# Estimation of Major Agricultural Crop with Effective Yield Prediction using Data Mining

Ms. B. ANGALAPARAMESWARI, B.E/M.TECH(CSE),

Department of Computer Science and Engineering

Mr. K. BHARATH, B.E, Student of Computer Science Engineering

Mr. S. SASIKUMAR, B.E, Student of Computer science and Engineering

St. Joseph College of Engineering, Sriperumbudur, Chennai.

## Abstract

Indian agriculture recognized for different types of crop production, mainly due to change in resources. India is a country of the rural economy Crop yield projections and estimation are critical areas of research used to make sure food safety around the world and the Indian financial system.

India is a global agricultural force. Each year we are getting a high production of yield from preceding year. This more production in successive years totally depends on impact of various crop factors, efficiency of estimation and forecasting technique. Crop forecasting and estimation are very important for developing new government policies that actually give impetus to crop production.

The country is cultivated about 63% is treated with rain, while 37% is irrigated in crop production is totally depended on number of factors like geography, weather, biological and economical.

Massive number of different sources of raw agricultural data is present, but analysis these facts are very complicated for yield estimation of crop and implement in big data mining Techniques.

Whereas most efficient forecasting analysis tools used in agriculture data are KNN, k-means, and SVM. ANN and SVM are the most widely used methods to explore the useful data extensively in the area of agriculture.

**Key Terms :** Effective yield prediction, Agricultural crops, Data mining, Machine learning, Regression analysis, Multivariate analysis, Remote sensing, Crop management practices, Weather data, Soil quality, Precision agriculture, Food security, Sustainability.

# Introduction

Indian agriculture recognized for different types of crop production, mainly due to change in resources. India is a country of the rural economy Crop yield projections and estimation are critical areas of research used to make sure food safety around the world and the Indian financial system. India is a global agricultural force. Each year we are getting a high production of yield from preceding year. This more production in successive years totally depends on impact of various crop factors, efficiency of estimation and forecasting technique. Crop forecasting and estimation are very important for developing new government policies that actually give impetus to crop production. Crop yield forecasting is of main concern for market participants from farmers to commercial trading companies, such as large agricultural companies, and non-commercial trading companies, such as hedge funds. Early season production forecast is key to the price discovery mechanism for those billion-dollar crops. Yield forecast has a significant impact on the market's positions according to the anticipated supply of crops and the given demand. Farmers are faced with making difficult decisions on how to remain productive and sustainable with changing climates and market economic pressure. Data mining may provide crucial role in decision-making for complex agricultural problems by identifying the hidden patterns from large and complex data. It understands the general trends of the effect of various factors influencing crop yield. The ability to predict the future crop yield enables farmers and other stakeholders to make the most appropriate decisions for their crop. The provision of accurate and timely information such as meteorological, soil, use of fertilizers, and pesticides can help farmers make the best decision for their cropping situations. This could benefit them in attaining greater crop productivity if the conditions are suitable or help them decrease the loss due to unsuitable conditions for the crop yield. The agricultural sector is confronted with the major challenge of increasing production to feed a growing and increasingly prosperous population in a situation of decreasing availability of natural

resources. Factors of particular concern are water shortages, declining soil fertility, effects of climate change and rapid decrease of fertile agricultural lands due to urbanization. The poor, especially the rural poor, are particularly vulnerable to the negative effects of extreme weather and natural disasters. Yet accurate forecasting and timely warning can mitigate the effects of natural disasters such as floods, and improved weather forecasting can improve crop yields and lessen the effects of severe weather or drought. ICT has a crucial role to play in all links of the chain, from detection to modelling and forecasting to advance warning and localization. Yet the vast majority of the poor in developing countries still have very poor access to such information and very little advance warning of adverse events. Accurate forecasting and the increasingly sophisticated computer models obtain more timely and accurate weather and natural disaster information.

