CROP RECOMMENDATION USING WEATHER AND SOIL CONTENT

NAME OF THE AUTHOR:

RAGHAV GUPTA

https://www.kaggle.com/code/theeyeschico/crop-analysis-and-prediction/notebook#MODEL-SELECTION

TABLE OF CONTENTS

➡️ Purpose Statement

➡️ Tell Your Story (with Data)

➡️ Model Selection

➡️ Conclusion

WHAT ARE
WE TALKING
ABOUT ?

OBJECTIVE

This project is the user creates a forecasting model to suggest the most suitable crops to grow on a given farm based on various parameters.

ABOUT DATASET

This dataset was build by augmenting datasets of rainfall, climate and fertilizer data available for India.

DATA FIELDS

N - Ratio of Nitrogen Content In Soil

P - Ratio of Phosphorous Content In Soil

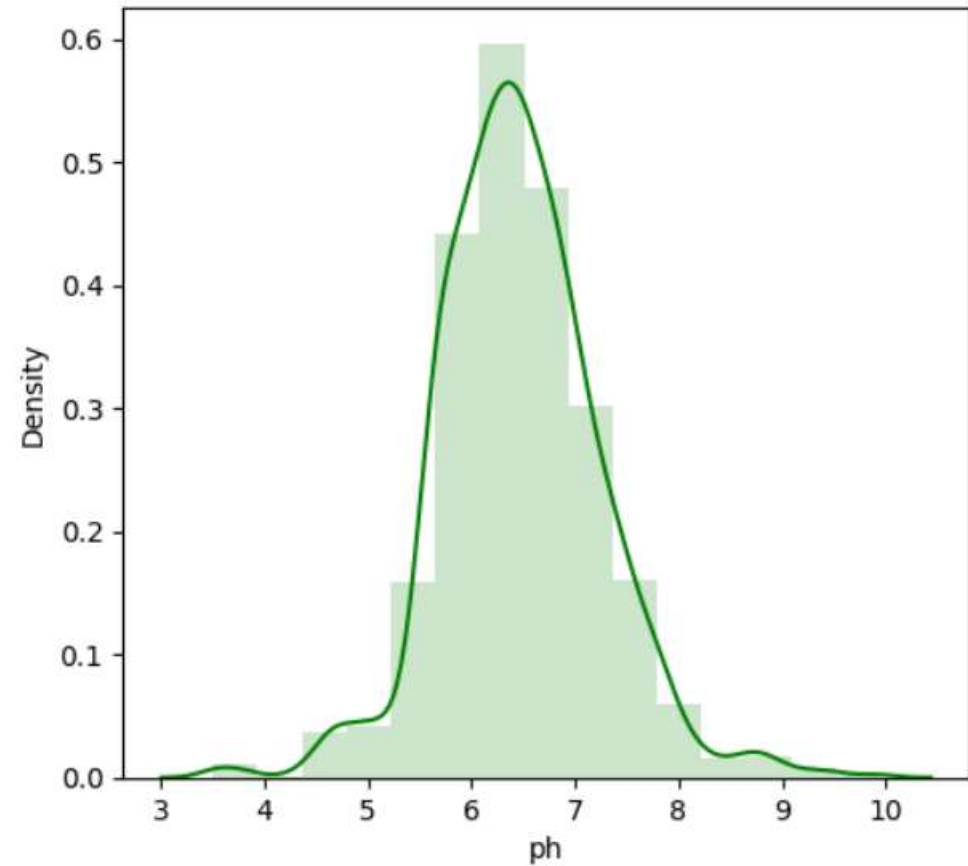K - Ratio of Potassium Content In Soil
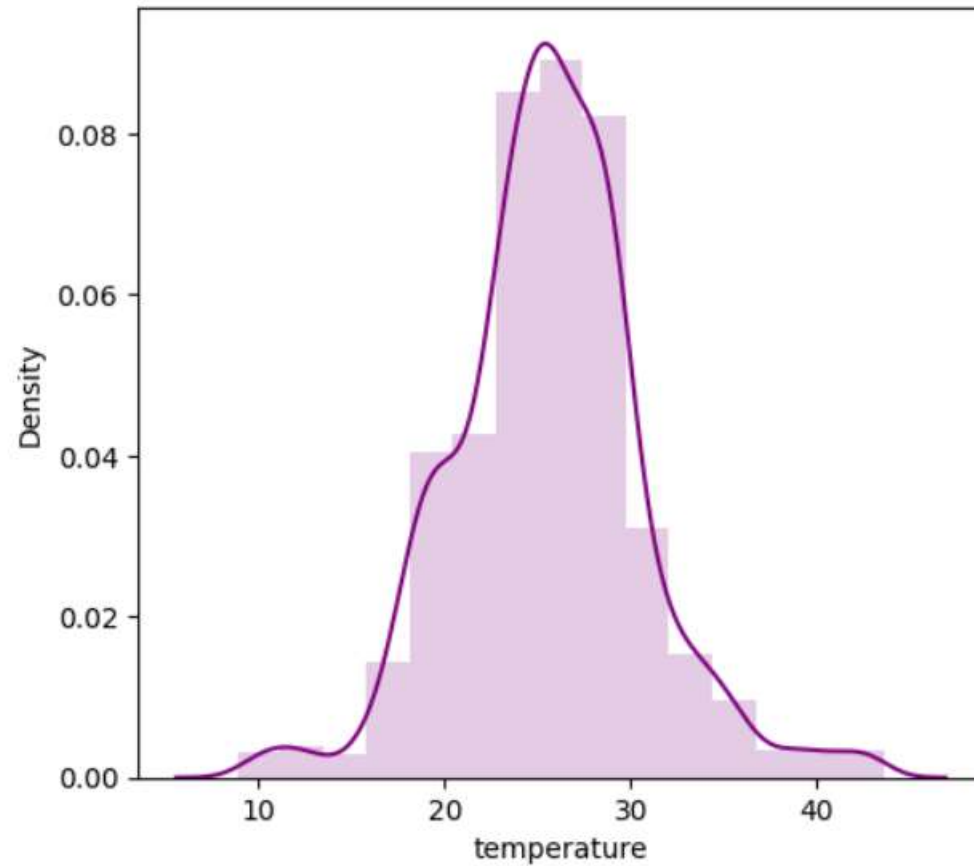
temperature - Temperature In Degree Celsius

