are shown as:

* The percentage of all the types of crops present on the label (i.e., target value) is plotted in the above graph.
* It represents the percentage of each crop to be predicted.
* From the above graph, we can also say that the distribution of every crop is the same.

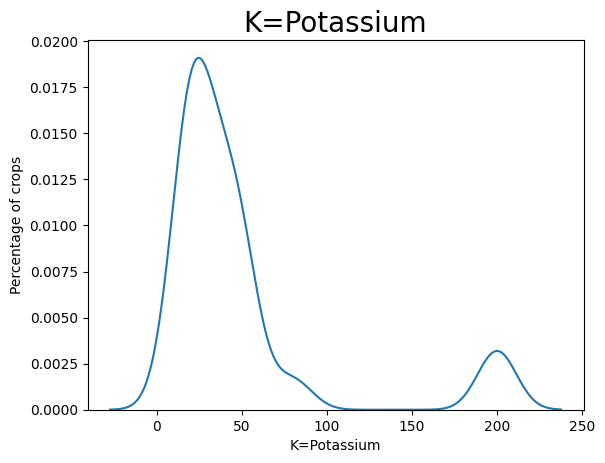
The distribution of various features over the percentage of crops to be selected is shown below:

****

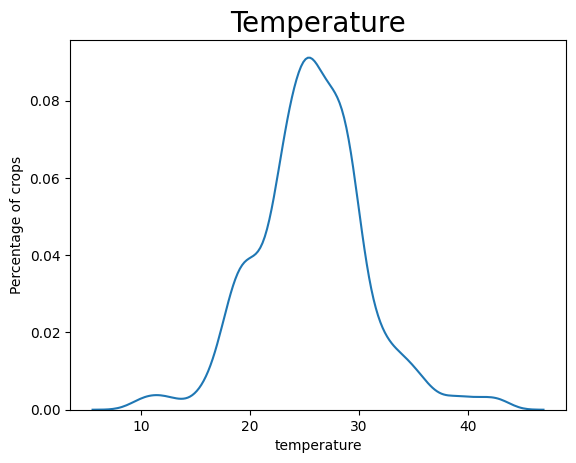
Graph 2 – Distribution of N=Nitrogen vs. Percentage of crops

****

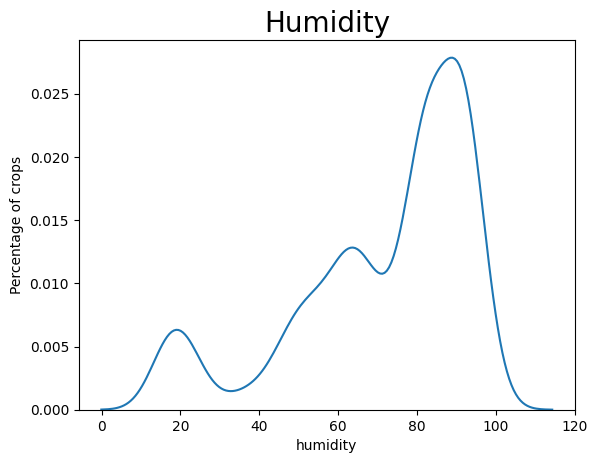
Graph 3 - Distribution of P=Phosphorous vs. Percentage of crops

****

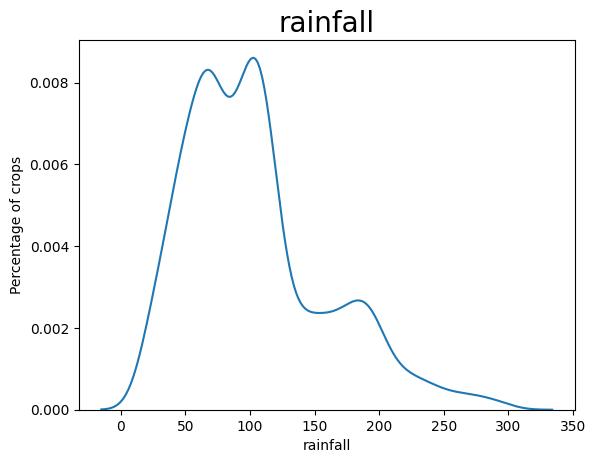
Graph 4 - Distribution of K=Potassium vs. Percentage of crops