The United Nations (UN) is working towards its goal to reduce the hunger of the poor in the world by sustainable development in various sectors. Improving the supply of good quality food at a global level is one of the strategies adopted by the UN Council to handle its goal. But, the uncontrolled explosion of the population demands new approaches to overcome the problem. Prediction of human population and crop production is one way to address the situation. The accurate crop yield predictions during the growing season afford many advantages to the policymakers and farmers viz., forecasting the market prices, planning the import and export, and minimizing the socioeconomic impact of crop loss. Timely decisions on food production strengthen national food security. Agricultural entrepreneurs and small holders also reap the benefits of such predictions since they could make decisions that are well informed. The crop yield prediction is a challenging task for the policy leaders due to the complexity of the data. The researchers working in agriculture and ago economics are widely interested in developing new mathematical methods that give better prediction using the available parameters. The research on this direction is involved in providing a correlation between the area of agriculture and crop yield, taking into consideration diverse variables in the environment, the quality of the soil, irrigation, and how land is used. These models are built on rules with available parameters. The experts involved have perceptive knowledge about the type of relationships found with the variables involved in both agricultural and the environment.

# Literature Survey

The creation of 'Big Data technologies' improves the overall efficiency of agriculture. These technologies employ innovative information technology and systems (El Bilali and Allahyari (2018)). There are a number of issues for creative architecture and systems, algorithms, and analytics to handle, extract values, and extract hidden knowledge from it, due to the scale, variety, and sophistication of agricultural data sets. Big data analysis has a vital role in making strategic decisions for growers, policymakers, agri-business and agro-input firms, banks, insurance companies and service providers. The key opportunity and challenge are to set a new standard in agriculture because the factors that influence agriculture will differ with the environment, region, and types of soil, culture and tradition.

The data collection is critical for accurate prediction in agricultural management. The precise data allows to obtain relevant information about the crop cultivated, agronomic methods adopted by farmer. (Srivastava and Marshall-Colon (2018)). The multimodal nature of data presents several challenges, such as improving data collection methods, effective and efficient statistical and data analysis techniques for understanding and supporting the functions of different agriculture verticals. Since Big Data in agriculture is not so sensitive, there are fewer security or privacy concerns and data mining can be guided using practices in agriculture (Kamilaris et al. (2017)).

The issues are system management and data collection automation, such that there are almost no costs (Sonka (2015)). Since on-farm data will usually remain in the hands of individual businesses, and investments are needed to move and consolidate data in a central network, Poppe etal. (2015) recommend cooperation between the area's Agricultural Business Centers and Data Exchange Facilities. The problem in this field is when the collaboration is closed, whether proprietary systems will also be closed or whether these will become more open. Undoubtedly the most critical issues facing big data governance are how to guarantee privacy and protection. The ability to easily retrieve the right sources of data is integral to assessing performance indicators and core processes needed for effective growth strategies.

As observed by Antle et al. (2017), users do not want models but are interested in the information they produce. Decision-making tools ought to have these models embedded in them

to have value for farm managers. Automating data collection using sensors on machinery and other mobile devices and web-based sources such as weather and economic data such as prices could be one improvement. A further development area is the interoperability of instruments with software for accounting and the preparation of tax invoices. That way, data entered once can then be used across various analytical tools in an integrated way. The available method of performing this incorporation manually on a case-by-case basis renders this sort of analysis expensive even in a small geographic area and makes integration difficult across vast areas.

Real-time information sharing between farmers and researchers enables service providers to provide real-time and customized services based on geography, plant, management methods, level of automation, method of irrigation, farm size and soil composition. They inform growers about the multiple options of agricultural production and, where required, act immediately. Sawant et al. (2016) used the business models PRIDE and KRISHI. The Dindori tehsil farmers of Maharashtra state's Nashik district, India, are being taught by researchers to get more performance within a short period of time. The average productivity increased from 64 percent in 2013–2014 to 112 percent in 2014–2015 due to the custom crop protocol, agro-advisory, and timely warnings. In the second year an increase of around 90 percent in farmer participation was observed.