humidity - Relative Humidity In %

ph - PH Value of The Soil

rainfall - Rainfall In mm

Compatibility of Temperature and PH Density

Coconut is a tropical crop and needs high humidity therefore explaining massive exports from coastal areas around the country.
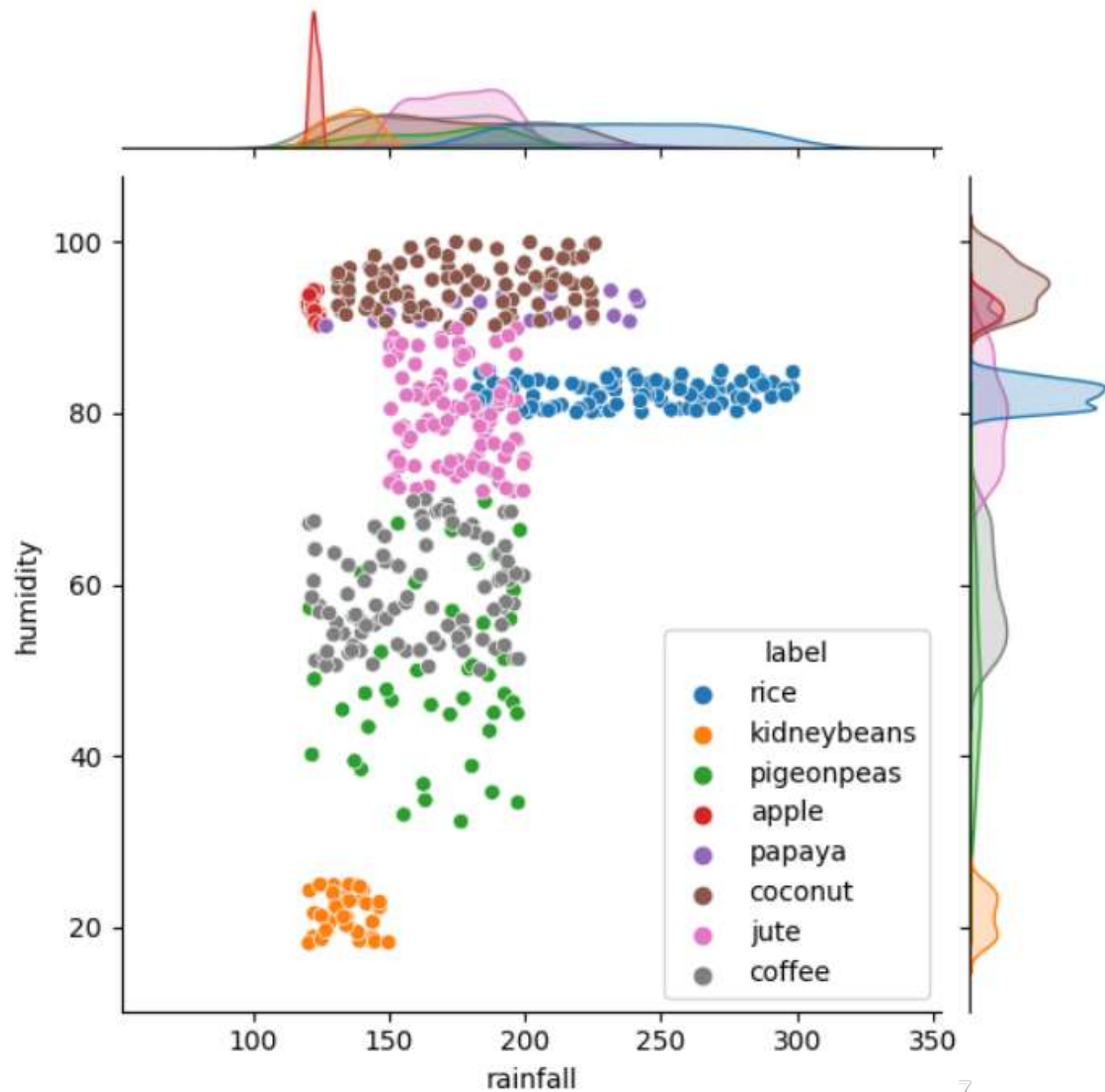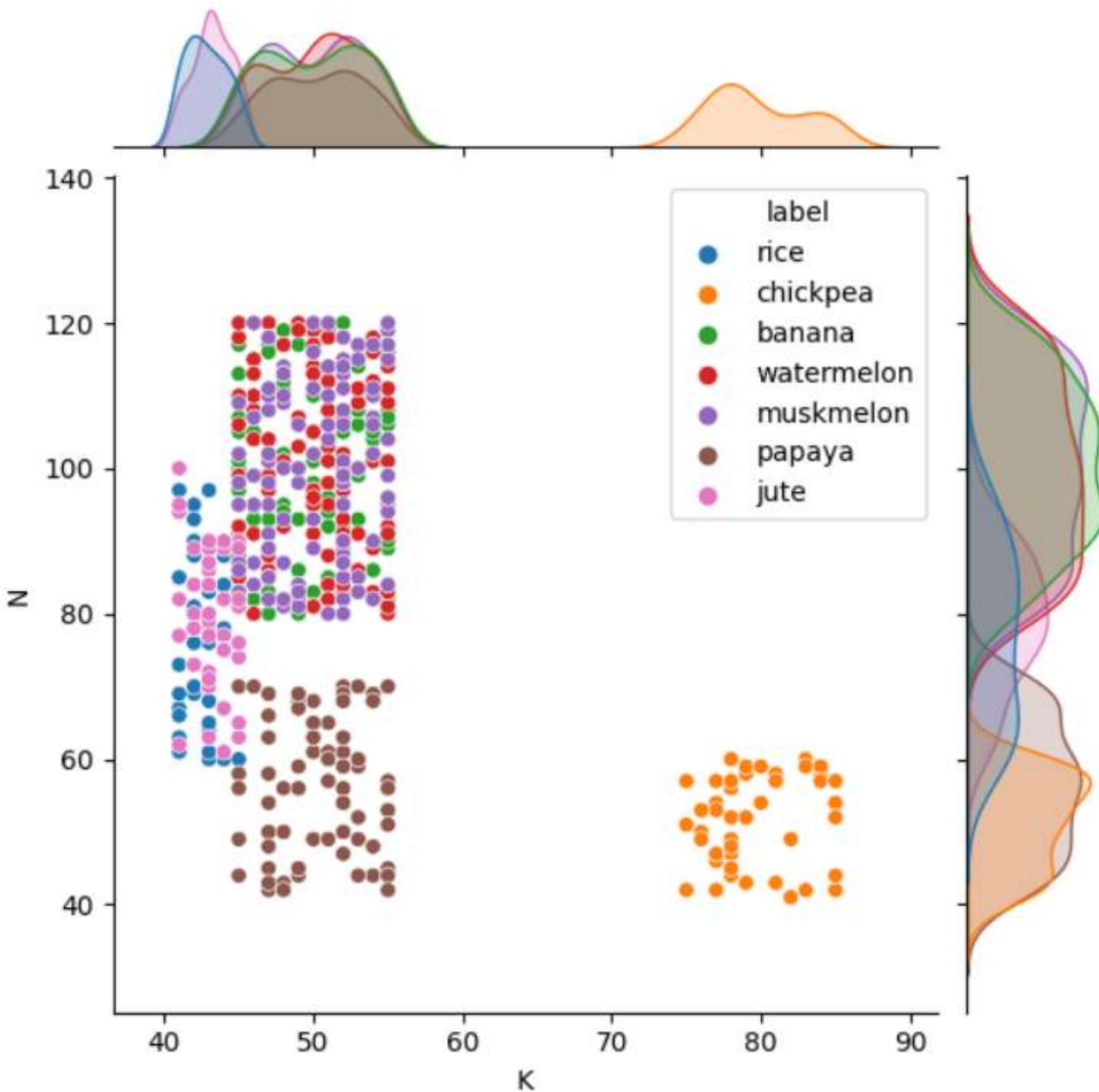
These soil ingredients directly affects nutrition value of the food.