****

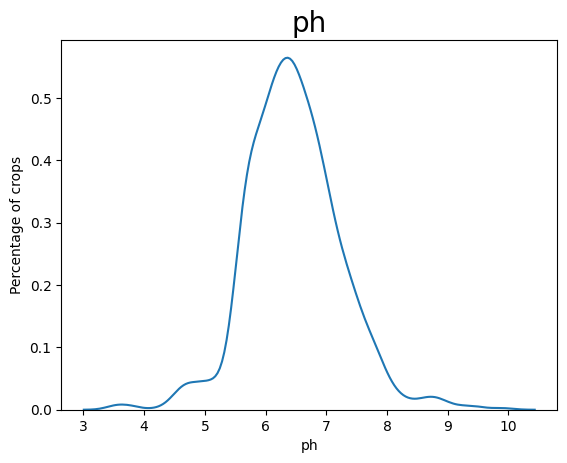
Graph 5 - Distribution of temperature vs. Percentage of crops

****

Graph 6 - Distribution of humidity vs. Percentage of crops



Graph 7 - Distribution of rainfall vs. Percentage of crops



Graph 8 - Distribution of pH vs. Percentage of crops

The discussion and analysis of various graphs shown above are:

The representation of the distribution of each feature or parameter to the percentage of the label or the target.

The graphs plotted are of different features vs. percentages of crops.

We have plotted 7 different graphs.

1. Distribution of N=Nitrogen vs. Percentage of crops - this graph shows the distribution of nitrogen over various crops.

2. Distribution of P=Phosphorous vs. Percentage of crops - this graph shows the distribution of phosphorous over various crops.

3. Distribution of K=Potassium vs. Percentage of crops - this graph shows the distribution of potassium over various crops.

4. Distribution of temperature vs. Percentage of crops - this graph shows the distribution of temperature over various crops.

5. Distribution of humidity vs. Percentage of crops - this graph shows the distribution of humidity over various crops.

6. Distribution of rainfall vs. Percentage of crops - this graph shows the distribution of rainfall over various crops.

7. Distribution of pH vs. Percentage of crops - this graph shows the distribution of pH over various crops.

Github link: <https://github.com/GundaSanjana/Dafe-code>

**7. CONCLUSION AND FUTURE SCOPE**

This document showed us how a suitable crop is predicted based on various parameters. We have developed an effective crop recommendation model to produce high yield here. We have performed EDA on the dataset. Then we applied different ML models to the dataset formed. The dataset selected for the model is a Real-time crop recommendation system.

The models applied to the dataset have shown a greater improvement in crop prediction and produced high yields. The accuracies of the predicted models were 97.9% for SVM, 99% for the random forest, 97% for KNN, 98.6% for the decision tree, 95% for the logistic regression, and 98% for the gradient boosting model. The best accuracy was found by the random forest. The above information helps predict the best crop suitable depending on various parameters.

The limitation of our model is it may lead to overfitting if the data is not distributed properly. Machine learning models have greatly improved crop recommendation by extracting meaningful data. Random forest provides prediction more accurately than any other model applied. Our future work is to explore the prediction based on IoT. We will aim to recommend a suitable crop using the IoT component. From now on, we also try to apply DL models instead of machine learning models. This can also help to produce optimal solutions.

**8**. **REFERENCES**

[1] Gopi, P. S. S., & Karthikeyan, M. (2023). Multimodal Machine Learning Based Crop Recommendation and Yield Prediction Model. Intell. Autom. Soft Comput, 36, 313-326.

[2] Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., ... & Uwamahoro, A. (2023). Crop yield prediction using machine learning models: the case of Irish potato and maize. Agriculture, 13(1), 225.