# System Design

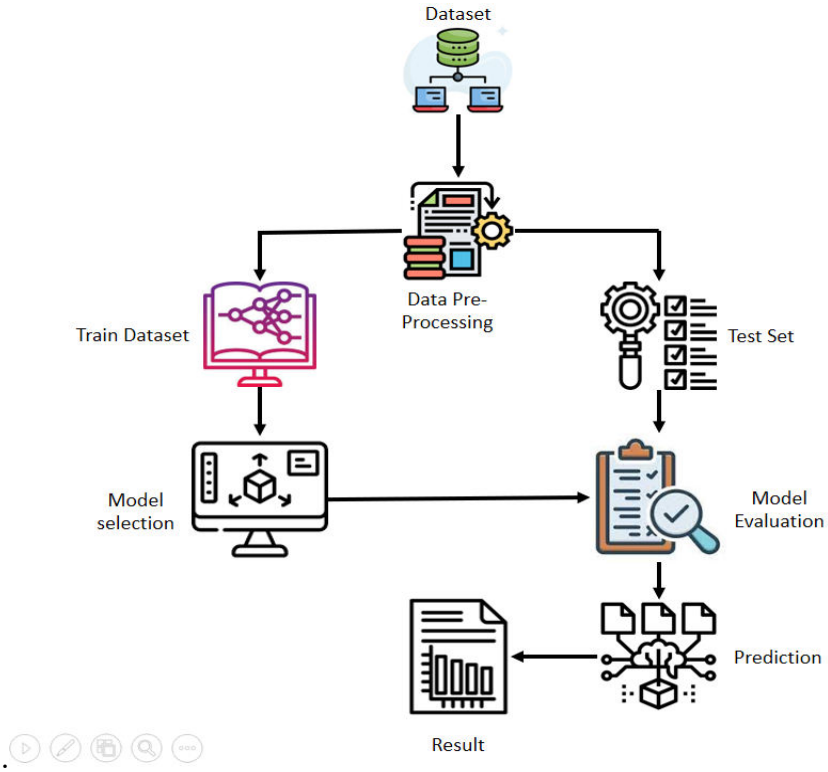
Data collection will collect data on weather patterns, soil quality, and crop management practices from various sources such as government agencies, research institutions, and individual farmers. The system will also gather real-time data on crop growth and development using remote sensing technologies such as satellite imagery and drones.

Data Processing involves the collected data will be preprocessed to remove noise and outliers and then transformed into a suitable format for analysis. The system will use statistical and computational techniques such as regression analysis and multivariate analysis to identify the factors that have the most significant impact on crop yield.

The machine learning algorithm will use to develop models that can predict crop yield based on the identified factors. The models will be trained on historical data and then tested on new data to evaluate their accuracy and effectiveness.

The user interface will have a user-friendly interface that allows farmers and policymakers to input their crop management practices and receive a prediction of crop

yield based on the models developed by the system. The interface will also provide real-time data on crop growth and development to help farmers optimize their yield prediction.




The system will be hosted on a cloud infrastructure to provide scalability, flexibility, and security. The cloud infrastructure will also allow for easy integration with other agricultural systems and technologies.

The system will have a feedback mechanism that allows farmers to provide feedback on the accuracy and effectiveness of the predictions. The feedback will be used to improve the models and make the predictions more accurate over time.

The system can include a visualization component that allows farmers and policymakers to view the prediction results in an intuitive and easy-to-understand format. This can include graphs, charts, and maps that show the predicted crop yield for different regions and crops. The visualization component can also help farmers identify areas that require additional attention and resources to optimize crop production.

The result of the machine learning model is the prediction can help farmers optimize their crop management practices, reduce waste, and increase profitability. For example, if the system predicts a lower-than-expected crop yield due to poor soil quality, farmers can adjust their fertilization and irrigation practices to improve soil quality and increase yield. This can result in improved crop quality and quantity, leading to higher profits for farmers and increased food security for the population.

# IMPLEMENTATION

**Importing Library Files:**

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from plotly.offline import init_notebook_mode, iplot

import plotly.graph_objs as go

init_notebook_mode(connected=True)

from plotly.graph_objs import *

%matplotlib inline
```

**Loading Dataset and Checking missing values:**

```
data=pd.read_csv(r'File
        name.csv',
        encoding='latin-1')

pd.set_option("display.max_col
        umn",1000)

data.shape

data.head()

len(data['Area'].unique())

a= data.Y1961

a

b= data.Y2013

b
```

```
data.Element.unique()