POTASSIUM

Potassium is a very important element for the growth and development of plants. It regulates the water balance in plants and helps them to grow, develop, and become more resistant to diseases.

Rainfall and moisture dissolve the potassium in the soil and it becomes readily available to plants.
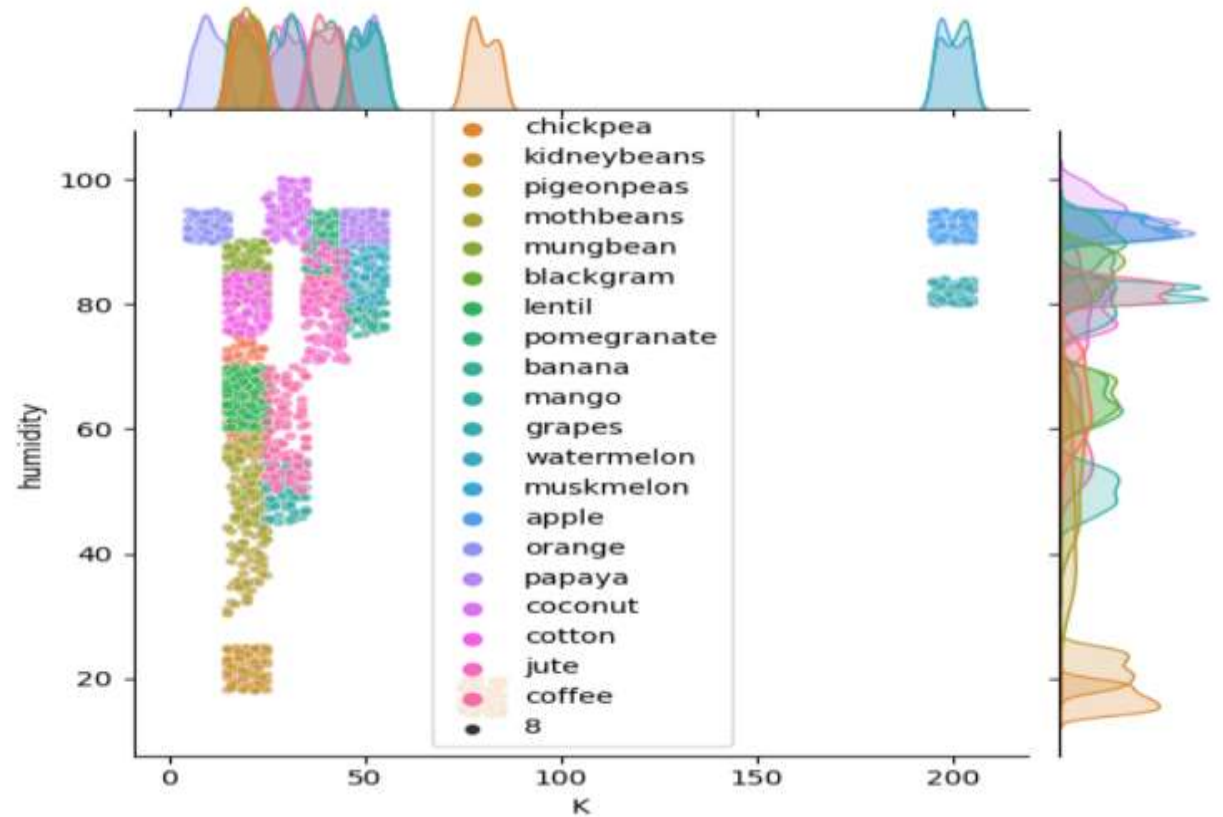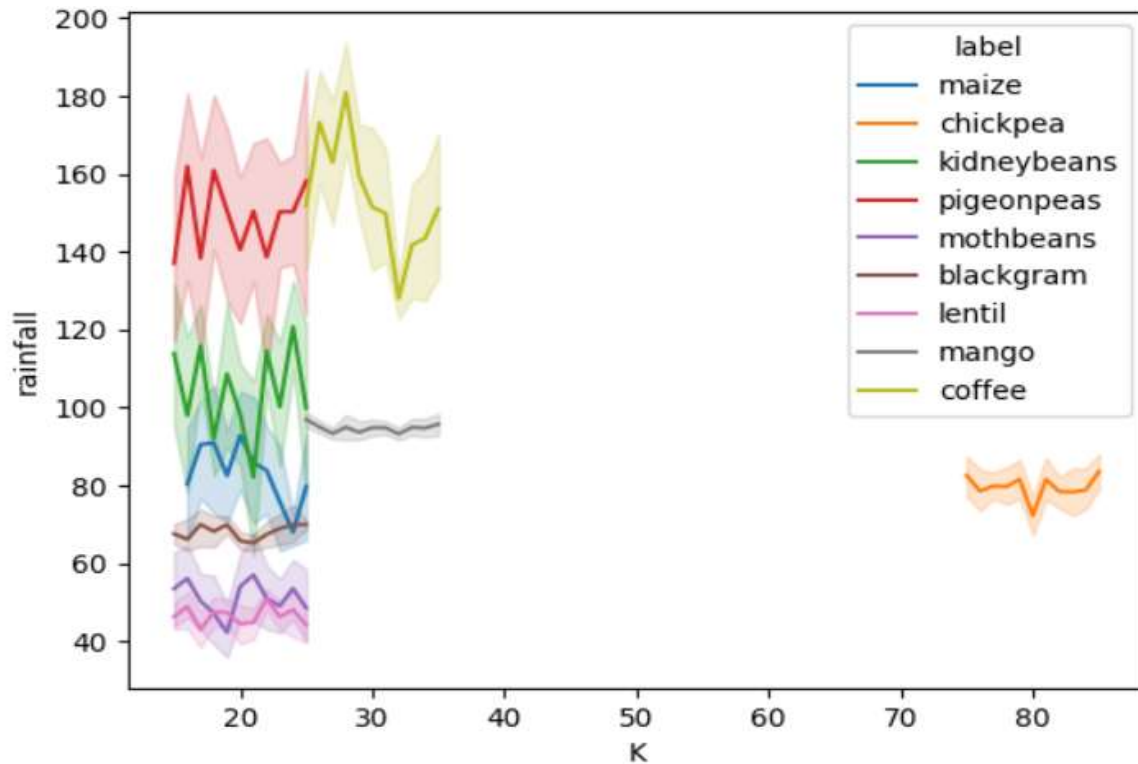
P plays an important role in the energy metabolism of plants.
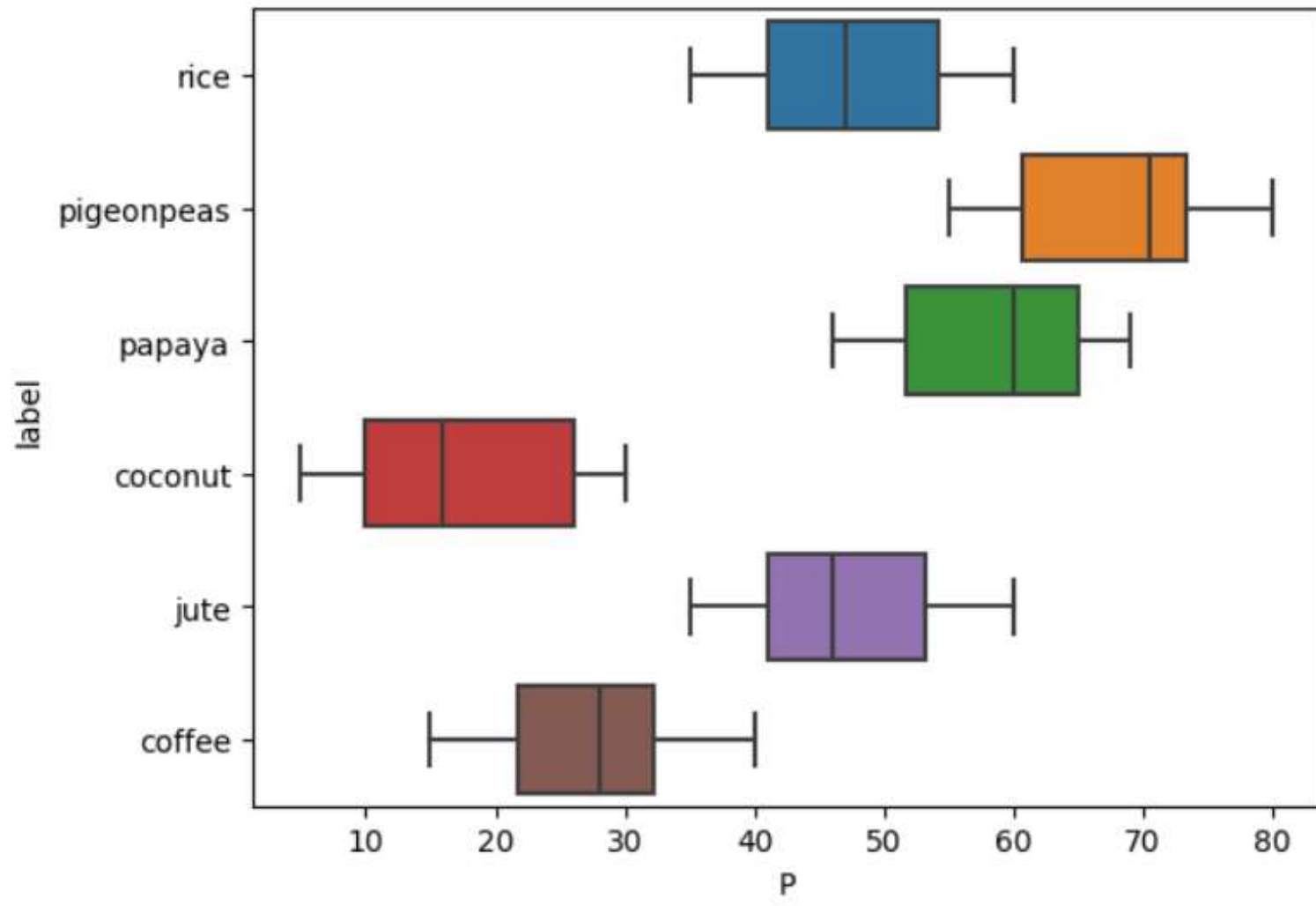
A graph showing how phosphorus levels change during periods of high rainfall.

