International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 09 – Issue 03 || May - Jun 2024 || PP. 72-76

Crop Recommendation System in India

Priyen Joshi[1], Sahil Mujawar[2], Mallinath Elekar[3]
[1,2,3]*MIT World Peace University*
Pune, Maharashtra, India

Abstract: Crops play an integral role in the lives of millions of people worldwide. This is especially true for an agriculture heavy country such as India. Not only do we depend on crops as a source of food, but avariety of crops contribute to maintaining soil fertility and structure.

Keywords: Crops, independent variable, dependent variable, confusion matrix, Python, Predictive Analytics

1. Introduction

Today, farming has come a long way from the traditional methods of the past, thanks to technological advancements like sensors, machinery, and IT. Currently, farms use cutting-edge technology such as temperature and moisture sensors, aerial imaging, GPS, and many IoT devices. These innovations help businesses and farmers work more efficiently, safely, and sustainably.

Digital agriculture and its technology have opened up new opportunities for data collection. Remote sensors, cameras, and connected devices can constantly gather data across entire farms, keeping an eye on things like The content present in the soil, temperature, humidity level as a percentage, soil pH value, and precipitation. The amount of data these sensors produce can be a lot to handle, but it gives farmers a better and more timely understanding of their farming environment.

Environmental data collected by remote sensors are processed by various machine learning algorithms and visual representations to provide insights to the farmer on decision-making and farm management. The more data input and the more advanced the algorithms get, the better they are at Cultivating crops. Therefore, farmers have the opportunity to improve their harvest by making more accurate decisions in the field.

In this regard, implementing a system that detects temperature, pH of the soil, and soil moisture levels, and then processing this data in specific algorithms, integrating it into a visual interface connected to various research modules, is able to predict the most feasible crop type with maximum gain for an agricultural land

2. Analytics

a) Descriptive Analytics

This step involves obtaining data on our variable (crops, in this case) and possible related factors (weather properties that may or may not affect crops).We, then visualize our dependent variable (crop) with each of the independent variables (other factors).We obtained our data from Kaggle.

The Kaggle dataset provides the selection of 22 crops across India. We use Python along with a Juypter notebook to extract data and gain the shape of a dataset's distribution in the form of a pivot table.From the system, we obtain data on the following:

Table1: Data Summarization

Label	Potassium	Nitrogen	Phosphorus	Humidity	pH	Rainfall	Temperature
apple	199.89	20.80	134.22	92.333383	5.929663	112.654779	22.630942
banana	50.05	100.23	82.01	80.358123	5.983893	104.626980	27.376798
blackgram	19.24	40.02	67.47	65.118426	7.133952	67.884151	29.973340
chickpea	79.92	40.09	67.79	16.860439	7.336957	80.058977	18.872847
coconut	30.59	21.98	16.93	94.844272	5.976562	175.686646	27.409892
coffee	29.94	101.20	28.74	58.869846	6.790308	158.066295	25.540477
cotton	19.56	117.77	46.24	79.843474	6.912675	80.398043	23.988958
grapes	200.11	23.18	132.53	81.875228	6.025937	69.611829	23.849575
jute	39.99	78.40	46.86	79.639864	6.732778	174.792798	24.958376
kidneybeans	20.05	20.75	67.54	21.605357	5.749411	105.919778	20.115085

lentil	19.41	18.77	68.36	64.804785	6.927932	45.680454	24.509052
maize	19.79	77.76	48.44	65.092249	6.245190	84.766988	22.389204
mango	29.92	20.07	27.18	50.156573	5.766373	94.704515	31.208770
mothbeans	20.23	21.44	48.01	53.160418	6.831174	51.198487	28.194920
mungbean	19.87	20.99	47.28	85.499975	6.723957	48.403601	28.525775
muskmelon	50.08	100.32	17.72	92.342802	6.358805	24.689952	28.663066
orange	10.01	19.58	16.55	92.170209	7.016957	110.474969	22.765725
papaya	50.04	49.88	59.05	92.403388	6.741442	142.627839	33.723859
pigeonpeas	20.29	20.73	67.73	48.061633	5.794175	149.457564	27.741762
pomegranate	40.21	18.87	18.75	90.125504	6.429172	107.528442	21.837842
rice	39.87	79.89	47.58	82.272822	6.425471	236.181114	23.689332
watermelon	50.22	99.42	17.00	85.160375	6.495778	50.786219	25.591767

b) Diagnostic Analytics

Diagnostic analytics is closely interlinked with descriptive to the point that in most cases, they are considered to be conducted concurrently. This involves using the visualizations developed earlier to understand the relationships between our independent and dependent variables.