Filling missing values:

len(data['ElementCode'].uniq
        ue())

len(data.Item.unique())

len(data['Item
        Code'].unique())

data['Sum Years'] = 0

for year in range(1962, 2014):

col = 'Y' + str(year)

   data['Sum Years'] =
        data['Sum Years'] +
        data[col]

el_size =
        data.groupby('Element
        ').agg('size')

el_size.values

el_size.index

sns.barplot(x=el_size.index,
        y=el_size.values,alpha=
        0.9)

plt.show()
```

**Outlier Removal Checking:**

```
for item, group in data.groupby(['Item', 'Area']):


item_area.append((item[0], item[1],
        group.Element.values.tolist()))

only_food = set()

only_feed = set()

food_and_feed = set()

list(map(lambda x: only_feed.add(x[0]),
        list(filter(lambda x: 'Food' not in x[2],
        item_area))));

list(map(lambda x: only_food.add(x[0]),
        list(filter(lambda x: 'Feed' not in x[2],
        item_area))));

list(map(lambda x: food_and_feed.add(x[0]),
        list(filter(lambda x: 'Feed' in x[2] and 'Food' in
        x[2], item_area))));

only_food.intersection(food_and_feed)

only_feed.intersection(food_and_feed)

only_feed.difference(food_and_feed)

only_food.difference(food_and_feed)
```

**Normalization of Data:**

```
for item, group in data_item_grouped:

 #    print(group[group['Sum Years'] ==
max_sum_items[item]]['Area'].values[0])
```

```
#     print(max_sum_items[item])

max_sum_items_area[item] = group[group['Sum
Years'] ==
max_sum_items[item]]['Area'].values[0]

max_sum_items = max_sum_items.to_dict()

max_sum_items_sorted =
sorted(max_sum_items.items(), key=lambda x:
x[1], reverse=True)

titles_areas = []

k, v in max_sum_items_sorted:

titles_areas.append(max_sum_items_area[k])

items = list(map(lambda x: x[0],
max_sum_items_sorted))

values = list(map(lambda x: x[1],
max_sum_items_sorted))

titles_areas_items = list(map(lambda x: "(" + x[0]
+ ")  ,  " + x[1], list(zip(titles_areas, items))))

fig, ax1 = plt.subplots()

sns.barplot(x=values[:20], y=items[:20], alpha=0.8
)

ax1.tick_params(labeltop=False, labelright=True)

ax_2 = ax1.twinx()

ax_2.set_yticks(list(range(20)))
```

```
ax_2.set_yticklabels(titles_areas[:20][::-1])

plt.show()
```

**Training and Testing the Data:**

```
iplot([go.Choropleth(locationmode='country names',locations=area_1961.index,
text=area_1961.index,z=area_1961.values)],filename='1961')
iplot([go.Choropleth( locationmode='country names', locations=area_2013.index,
text=area_2013.index,z=area_2013.values)],filename='2013')
data
data.columns
def label_encoding(categories):categories = list(set(list(categories.values)))mapping = {}
for idx in range(len(categories)):
mapping[categories[idx]] = idx
return mapping
data['Area Abbreviation'] = data['Area Abbreviation'].map(label_encoding(data['Area
Abbreviation']))
data.head(10)
```

**SMOTE Technique**

```
data['Item'] = data['Item'].map(label_encoding(data['Item']))

data['Area'] = data['Area'].map(label_encoding(data['Area']))

data.head(10)

data['Element'] = data['Element'].map(label_encoding(data['Element']))

data.head(10)

X = data[['Area Abbreviation', 'Area Code', 'Area', 'Item Code', 'Element Code']].values