PH

The solubility of nutrient elements taken up by plant roots varies at different pH ranges

PH	Reaction Class
< 7	Acidic
7	Neutral
7 <	Alkaline
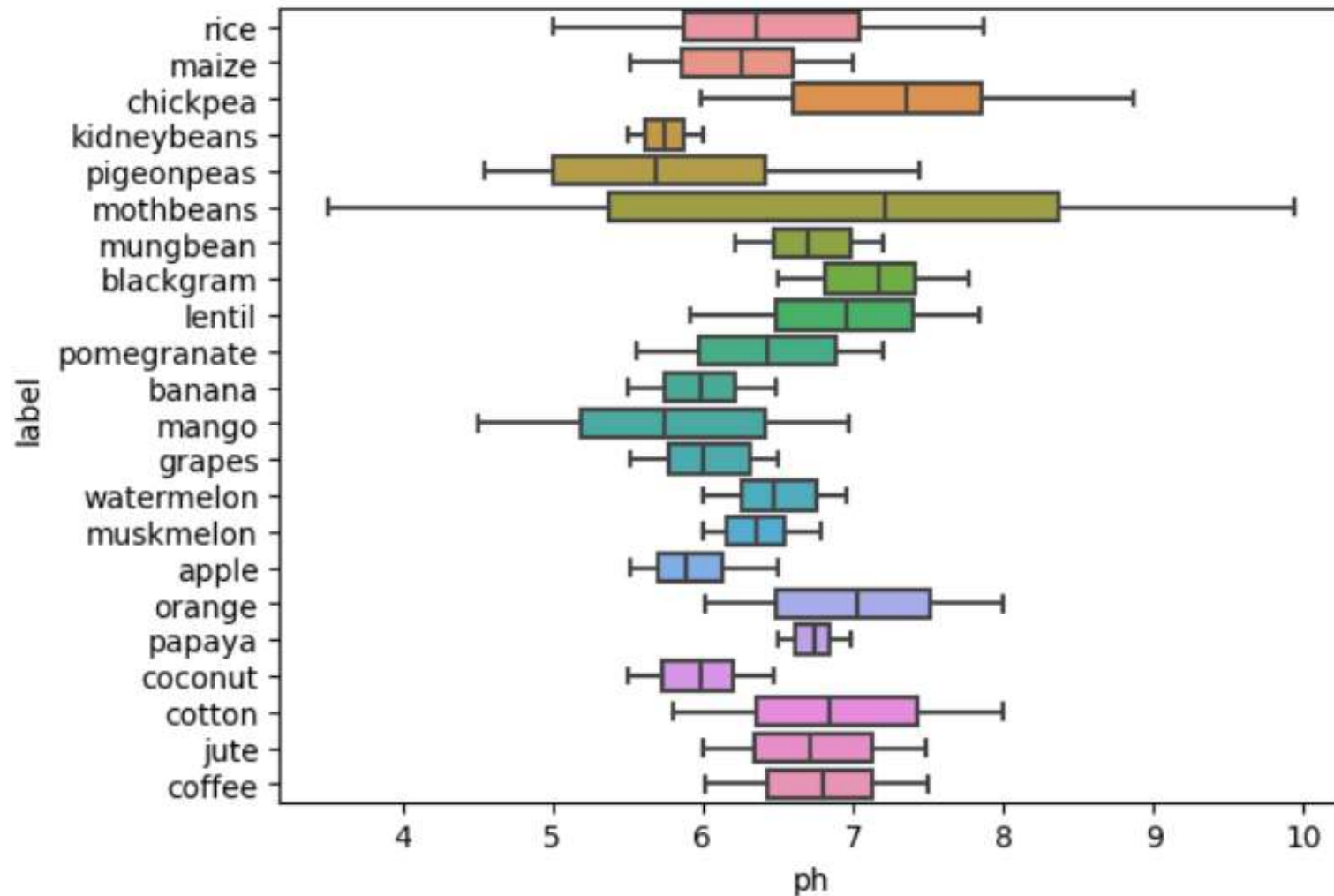
Mothbeans can grow in soil types with a wider pH range, while papaya and kidneybeans grow in more specific soil types.