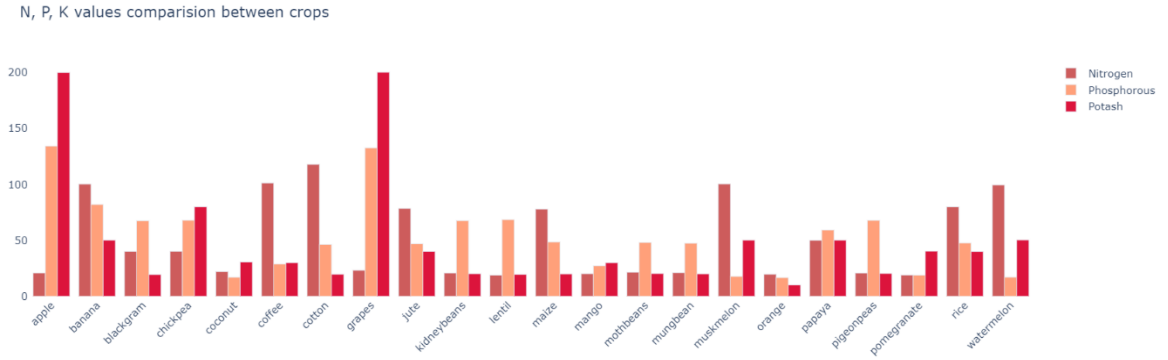

Figure 1: Relationship between dependent variable (crops) and independent variable (N,P,K)

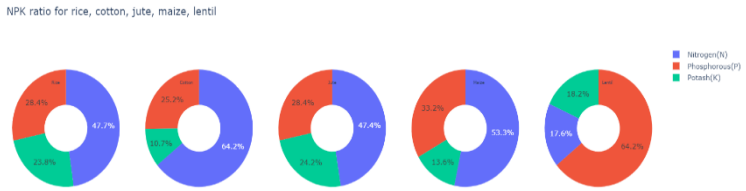

Figure 2: NPK ratio for rice, cotton, jute, maize, and lentils

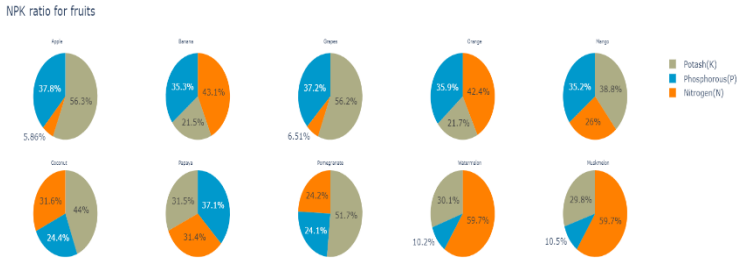

Figure 3: NPK ratio for fruits

To achieve the correlation between the two features, we use a heatmap to better understand it
A value close to 1 indicates a strong positive correlation between the two features.
A value close to -1 indicates a strong negative correlation between the two features.
A value close to 0 indicates little or no correlation between the two features.

For example, the correlation between temperature (K) and humidity is -0.23, which suggests a weak negative correlation. This means that as the temperature increases, humidity tends to decrease slightly, and vice versa.

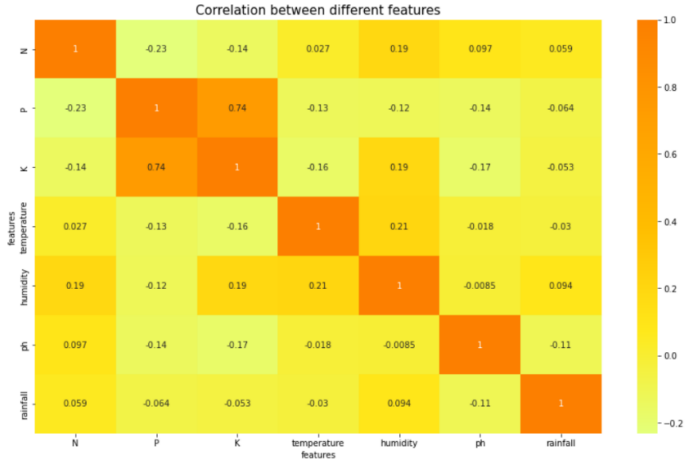

Figure 4: Heatmap (Correlation between two features)