[3] Senapaty, M. K., Ray, A., & Padhy, N. (2023). IoT-Enabled Soil Nutrient Analysis and Crop Recommendation Model for Precision Agriculture. Computers, 12(3), 61.

[4] Bakthavatchalam, K., Karthik, B., Thiruvengadam, V., Muthal, S., Jose, D., Kotecha, K., & Varadarajan, V. (2022). IoT framework for measurement and precision agriculture: predicting the crop using machine learning algorithms. Technologies, 10(1), 13.

[5] Bouguettaya, A., Zarzour, H., Kechida, A., & Taberkit, A. M. (2022). Deep learning techniques to classify crops through UAV imagery: A review. Neural Computing and Applications, 34(12), 9511-9536.

[6] Bondre, D. A., & Mahagaonkar, S. (2019). Prediction of crop yield and fertilizer recommendation using machine learning algorithms. International Journal of Engineering Applied Sciences and Technology, 4(5), 371-376.

[7] Nischitha, K., Vishwakarma, D., Ashwini, M. N., & Manjuraju, M. R. (2020). Crop prediction using machine learning approaches. International Journal of Engineering Research & Technology (IJERT), 9(08), 23-26.

[8] Palanivel, K., & Surianarayanan, C. (2019). An approach for prediction of crop yield using machine learning and big data techniques. International Journal of Computer Engineering and Technology, 10(3), 110-118.

[9] Darwin, B., Dharmaraj, P., Prince, S., Popescu, D. E., & Hemanth, D. J. (2021). Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review. Agronomy, 11(4), 646.

[10] Rajak, R. K., Pawar, A., Pendke, M., Shinde, P., Rathod, S., & Devare, A. (2017). Crop recommendation system to maximize crop yield using machine learning technique. International Research Journal of Engineering and Technology, 4(12), 950-953.

[11] Sani, S. R., Ummadi, S. V. S., Thota, S., Muthineni, N., Swargam, V. S. S., & Ravella, T. S. (2023, May). Crop Recommendation System using Random Forest Algorithm in Machine Learning. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 501-505). IEEE.

[12] Dighe, D., Joshi, H., Katkar, A., Patil, S., & Kokate, S. (2018). Survey of crop recommendation systems. IRJET, 5, 476-481.

[13] Priyadharshini, A., Chakraborty, S., Kumar, A., & Pooniwala, O. R. (2021, April). Intelligent crop recommendation system using machine learning. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 843-848). IEEE.

[14] Desai, S., Joshi, M., & Mane, P. (2022, December). Intelligent Crop Recommendation System Using Machine Learning Algorithms. In International Advanced Computing Conference (pp. 277-289). Cham: Springer Nature Switzerland.

[15] Anguraj, K., Thiyaneswaran, B., Megashree, G., Shri, J. P., Navya, S., & Jayanthi, J. (2021). Crop recommendation on analyzing soil using machine learning. Turkish Journal of Computer and Mathematics Education, 12(6), 1784-1791.

[16] Kulkarni, N. H., Srinivasan, G. N., Sagar, B. M., & Cauvery, N. K. (2018, December). Improving crop productivity through a crop recommendation system using an ensembling technique. In *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)* (pp. 114-119). IEEE.

[17] Madhuri, J., & Indiramma, M. (2021). Artificial neural networks based integrated crop recommendation system using soil and climatic parameters. *Indian Journal of Science and Technology*, *14*(19), 1587-1597.

[18] Pawar, M., & Chillarge, G. (2018, April). Soil toxicity prediction and recommendation system using data mining in precision agriculture. In *2018 3rd International Conference for Convergence in Technology (I2CT)* (pp. 1-5). IEEE.

[19] Pande, S. M., Ramesh, P. K., ANMOL, A., Aishwarya, B. R., ROHILLA, K., & SHAURYA, K. (2021, April). Crop recommender system using machine learning approach. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1066-1071). IEEE.

[20] Jagadeeswari, M., Manikandababu, C. S., Kavya, S., Prasanna, R., & Madumitha, S. (2022, September). Artificial Intelligence-based Crop Recommendation System. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1127-1133). IEEE.