Analysis of Total Geographical Land Use and Prediction of Crops using Various ML Models

Submitted in partial fulfilment for the course

Data Visualisation (CSE3020)

by

20BCE1317 (Jyothssena GS)

20BCE1360 (Prathiba Lakshmi Narayan)

20BCE1043 (Vishal N)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2023

Bonafide Certificate

We hereby declare that the project entitled "**Analysis and Prediction of Total Geographical Land Use**" submitted by us in partial fulfilment of the requirements of the course "**Data Visualisation (CSE3020)**", is a record of bonafide work carried out by the both of us, under the course faculty of **Prof. Pattabiraman V**. We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for any other course of this institute or of any other institute or university.

Vishal N
20BCE1043
SCOPE
VIT Chennai

Prathiba Lakshmi Narayan
20BCE1360
SCOPE
VIT Chennai

Jyothssena GS
20BCE1317
SCOPE
VIT Chennai

Acknowledgement

We wish to express our sincere thanks and deep sense of gratitude to our professor, **Dr. Pattabiraman V,** SCOPE, VIT Chennai, and his associates, for their consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

No words can adequately express our sincerest thanks to all those who have helped us in making this project a success.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

Table of Contents

Abstract

Objective

Agriculture is one of the primary sources of income for millions of people in India. With over 60% of the population engaged in agriculture, it is a crucial sector that contributes significantly to the nation's economy. However, crop yield and quality depend on various factors such as soil fertility, weather patterns, and nutrient availability. With the help of modern technologies and data analysis techniques, farmers and policymakers can optimise crop production by identifying the most suitable crops for a particular region and understanding the impact of various environmental factors.

Methods Used

We use methods like Deep Learning, XGBoost, Multiple Regression, Ensemble Learning, KNN, SVC, Random Forest and Gradient Boosting to model and predict the crop yield based on the aforementioned factors.

By identifying the most suitable crops for a particular region and understanding the impact of environmental factors, farmers and policymakers can make informed decisions that can lead to improved crop yields, food security, and sustainable agriculture.

Outcome

The project can help in identifying crop patterns, predicting crop yields, evaluating model performance, understanding factors influencing crop growth, creating interactive visualisations, and providing decision support for farmers, policymakers, and other stakeholders involved in crop management and planning. These outcomes can lead to improved crop yields, optimised resource utilisation, and more sustainable agricultural practices.

Scope

This project involves various tasks such as data preparation, model building, model evaluation, model interpretation, deployment, and documentation. Data preparation involves collecting, cleaning, exploring, and engineering the data. Model building includes implementing and training machine learning algorithms using Python libraries, and hyperparameter tuning. Model evaluation involves assessing model performance using evaluation metrics. Model interpretation includes interpreting model results and visualising them for insights. Deployment involves integrating models into a production environment or making them accessible to end-users. Finally, a detailed documentation of the entire process is made.

Introduction

In this data visualisation project, we aim to analyse the crops grown in different states of India by considering multiple factors such as NPK values, rainfall, pH, and humidity. The dataset has been preprocessed and cleaned, ensuring that it is suitable for further analysis. The project employs various machine learning algorithms such as Linear Regression, Ensemble Learning, KNN, SVC, and Random Forest to model and predict the crop yield based on the aforementioned factors.

The use of machine learning algorithms enables us to identify patterns and relationships between different variables and the crop yield. By using these algorithms, we can develop predictive models that can help us determine the most suitable crops for a particular region. This can help farmers make informed decisions about crop cultivation, including the selection of seeds, fertilisers, and other inputs.

Finally, the visualisation of results using Tableau and Python enables us to communicate the insights gained from the analysis in a visually appealing and easy-to-understand manner. By providing clear visualisations, farmers and policymakers can easily interpret and act upon the findings, leading to better crop yields and improved agricultural practices.

Furthermore, policymakers can use the insights gained from this analysis to optimise crop production and promote sustainable agriculture. With the ability to forecast crop yields accurately, policymakers can plan and allocate resources efficiently, reducing food waste and promoting food security. Additionally, this analysis can help identify areas where crops are not performing optimally and recommend interventions to address the issue.

In conclusion, this data visualisation project can have a significant impact on the agriculture sector in India. By identifying the most suitable crops for a particular region and understanding the impact of environmental factors, farmers and policymakers can make informed decisions that can lead to improved crop yields, food security, and sustainable agriculture.

Literature Review

1. Crop Prediction using Machine Learning | IEEE Conference Publication [1]

   The paper "Crop Prediction using Machine Learning" by M. Kalimuthu, P. Vaishnavi, and M. Kishore, published in the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), presents a machine learning-based approach to predict crop yield using weather and soil data.

   The study involves collecting seed data of crops and appropriate parameters like temperature, humidity, and moisture content to facilitate successful crop growth. Additionally, the authors are developing a mobile application for Android that enables users to enter parameters like temperature and their location to initiate the prediction process.

   The authors use Naive Bayes, a supervised learning algorithm that is proposed by the authors to guide beginner farmers in sowing reasonable crops.

   Using Naive Bayes for crop prediction is advantageous as it is easy to implement and requires less computational resources compared to other complex algorithms. It is suitable for multi-class classification problems and can handle a large number of features. In the case of crop-prediction, India has the capability to grow many different crops and thus naive bayes can be beneficial

   While there are advantages, Naive Bayes works under the assumption that the features are independent. Feature independence is not always true in real-world problems, thus leading to lower accuracy when features are highly correlated. Another concern is that it may be sensitive to irrelevant features that do not contribute to classification but increase the complexity of the model.

2. Patterns and ecological implications of agricultural land-use changes: a case study from central Himalaya, India - ScienceDirect [2]

   The study by Semwal et al. (2004) focuses on agricultural land-use changes in the central Himalayan region of India and their ecological implications. The authors use a participatory survey approach to collect data on land-use patterns and changes, in addition to the information available from existing land-use maps and satellite imagery.

   The study finds that participatory survey data complement existing land-use maps and satellite imagery, providing a more comprehensive understanding of land-use patterns and changes. However, the authors note that reconstruction of the past based on farmers' perceptions is limited to variables that are traditionally quantified and to a time scale that is within the range of human memory (30 years in this study).

   The study highlights the importance of involving local communities in land-use monitoring and management, as they possess valuable knowledge and perceptions that are often overlooked by traditional mapping methods. The authors argue that participatory approaches can provide a more accurate and complete picture of land-use changes and their ecological implications, which can aid in designing effective policies and interventions.