Feature Importance

We see that the most important feature is rainfall, the least important is temperature
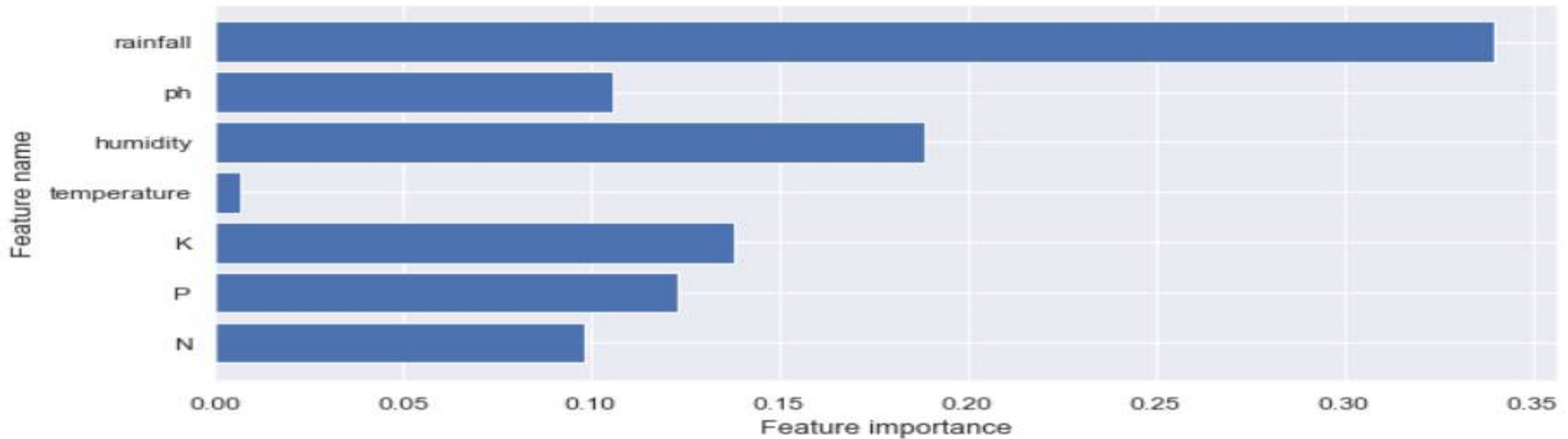
DATA PRE-PROCESSING & FEATURE SCALING