c) Predictive Analytics

The third and last step that we have taken is integrating data into an appropriate program so that we can come up with a prediction model. Considering the enormous amount of data that we took into account, standard analytical programs like MATLAB and Microsoft Excel were inappropriate. For this reason, we imported the data by the use of Python programming language and its 'pandas' framework. The first stage in creating a predictive model is data cleaning. This entails eliminating outliers and undefined null values.

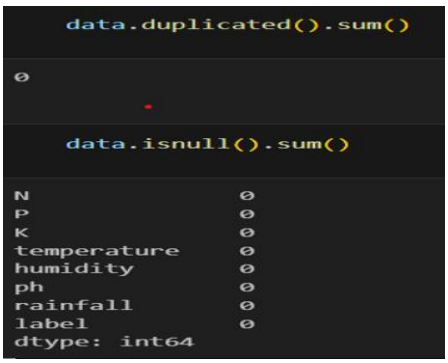

Figure 5: Analyzing null values

As observed, the dataset contains no null values.

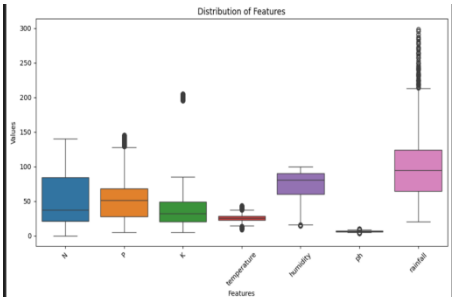

Figure 6: Analysis of outliers

The line inside the box represents the median value of the data.
Whiskers: The lines extending from the box represent the range of the data within 1.5 times the interquartile range (IQR) from the quartiles.
Outliers: Points outside the whiskers are considered outliers and are plotted as individual

Afterscaling our data, we constructed a random forest regression model in the Jupyter Notebook

3. Prediction Model and Results

After scaling our data, we proceeded to construct multiple prediction models in Jupter Notebook using the dataset obtained from Kaggle. After thorough testing, we determine that a random forest regression model has the best performance out of 'LightGBM', 'Decision Tree', and 'Logistic Regression'.

The 'train_test_split' function fromscikit-learn splits your data into training and testing sets, with 33% of the data reserved for testing for machine learning tasks. It accepts input features and target variables and splits these into separate training and testing sets. The 'test_size' defines the percentage of the dataset to be used for testing; meanwhile, the shuffle ensures that randomization is done prior to splitting. Setting 'random_state' allows reproducibility. An example usage would be by importing 'train_test_split', passing X and y, and storing the split data into 'X_train', 'X_test', 'y_train', and 'y_test'.

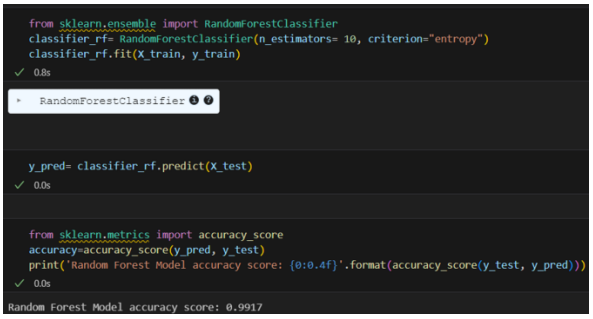

Figure 7: Accuracy score of the model

To judge the performance of our model, a confusion matrix is used in classification tasks to evaluate the performance.