   Overall, the study by Semwal et al. (2004) contributes to the growing body of literature on participatory approaches to land-use monitoring and management, and emphasises the importance of incorporating local knowledge and perceptions in land-use planning and decision-making.

3. Crop Yield Prediction based on Indian Agriculture using Machine Learning | IEEE Conference Publication [3]

   The paper by Nishant et al. (2020) aims to predict crop yield in India using machine learning techniques. The authors emphasise the importance of agriculture in India and how crop yield prediction can help farmers plan their crop production, manage resources, and improve their economic conditions.

   The study uses simple parameters such as state, district, season, and area to predict the yield of various crops in India. The authors employ advanced regression techniques such as Kernel Ridge, Lasso, and ENet algorithms to predict the crop yield. Additionally, the study uses the concept of stacking regression to enhance the accuracy of the algorithms.

   The study is significant as it applies machine learning techniques to the context of Indian agriculture, which can help farmers make informed decisions and improve crop productivity. The use of advanced regression techniques and stacking regression can help improve the accuracy of crop yield prediction, which is crucial for crop management and resource planning.

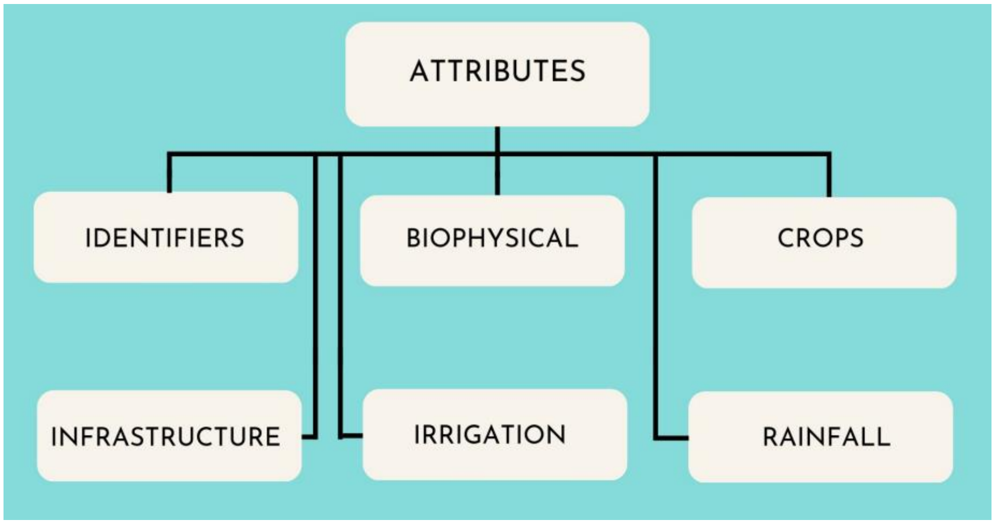
   Overall, the study by Nishant et al. (2020) provides valuable insights into the application of machine learning techniques for crop yield prediction in the context of Indian agriculture. The findings of the study can help farmers, policymakers, and researchers to make informed decisions regarding crop management and resource planning.

Materials and Methods

Datasets

The sources for our datasets include a dataset built on data collected by ICRISAT, in collaboration with TATA-Cornell Institute, from http://data.icrisat.org/dld/src/crops.html. We are also using a soil nutrients dataset from Kaggle, taken from https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset.

1. Dataset 1



The identifiers are id, state and district. Biophysical attributes are total area, forest area etc. Infrastructure attributes are the number of banks and post offices. Crop attributes have area, yield and production of various crops. In addition to that, rainfall levels were recorded for each month.

2. Dataset 2

The second dataset has crops with corresponding NPK- Nitrogen, Phosphorous, Potassium, humidity values, temperature, pH and rainfall. The various crops were rice, coconut, jute etc.

Preprocessing

Before we perform any machine learning algorithm, we need to preprocess and explore the data to get a better understanding of the same.

The unwanted columns were first removed. The data also had inconsistencies and outliers which can skew the results. This was seen in the crop yield column where there were a lot of 0 values. To eliminate this, we took the average yield of that particular crop in that particular district for the previous and next year and replaced the 0 values with this.

We then had to split our datasets into training and testing datasets. To do this, we first take the feature variables and the target variables separately. Here, the target variable is the crop grown while the feature variables are Nitrogen, Phosphorus and Potassium percentages, temperature, humidity, pH and rainfall.

We then need to perform feature scaling to ensure that all the features variables in the dataset are on a similar scale range. It ensures avoidance of bias in the data model and hence improves performance. It also helps us interpret and compare the data better. We used the MinMaxscaler from the sklearn library to transform the dataset.

Models and Algorithms Used

The different machine learning models that were used include - Multiple regression, SVC, Decision tree, Random forest, Gradient boosting and ensemble learning and voting based on decision tree, random forest and logistic regression.

We have also implemented a simple sequential deep learning model using the Tensorflow library. A sequential 3-layer deep learning algorithm is a neural network model with three stacked layers: input, hidden, and output. The input layer prepares the data, the hidden layer learns features using activation functions, and the output layer produces the final prediction. The model is trained on labelled data using techniques like gradient descent and backpropagation. It's used for tasks like image recognition, natural language processing, and speech recognition, by automatically learning patterns from data.

Multiple regression is a statistical technique used to model the relationship between two or more independent variables and a single dependent variable. It aims to predict the value of the dependent variable based on independent variables using a linear equation. Multiple regression is commonly used for predicting outcomes or explaining the variability in data.

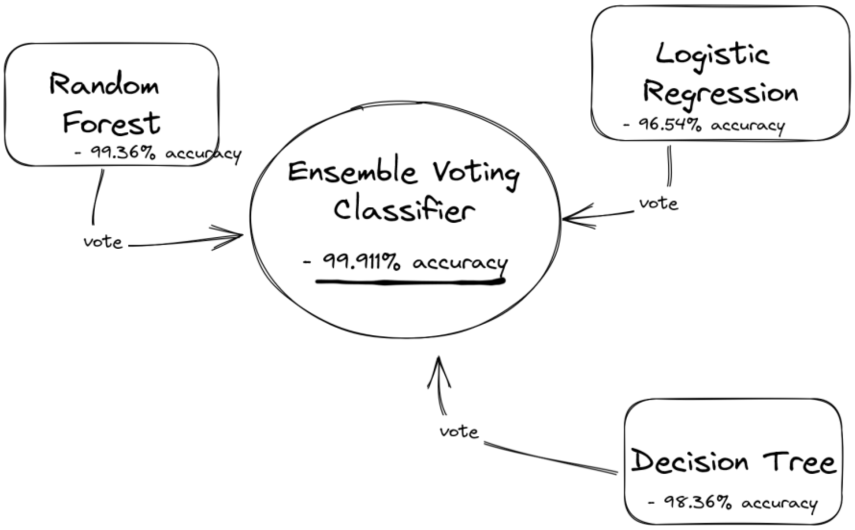
Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It works by finding a hyperplane that best separates the data into different classes or predicts the output values. SVM is known for its ability to handle non-linear data by using different kernel functions. It is widely used for classification tasks, especially when dealing with complex and non-linear data.