Make the data ready for machine learning model

```python
c=df.label.astype('category')
targets = dict(enumerate(c.cat.categories))
df['target']=c.cat.codes

y=df.target
X=df[['N','P','K','temperature','humidity','ph','rainfall']]
```
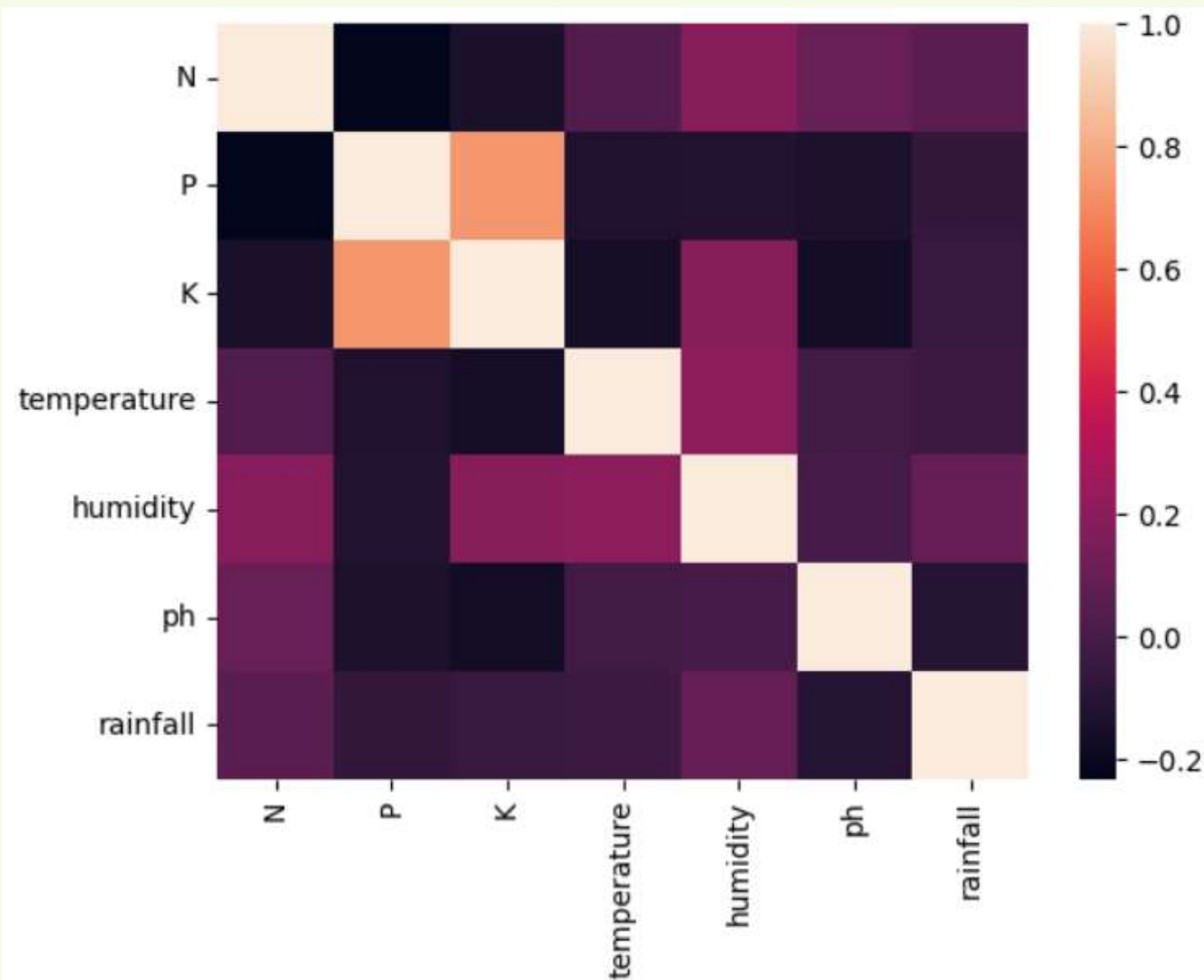
Positive correlation between P and K

Enables the design of systems optimized for projects to run as fast as possible.

```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler

X_train, X_test, y_train, y_test = train_test_split(X, y,random_state=1)

scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)

# we must apply the scaling to the test set as well that we are computing for the training
set
X_test_scaled = scaler.transform(X_test)
```

MODEL SELECTION

97.82%	**K-Nearest Neighbours**

98-98.5%	**Hyperparameter Tuning**

CLASSİFİCATİON USİNG SUPPORT VECTOR CLASSİFER (SVC)

97.45%

98.91%

98.91%

Linear Kernel
Accuracy

Rbf Kernel
Accuracy

Poly Kernel
Accuracy

The param_grid parameter is defined as a dictionary containing the possible values of hyperparameters. In this example, the parameter dictionary contains the values of the C and gamma hyperparameters.

```python
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV

parameters = {'C': np.logspace(-3, 2, 6).tolist(), 'gamma': np.logspace(-3, 2, 6).tolist()}
# 'degree': np.arange(0,5,1).tolist(), 'kernel':['linear','rbf','poly']

model = GridSearchCV(estimator = SVC(kernel="linear"), param_grid=parameters, n_jobs=-1,
cv=4)
model.fit(X_train, y_train)
```

Output:

```
GridSearchCV(cv=4, estimator=SVC(kernel='linear'), n_jobs=-1,
             param_grid={'C': [0.001, 0.01, 0.1, 1.0, 10.0, 100.0],
                         'gamma': [0.001, 0.01, 0.1, 1.0, 10.0, 100.0]})
```