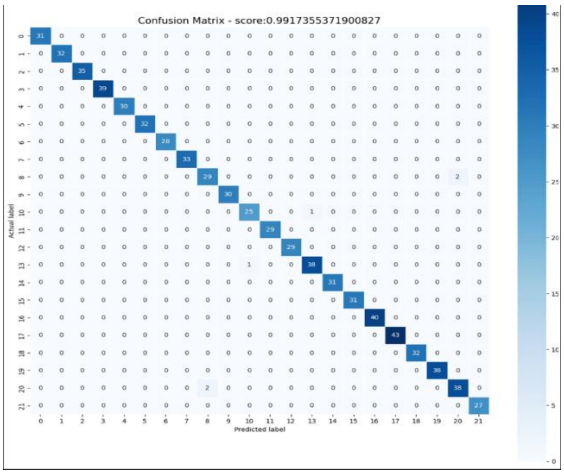

Figure 8: Confusion Matrix

These metrics assess how well the model can correctly classify instances across different classes. Here, high scores denote accurate predictions in the actual outputs. A precision score of 1.00 signifies that all of the predictions for the class were correct.

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```
0.0s

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	31
banana	1.00	1.00	1.00	32
blackgram	1.00	1.00	1.00	35
chickpea	1.00	1.00	1.00	39
coconut	1.00	1.00	1.00	30
coffee	1.00	1.00	1.00	32
cotton	1.00	1.00	1.00	28
grapes	1.00	1.00	1.00	33
jute	0.94	0.94	0.94	31
kidneybeans	1.00	1.00	1.00	30
lentil	0.96	0.96	0.96	26
maize	1.00	1.00	1.00	29
mango	1.00	1.00	1.00	29
mothbeans	0.97	0.97	0.97	39
mungbean	1.00	1.00	1.00	31
muskmelon	1.00	1.00	1.00	31
orange	1.00	1.00	1.00	40
papaya	1.00	1.00	1.00	43
pigeonpeas	1.00	1.00	1.00	32
pomegranate	1.00	1.00	1.00	38
rice	0.95	0.95	0.95	40
watermelon	1.00	1.00	1.00	27
accuracy			0.99	726
macro avg	0.99	0.99	0.99	726
weighted avg	0.99	0.99	0.99	726

Figure 9: Classification Report

Our model can recommend a crop according to the independent variables input by the user.

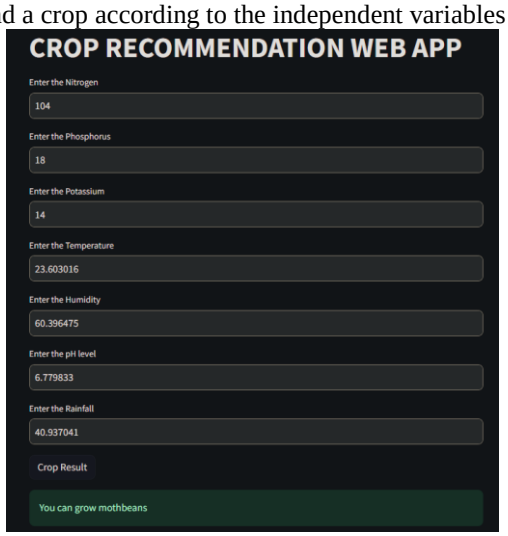

Figure 10: GUI

We utilized Python and its “streamlit” package to build a basic GUI that allows the user to inputindependent values and recommends the output with the click of a button.

4. Future Scope

It is capable of many more functions for the system. For now, it takes the required environmental parameters as input and suggests a crop that would be perfect to grow. Currently, the system takes inputs from all environmental factors; however, as an extra feature, an algorithm can be used to predict one factor based on two other factors. For example, predict the pH of the soil given sunshine and soil moisture.

5. Conclusion

Through this research project, we studied the importance of nutrient requirements and favorable environmental conditions, for different crops and devised a solution to overcome these by integrating predictive analytics techniques from the new and upcoming fields of machine learning and big data analytics.

6. References

[1] Kamilaris, Andreas, Francesc X. Prenafeta-Boldú. "A review of the use of convolutional neural networks in agriculture." The Journal of Agricultural Informatics, vol. 10, no. 2 (2019)

[2] Wolfert, Sjaak, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt. "Big data in smart farming – A review." Agricultural Systems 153 (2017): 69-80.

[3] Khan, Zohra, Amir Manzoor, and AdeelRafiq. "Application of Predictive Analytics in Agriculture: Techniques, Challenges, and Future Directions." IEEE Access 9 (2021): 125593-125609.

[4] Wolfert, Sjaak, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt. "Big data in smart farming – A review." Agricultural Systems 153 (2017): 69-80.