y = data[['Element']].values
```

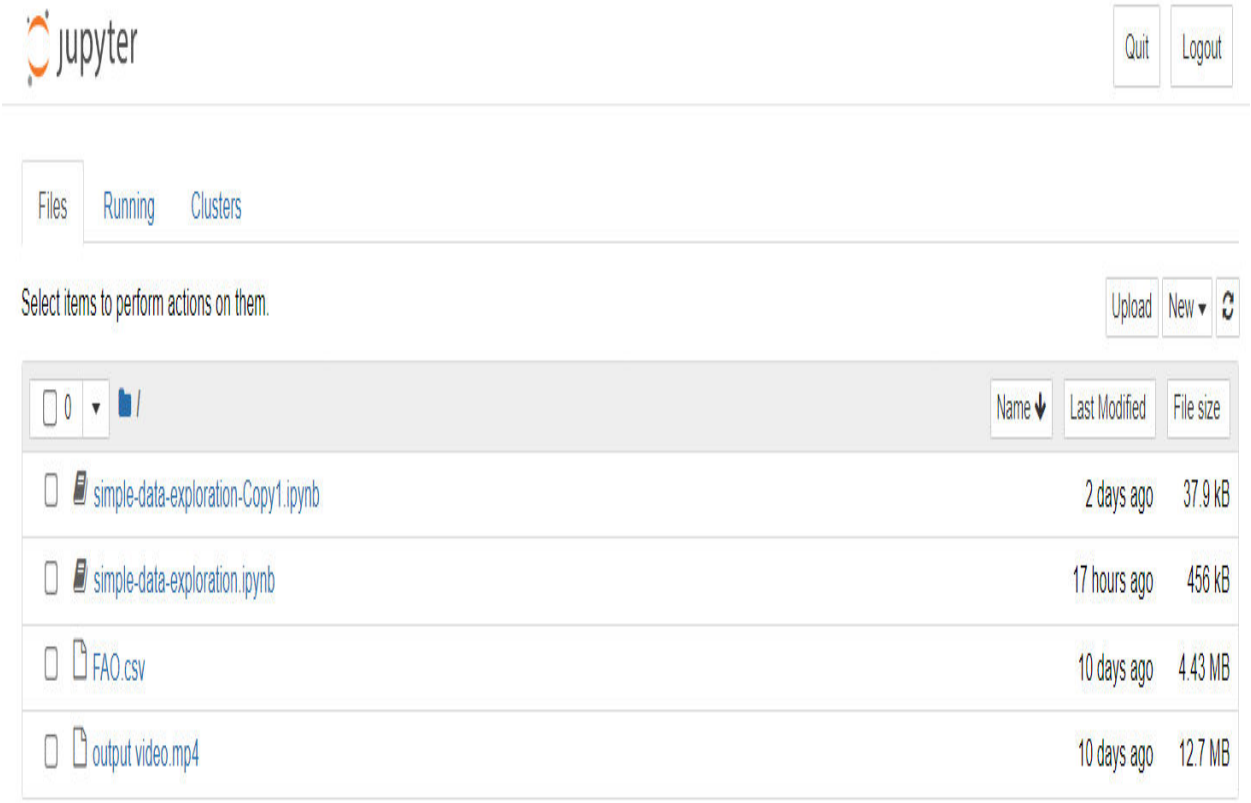
**Algorithm selection**

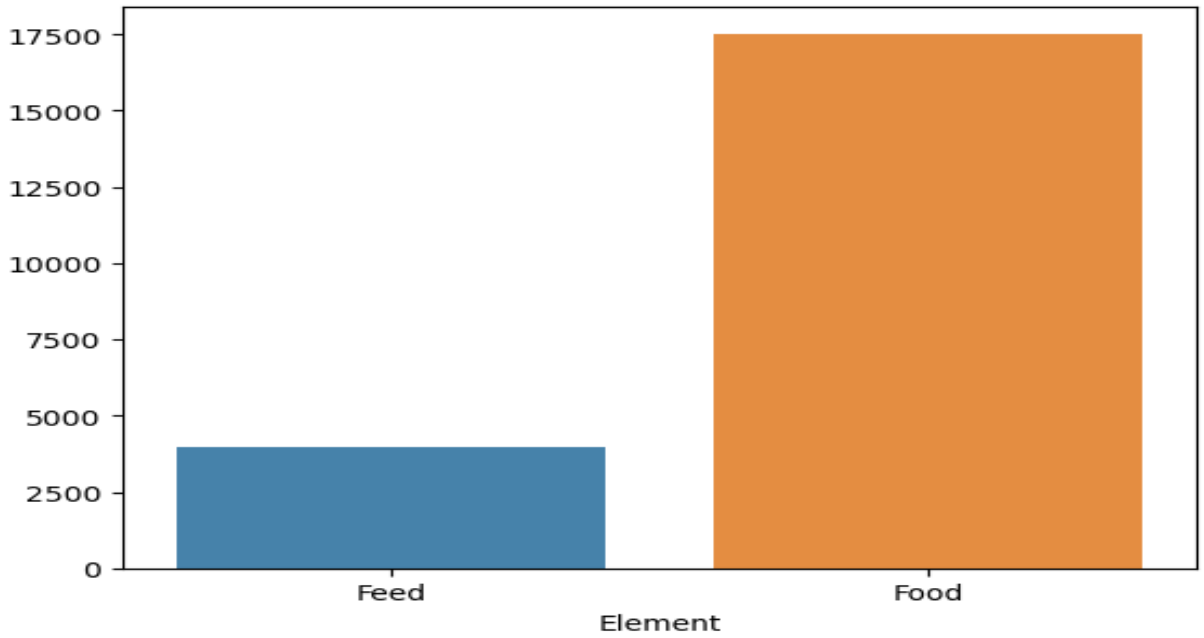
```
from sklearn.naive_bayes import GaussianNB