Decision tree is a popular machine learning algorithm used for both classification and regression tasks. It works by recursively splitting the data based on feature values to create a tree-like structure, where each node represents a decision based on a feature value, and each leaf node represents a prediction. Decision trees are easy to understand and interpret, and can handle both categorical and continuous data.

Gradient boosting is an ensemble learning technique that uses a collection of weak models, typically decision trees, to create a strong predictive model. It works by sequentially building and combining decision trees in a boosting fashion, where each subsequent tree is built to correct the errors of the previous trees. Gradient boosting is known for its high predictive accuracy and ability to handle complex data, making it popular for various machine learning tasks, including classification, regression, and ranking.

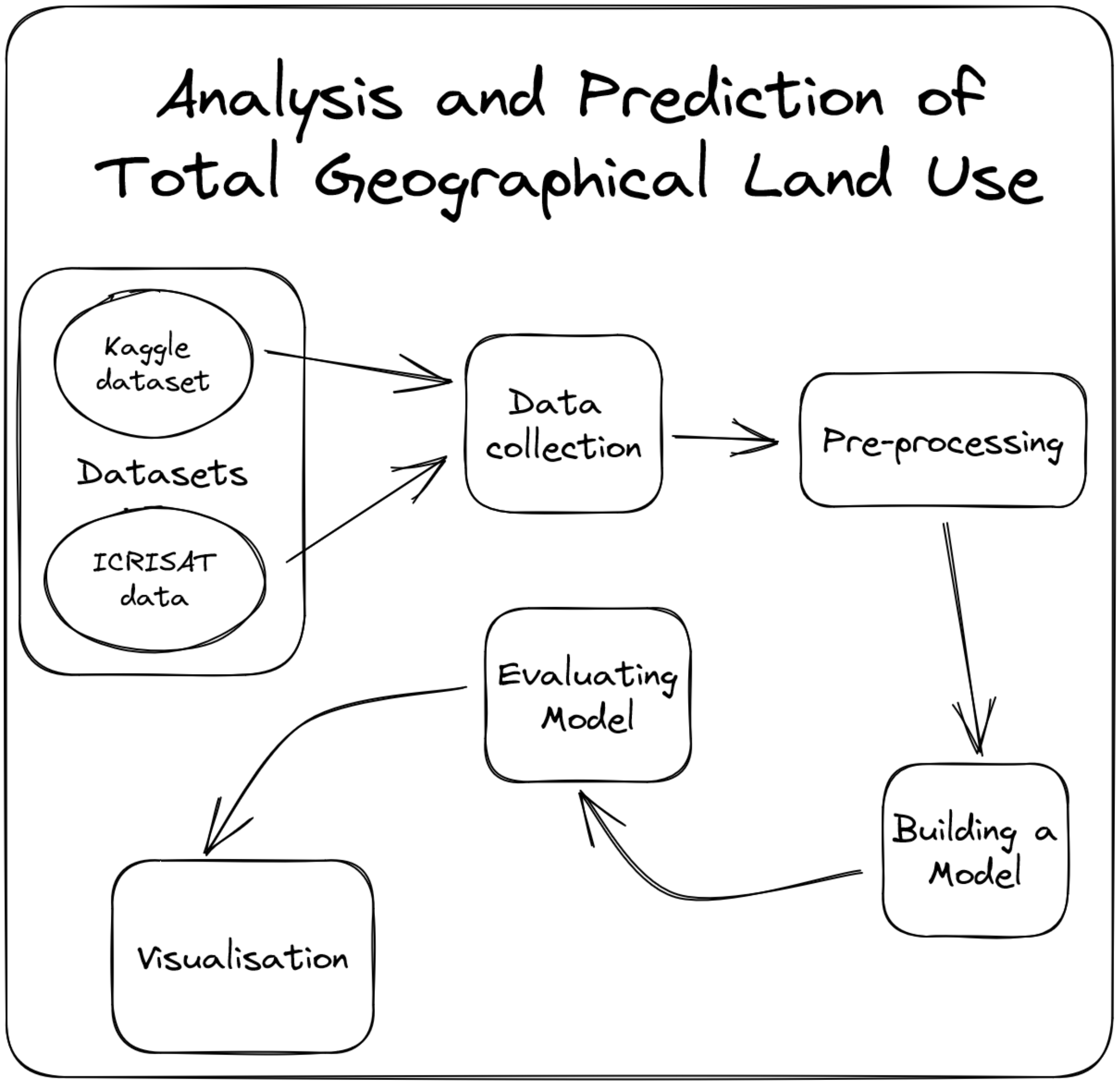
XGBoost is a gradient boosting algorithm used for classification and regression tasks. It builds an ensemble of decision trees sequentially, correcting errors from previous trees. It uses regularisation techniques to prevent overfitting and offers features like parallelization and early stopping for efficiency. Training involves feeding labelled data and tuning hyperparameters for optimal performance. XGBoost is popular for its accuracy and efficiency in various machine learning tasks.

Ensemble voting algorithm is a machine learning technique that combines the predictions of multiple base models to make a final prediction. It typically involves using different base models, such as decision trees, logistic regression, and random forests, and combining their predictions using a voting scheme. Ensemble voting algorithms can improve the overall performance and accuracy of the model by leveraging the strengths of different base models and reducing the impact of individual model's weaknesses.



Our ensemble voting classifier

Architecture



Our Architecture Diagram

Proposed Works

Novelty

In this research, we have taken a novel approach by combining multiple datasets, including monthly rainfall and land use data, to create a new dataset covering a 50-year period. This type of data is not available directly, which has allowed us to gain new and valuable insights that have not been explored before. By exploring various machine learning models, such as XGBoost and different types of voting and classification algorithms, we have assessed their accuracies and compared their results to determine the best-performing model. We have visualised these results in Tableau, presenting them in a creative and easy-to-understand manner that allows us to forecast future trends based on previous years' patterns.