Best Score and Parameters

98.66%
{'C': 1.0, 'gamma': 0.001}

ENSEMBLE LEARNING

98.73%

Classifying using Decision Tree

97.27% - TEST
97.15% - TRAIN

Classification using Random Forest
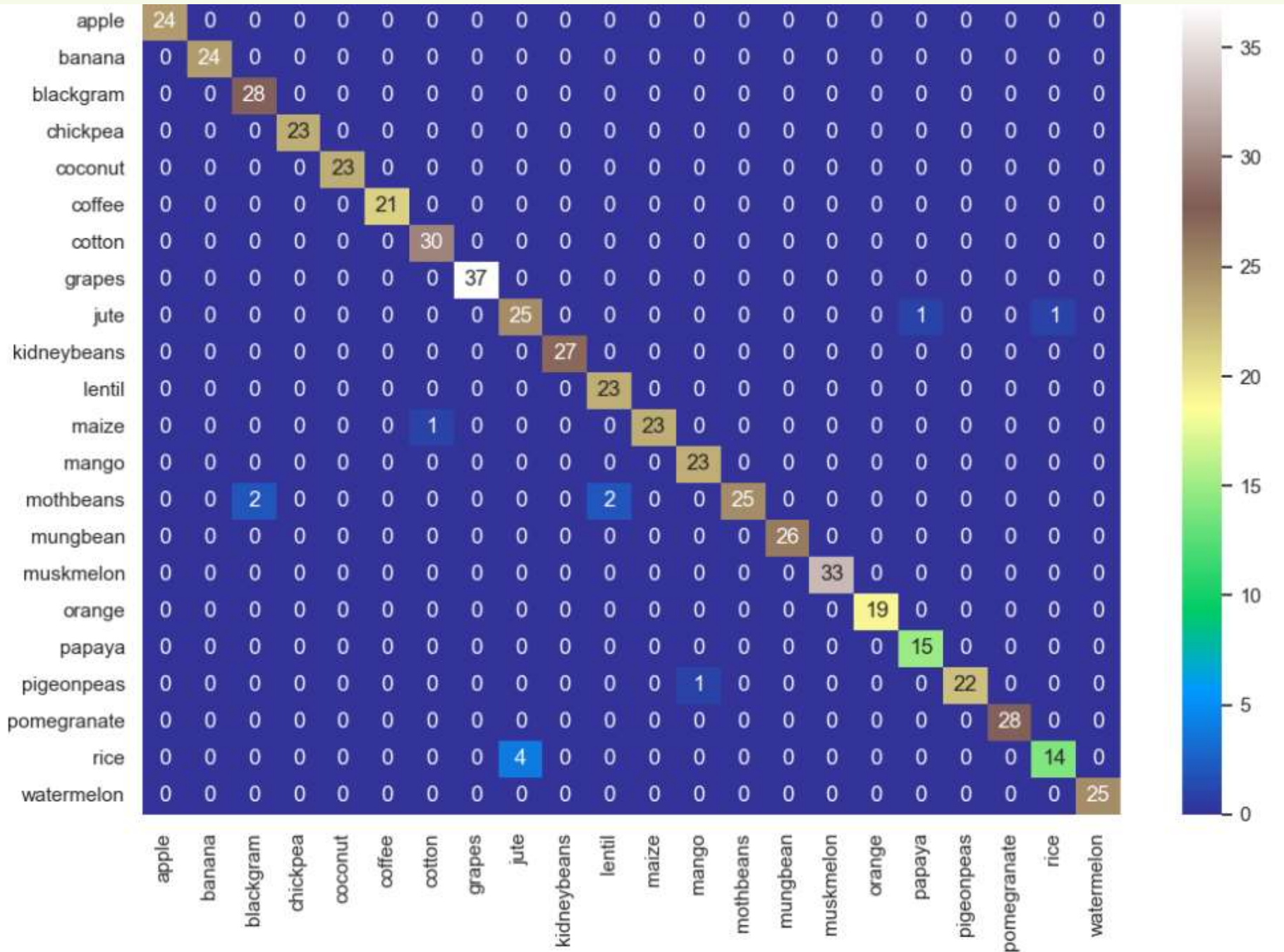
99.45%

Classification using Gradient Boosting

Confusion Matrix

It made a total of 4 incorrect predictions for mothbeans

Classification Report

It is used to measure the model's precision

RandomForestClassifier Classification Report

	precision	recall	f1	support
watermelon	1.000	1.000	1.000	25
rice	0.842	0.889	0.865	18
pomegranate	1.000	1.000	1.000	28
pigeonpeas	1.000	1.000	1.000	23
papaya	1.000	1.000	1.000	15
orange	1.000	1.000	1.000	19
muskmelon	1.000	1.000	1.000	33
mungbean	1.000	1.000	1.000	26
mothbeans	1.000	0.655	0.792	29
mango	1.000	1.000	1.000	23
maize	1.000	1.000	1.000	24
lentil	0.920	1.000	0.958	23
kidneybeans	1.000	1.000	1.000	27
jute	0.923	0.889	0.906	27
grapes	1.000	1.000	1.000	37
cotton	1.000	1.000	1.000	30
coffee	1.000	1.000	1.000	21
coconut	1.000	1.000	1.000	23
chickpea	1.000	1.000	1.000	23
blackgram	0.778	1.000	0.875	28
banana	1.000	1.000	1.000	24
apple	1.000	1.000	1.000	24

WHICH IS THE BEST MODEL ?

Score	Model
99.45%	Gradient Boosting
98.91%	RBF Kernel
98.91%	Poly Kernel
98.73%	Desicion Tree
97.82%	KNN
97.45%	Linear Kernel
97.27%	RF Test Set
97.15%	RF Training Set

LİCENSE OF THE PROJECT

The Apache 2 license is an open-source software license developed by the Apache Software Foundation. This license is a part of the free software movement and is a type of license that software developers can use for open-source software projects

The Apache 2 license provides project users with a free license to access, modify, redistribute, and commercially use the project's source code while protecting patents and copyrights.

MY OPINION?

The project data set may vary depending on the study area. In our dataset there are 3 chemical properties called macro elements and soil analysis can be done for micro elements. Soil salinity is also an important factor for plants and we can also assess microbial life. In addition, the cross-validation method could have been used and may have caused overfitting as the accuracy values were very close and high. The project includes well-designed visualizations, I think model selection and improvements are sufficient

THANKS

Edanur Sunay

edanursunay.1@gmail.com

https://github.com/edanursunay/holistic
_data_science_project_1