def get_accuracy(y_true, y_preds):

   true_negative, false_positive, false_negative, true_positive = confusion_matrix(y_true,
y_preds).ravel()
```

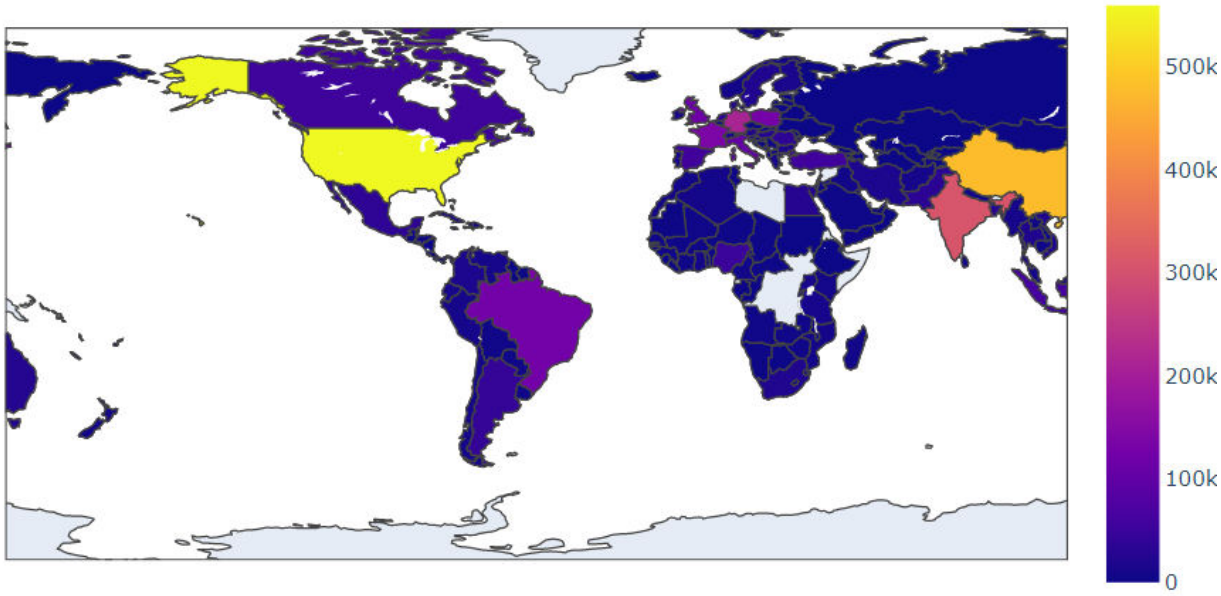
```
    accuracy = (true_positive + true_negative)/(true_negative + false_positive +
false_negative + true_positive)
    return accuracy
naive_b = GaussianNB()
naive_b.fit(X_train, y_train)
from sklearn.metrics import confusion_matrix
models = [naive_b]
acc = []
for model in models:
    preds_val = model.predict(X_val)
    accuracy = get_accuracy(y_val, preds_val)
    acc.append(accuracy)
model_name = ['Naive Bayes Accuracy']
accuracy = dict(zip(model_name, acc))
print(accuracy)
predicted = naive_b.predict(X_val)
cn_matrix = confusion_matrix(y_val, predicted)
print(cn_matrix)
import seaborn as sns
sns.heatmap(cn_matrix, annot=True)
from sklearn.metrics import classification_report
predicted = model.predict(X_val)
report = classification_report(y_val, predicted)
print(report)
from sklearn import datasets, svm
```

```
from sklearn.metrics import accuracy_score

x, y = X_train, y_train

clf_predict = svm.SVC(C=7120.0, cache_size=200, class_weight=None, coef0=0.0,

  decision_function_shape='ovr', degree=3, gamma=6.191, kernel='rbf',

  max_iter=-1, probability=False, random_state=None, shrinking=True,

  tol=0.001, verbose=False)

clf_predict.fit(x, y)

print("\n",clf_predict.predict(X_test[0:]))

print("\nAccuracy SVM : "+ str(round(accuracy_score(clf_predict.predict(X_test[0:]),
y_test[0:])*100, 1)))

predicted = clf_predict.predict(X_test[0:])