Our use of deep learning with sequential neural networks for crop recommendation datasets is a novel approach. This technique allows us to capture the complex data patterns and temporal dynamics involved in crop recommendation tasks effectively. Deep learning models can handle large-scale data from multiple sources and provide real-time recommendations, making them a promising and innovative approach for agricultural applications. They also have the potential to overcome limitations in human expertise by automatically learning from data. Our research highlights the potential for using this approach to enhance the accuracy, timeliness, and efficiency of crop recommendation systems.

To visually represent our findings, we have utilised a variety of tools and techniques, including over 10 tree maps showing different biophysical attributes, geographical maps displaying states with the most rainfall, and comparison graphs between different attributes in the dataset. This approach offers a unique way of presenting and interpreting the data, making it more accessible to a wider audience. Overall, our research contributes to the development of innovative and practical solutions for agricultural applications by leveraging novel data sources, machine learning techniques, and data visualisation tools.

Project Contributions

1. GS Jyothssena

a. Data Collection

- Different datasets based on agriculture were analysed and the most suitable ones were chosen for this project

b. Support Vector Classifier

- Linear kernel was giving satisfactory results, but used fine tuning to give better accuracies. Tested poly-kernel by tweaking parameters, but it lead to intensive overfitting.

c. LogisticRegression

- Implemented Logistic Regression that was imported from sklearn package to predict the crop based on environment conditions. This model was further used in the voting algorithm implementation, in both the soft and hard methods.

d. Tableau Visualization

- Various types of visualisations were used to understand the datasets in Tableau

2. Prathiba Lakshmi Narayan

a. Pre-processing

- The unnecessary columns were removed and zeros and outliers were handled.

b. Random Forest

- Built a RF model with 50 estimators on a random subsection of the dataset, to achieve the highest accuracy we had - ~99%.

c. Gradient Boosting

- Researching newer models and on improving the low accuracy we initially had, settled on using Gradient boosting.

d. XGBoost

- Built a model (xg_model) that uses the gradient boosting algorithm, to achieve better accuracies. Used the open source XGBoost library.

3. Vishal N

a. Decision Tree

- Implemented the algorithm by using the DecisionTreeClassifier from the scikit-learn module. This model was further used in the voting algorithm implementation, in both the soft and hard methods.

b. Voting Algorithm (Soft and hard)

- Using the EnsembleVoter built into scikit-learn, it was possible to implement a hard voter by assigning higher weightage to more accurate models like random forest. We also built a soft model using the same class to compare accuracies.

c. Deep Learning

- Using Keras from Tensorflow, implemented a deep learning model that iteratively improved (30 epochs) to achieve an accuracy of ~98%.

d. Python Visualisation

- Using the popular seaborn module, created beautiful visualisations of the correlations between the attributes and other interesting observations from the dataset, such as how humidity levels in the atmosphere affect Potassium levels.

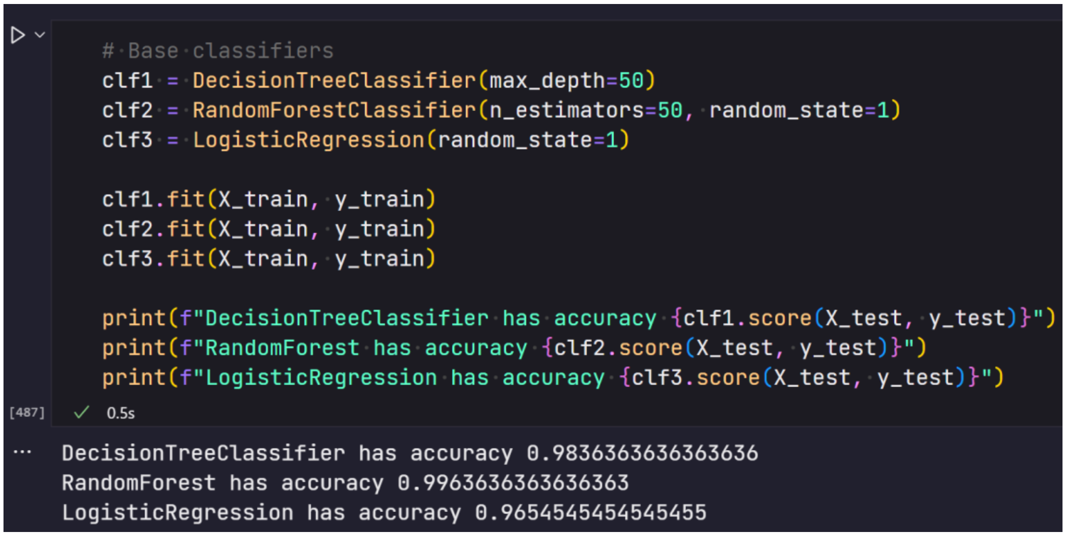
Results and Discussion

Results

Figures and Comparison Tables

Model	Accuracy
Decision Tree	98.36%
Random Forest	99.36%
Logistic Regression	96.54%
Support Vector Classifier	98.72%
XGBoost	99.45%
DeepLearning	97.58% (after 30 epochs)
Ensemble - Voting (Theoretical accuracy)	99.91%
Ensemble - Voting (Real Life accuracy)	98.18%

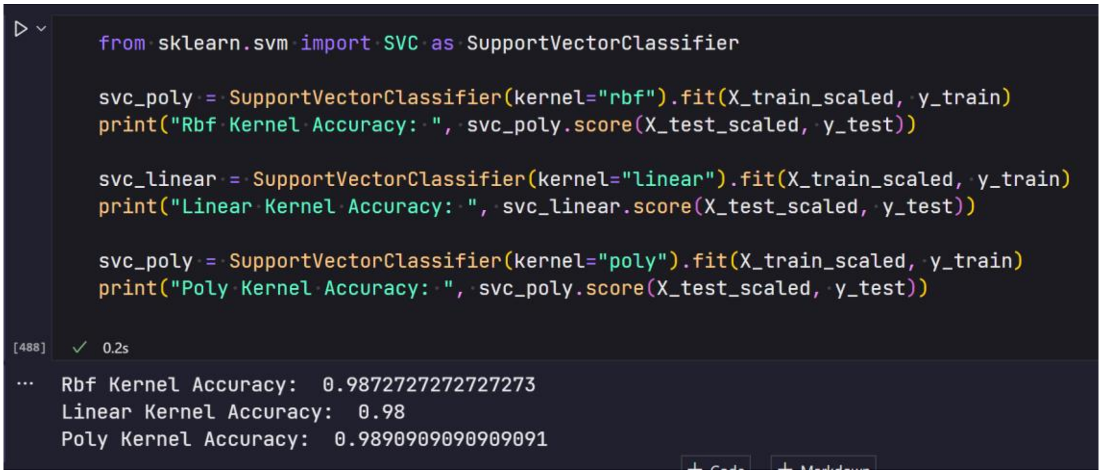
Table 1: Comparison of accuracies of all models we have used in our project.



```python
# Base classifiers
clf1 = DecisionTreeClassifier(max_depth=50)
clf2 = RandomForestClassifier(n_estimators=50, random_state=1)
clf3 = LogisticRegression(random_state=1)

clf1.fit(X_train, y_train)
clf2.fit(X_train, y_train)
clf3.fit(X_train, y_train)

print(f"DecisionTreeClassifier has accuracy {clf1.score(X_test, y_test)}")
print(f"RandomForest has accuracy {clf2.score(X_test, y_test)}")
print(f"LogisticRegression has accuracy {clf3.score(X_test, y_test)}")
```

```
DecisionTreeClassifier has accuracy 0.9836363636363636
RandomForest has accuracy 0.9963636363636363
LogisticRegression has accuracy 0.9654545454545455
```

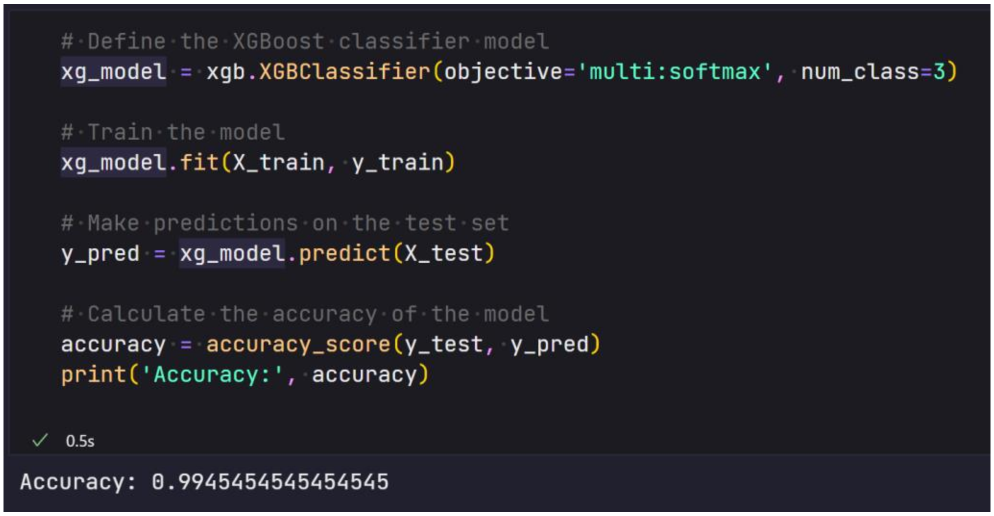
Fig. 1



```python
from sklearn.svm import SVC as SupportVectorClassifier

svc_poly = SupportVectorClassifier(kernel="rbf").fit(X_train_scaled, y_train)
print("Rbf Kernel Accuracy: ", svc_poly.score(X_test_scaled, y_test))

svc_linear = SupportVectorClassifier(kernel="linear").fit(X_train_scaled, y_train)
print("Linear Kernel Accuracy: ", svc_linear.score(X_test_scaled, y_test))

svc_poly = SupportVectorClassifier(kernel="poly").fit(X_train_scaled, y_train)
print("Poly Kernel Accuracy: ", svc_poly.score(X_test_scaled, y_test))
```

```
Rbf Kernel Accuracy:  0.9872727272727273
Linear Kernel Accuracy:  0.98
Poly Kernel Accuracy:  0.9890909090909091
```

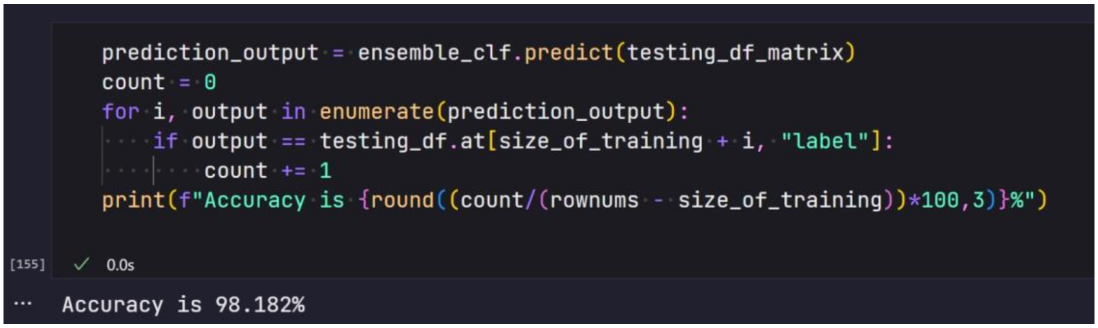
Fig. 2



```python
# Define the XGBoost classifier model
xg_model = xgb.XGBClassifier(objective='multi:softmax', num_class=3)

# Train the model
xg_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = xg_model.predict(X_test)

# Calculate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)
```

```
Accuracy: 0.9945454545454545
```