matrix = confusion_matrix(y_test[0:], predicted)

print(matrix)

import seaborn as sns

sns.heatmap(matrix, annot=True)

from sklearn.metrics import classification_report

predicted = model.predict(X_test[0:])

report = classification_report(y_test[0:], predicted)

print(report)
```

## SNAPSHOTS


jupyter
Quit
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
0
/
Name
Last Modified
File size
simple-data-exploration-Copy1.ipynb
2 days ago
37.9 kB
simple-data-exploration.ipynb
17 hours ago
456 kB
FAO.csv
10 days ago
4.43 MB
output video.mp4
10 days ago
12.7 MB


17500
15000
12500
10000
7500
5000
2500
0
Feed
Food
Element


500k
400k
300k
200k
100k
0

## CONCLUSION

The work demonstrated the potential use of data mining techniques in predicting the crop yield based on the input parameters average rainfall and area of field. The developed webpage is user friendly and the accuracy of predictions. The districts selected in the study indicating higher accuracy of prediction. The user-friendly web page developed for predicting crop yield can be used by any user by providing average rainfall and area of that place.

## FUTURE ENHANCEMENTS

The future work aimed at the analysis of the entire set of data and will be devoted to suitable strategies for improving the efficiency of the proposed algorithm. Use of such kind of approach to forecasting is not restricted to agriculture alone. The clustering and regression is one of the capable tool in field of data mining which can be used in several different ways.

## REFERENCES:

[1] Rajshekhar Borate., "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in India", International Journal of Novel Research in Computer Science and Software Engineering, April 2016.

[2] D Ramesh, B Vishnu Vardhan, "Analysis of Crop Yield Prediction using Data Mining Techniques", International Journal of Research in Engineering and Technology (IJRET),Vol.4, 2015.

[3] Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004.

[4] Djodiltachoumy, S. "Analysis of Data Mining Techniques for Agriculture Data." International Communications 4.2 (2016): 1311- 1313.WE

[5] Mucherino, Antonio, PetraqPapajorgji, and Panos M. Pardalos. "A survey of data mining techniques applied to agriculture." Operational Research 9.2 (2009): 121-140.

[6] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ...& Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[7] Hand, D. J. (2006). Data Mining. Encyclopedia of Environ metrics, 2.

[8] Ramesh, D., &Vardhan, B. V. (2015). Analysis of crop yield prediction using data mining techniques. International Journal of Research in Engineering and Technology, 4(1), 47-473.

[9] Ahamed, AT M. Shakil, et al. "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh." Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on. IEEE, 2015.

[10] Sangeetha, A., and M. Ravichandran. "A Survey on Data Mining Approaches to Handle Agricultural Data." Data Mining and Knowledge Engineering 8.9 (2016): 286-290.

[11] López, C. J. Á., Valiño, J. A. R., & Pérez, M. M. (2008). Typology, classification and characterization of farms for agricultural production planning. Spanish Journal of Agricultural Research, (1), 125-136.

[12] Sastri, A. S. R. A. S., Rao, B. R., Krishna, Y. R., & Rao, G. G. S. N. (1982). Agricultural droughts and crop production in the Indian arid zone. Archives for meteorology, geophysics, and bioclimatology, Series B, 31(4), 405-411.

[13] Allen, R. C. (1999). Tracking the agricultural revolution in England. Economic history review, 209-235.

[14] Wackernagel, M., Schulz, N. B., Deumling, D., Linares, A. C., Jenkins, M., Kapos, V., ...& Randers, J. (2002). Tracking the ecological overshoot of the human economy. Proceedings of the national Academy of Sciences, 99(14), 9266-9271.

[15] Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic acids research, 38(suppl_2), W64-W70.

## AUTHOR 1

Mrs.B.Angalaparameswari M.E is a Assistant Professor in department of computer science and engineering at St.Joseph college of engineering,Sriperumbudur,Chennai. She has done his M.E CSE in MGR University in the year 2015.

Mrs.B.Angalaparameswari has 1 year of teaching experience in this college.

## AUTHOR 2

Mr. S. Sasikumar B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops and Seminars in Python, Machine Learning.I got a certificate from ICT academy in Data Analyst course.

## AUTHOR 3

Mr. K. Bharath B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops and Seminars in the area of Python and Machine Learning. .I got a certificate from ICT academy in Data Analyst course.