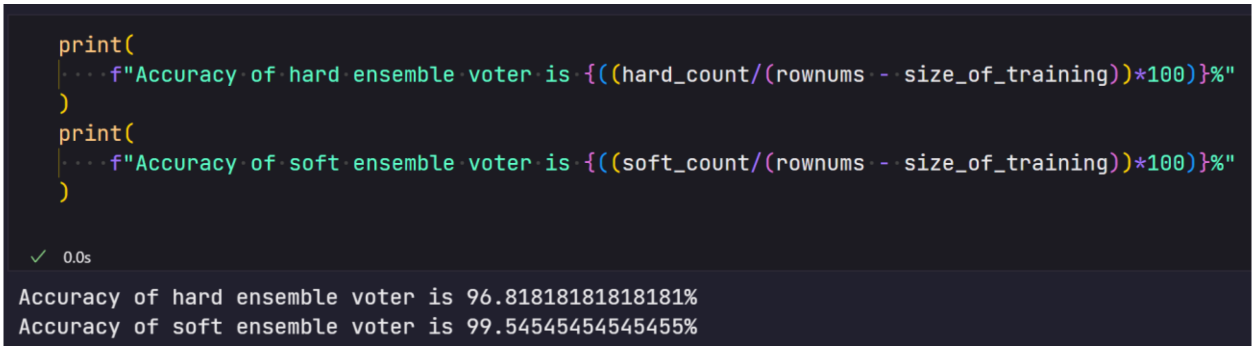
Fig 3.



```python
prediction_output = ensemble_clf.predict(testing_df_matrix)
count = 0
for i, output in enumerate(prediction_output):
    if output == testing_df.at[size_of_training + i, "label"]:
        count += 1
print(f"Accuracy is {round((count/(rownums - size_of_training))*100,3)}%")
```

```
Accuracy is 98.182%
```

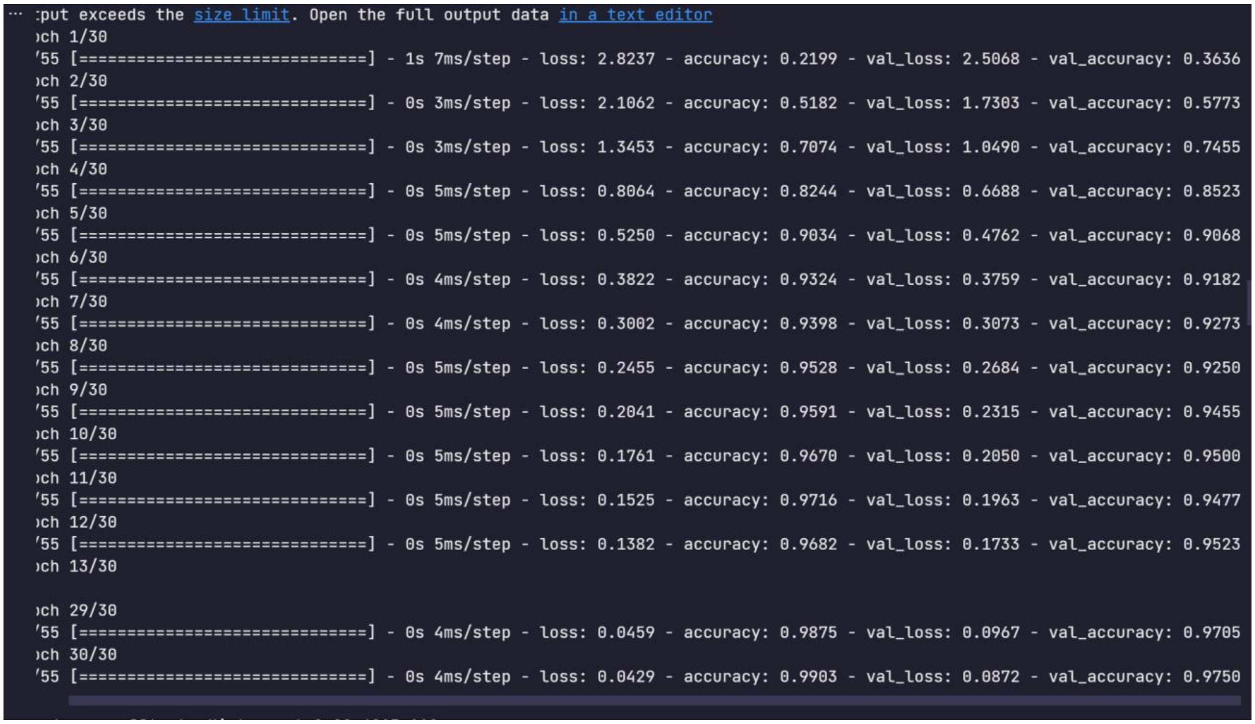
Fig 4.



```python
print(
    f"Accuracy of hard ensemble voter is {((hard_count/(rownums - size_of_training))*100)}%"
)
print(
    f"Accuracy of soft ensemble voter is {((soft_count/(rownums - size_of_training))*100)}%"
)
```

```
Accuracy of hard ensemble voter is 96.81818181818181%
Accuracy of soft ensemble voter is 99.54545454545455%
```

Fig. 5



```
··· :put exceeds the size limit. Open the full output data in a text editor
   ›ch 1/30
   '55 [==============================] - 1s 7ms/step - loss: 2.8237 - accuracy: 0.2199 - val_loss: 2.5068 - val_accuracy: 0.3636
   ›ch 2/30
   '55 [==============================] - 0s 3ms/step - loss: 2.1062 - accuracy: 0.5182 - val_loss: 1.7303 - val_accuracy: 0.5773
   ›ch 3/30
   '55 [==============================] - 0s 3ms/step - loss: 1.3453 - accuracy: 0.7074 - val_loss: 1.0490 - val_accuracy: 0.7455
   ›ch 4/30
   '55 [==============================] - 0s 5ms/step - loss: 0.8064 - accuracy: 0.8244 - val_loss: 0.6688 - val_accuracy: 0.8523
   ›ch 5/30
   '55 [==============================] - 0s 5ms/step - loss: 0.5250 - accuracy: 0.9034 - val_loss: 0.4762 - val_accuracy: 0.9068
   ›ch 6/30
   '55 [==============================] - 0s 4ms/step - loss: 0.3822 - accuracy: 0.9324 - val_loss: 0.3759 - val_accuracy: 0.9182
   ›ch 7/30
   '55 [==============================] - 0s 4ms/step - loss: 0.3002 - accuracy: 0.9398 - val_loss: 0.3073 - val_accuracy: 0.9273
   ›ch 8/30
   '55 [==============================] - 0s 5ms/step - loss: 0.2455 - accuracy: 0.9528 - val_loss: 0.2684 - val_accuracy: 0.9250
   ›ch 9/30
   '55 [==============================] - 0s 5ms/step - loss: 0.2041 - accuracy: 0.9591 - val_loss: 0.2315 - val_accuracy: 0.9455
   ›ch 10/30
   '55 [==============================] - 0s 5ms/step - loss: 0.1761 - accuracy: 0.9670 - val_loss: 0.2050 - val_accuracy: 0.9500
   ›ch 11/30
   '55 [==============================] - 0s 5ms/step - loss: 0.1525 - accuracy: 0.9716 - val_loss: 0.1963 - val_accuracy: 0.9477
   ›ch 12/30
   '55 [==============================] - 0s 5ms/step - loss: 0.1382 - accuracy: 0.9682 - val_loss: 0.1733 - val_accuracy: 0.9523
   ›ch 13/30

   ›ch 29/30
   '55 [==============================] - 0s 4ms/step - loss: 0.0459 - accuracy: 0.9875 - val_loss: 0.0967 - val_accuracy: 0.9705
   ›ch 30/30
   '55 [==============================] - 0s 4ms/step - loss: 0.0429 - accuracy: 0.9903 - val_loss: 0.0872 - val_accuracy: 0.9750
```

Fig. 6

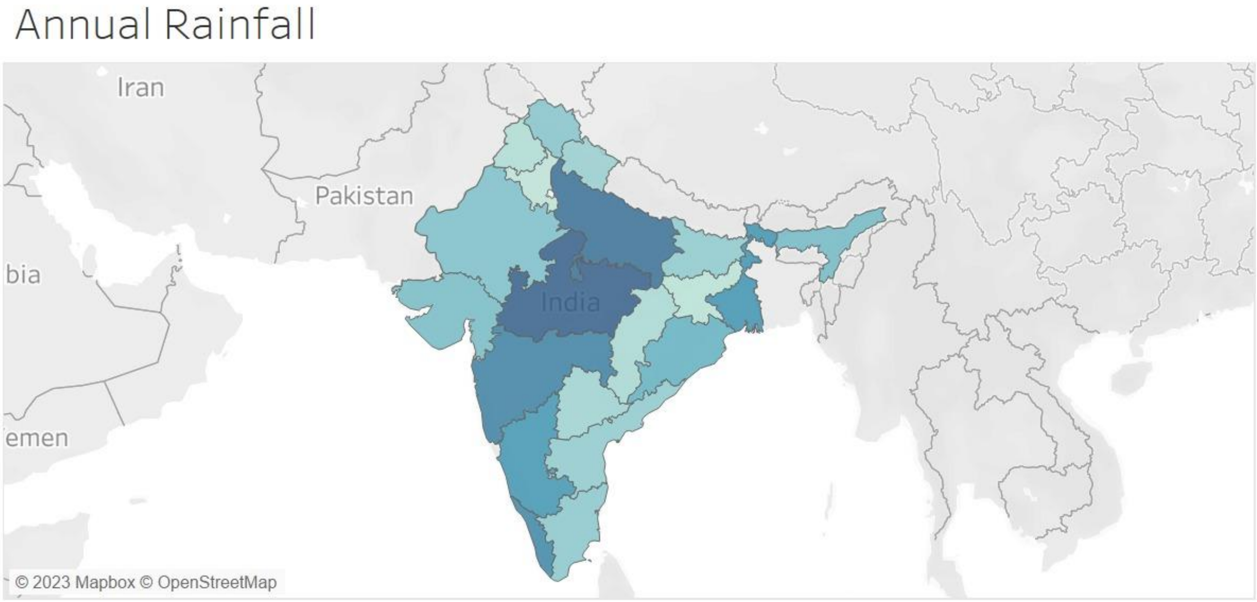


Fig. 7

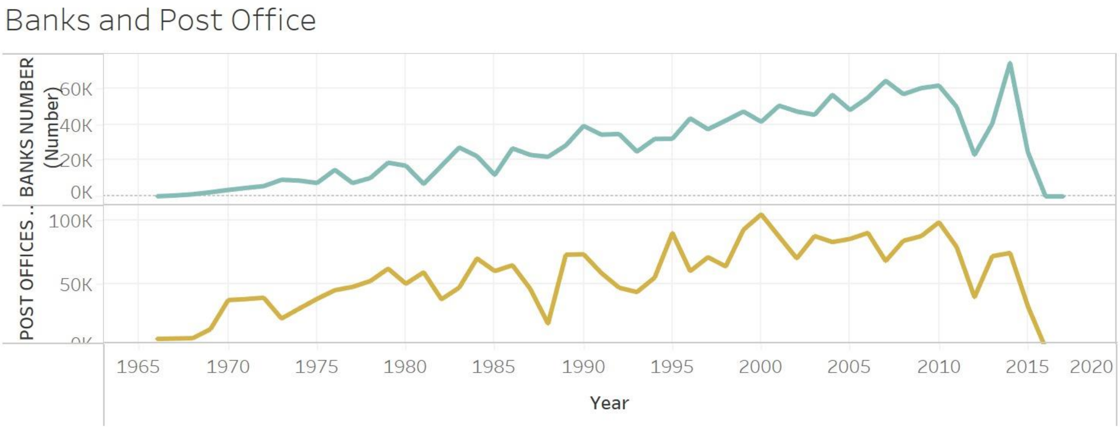

Banks and Post Office
BANKS NUMBER (Number)
60K
40K
20K
0K
POST OFFICES ..
100K
50K
0K
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

Fig 8.

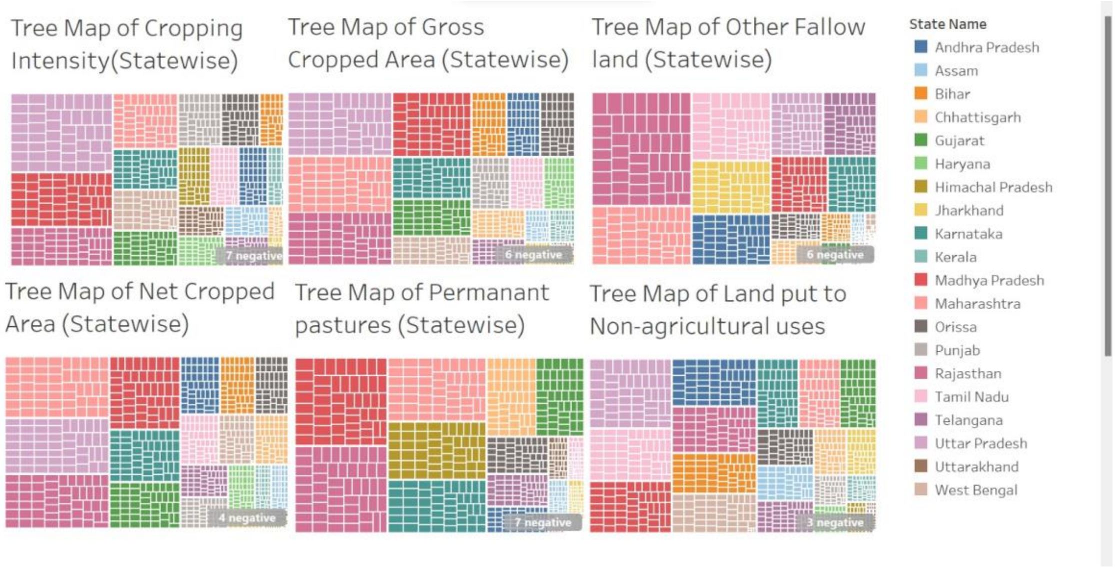

Tree Map of Cropping Intensity(Statewise)
Tree Map of Gross Cropped Area (Statewise)
Tree Map of Other Fallow land (Statewise)
7 negative
6 negative
6 negative
Tree Map of Net Cropped Area (Statewise)
Tree Map of Permanant pastures (Statewise)
Tree Map of Land put to Non-agricultural uses
4 negative
7 negative
3 negative
State Name
Andhra Pradesh
Assam
Bihar
Chhattisgarh
Gujarat
Haryana
Himachal Pradesh
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Orissa
Punjab
Rajasthan
Tamil Nadu
Telangana
Uttar Pradesh
Uttarakhand
West Bengal

Fig. 9a

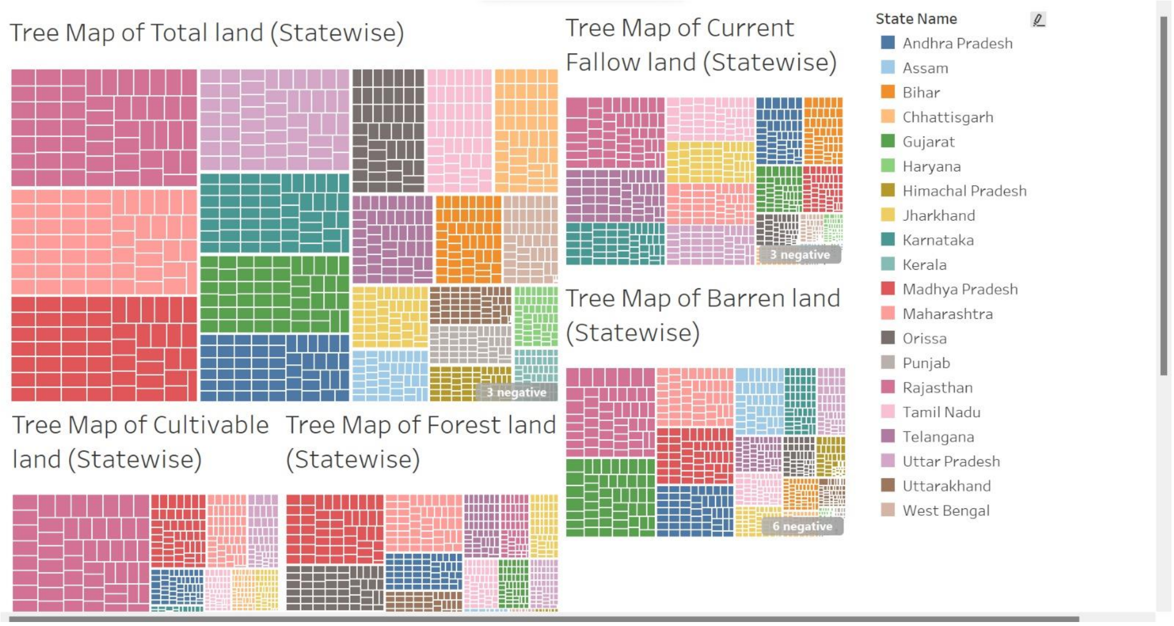


Fig. 9b

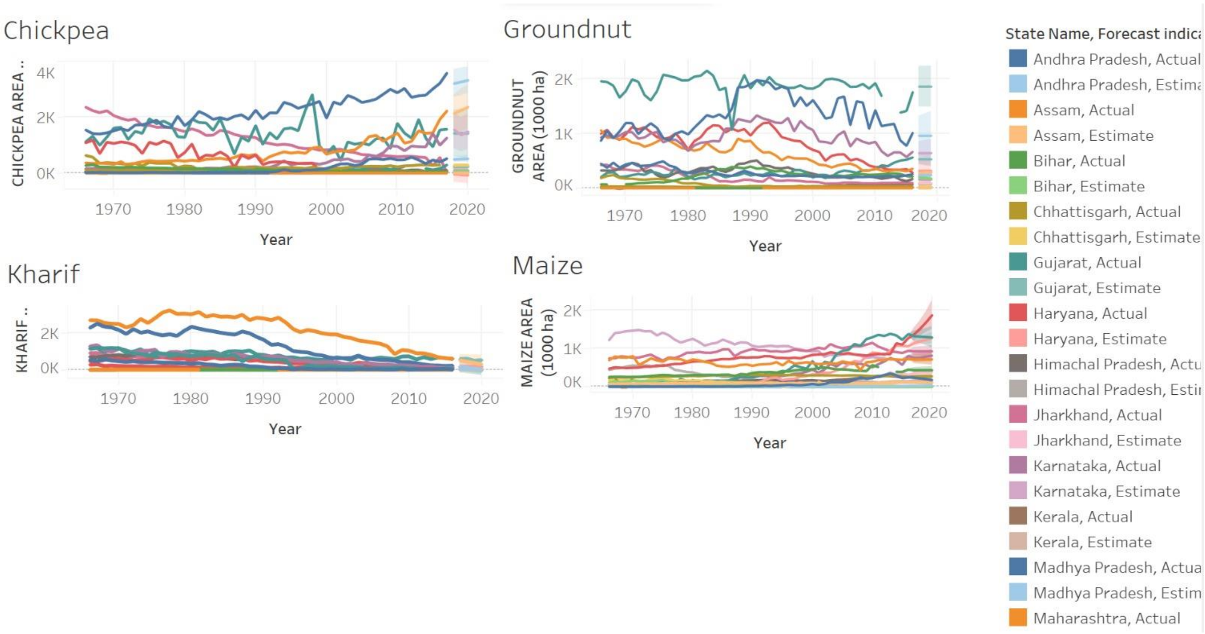


Fig. 10a

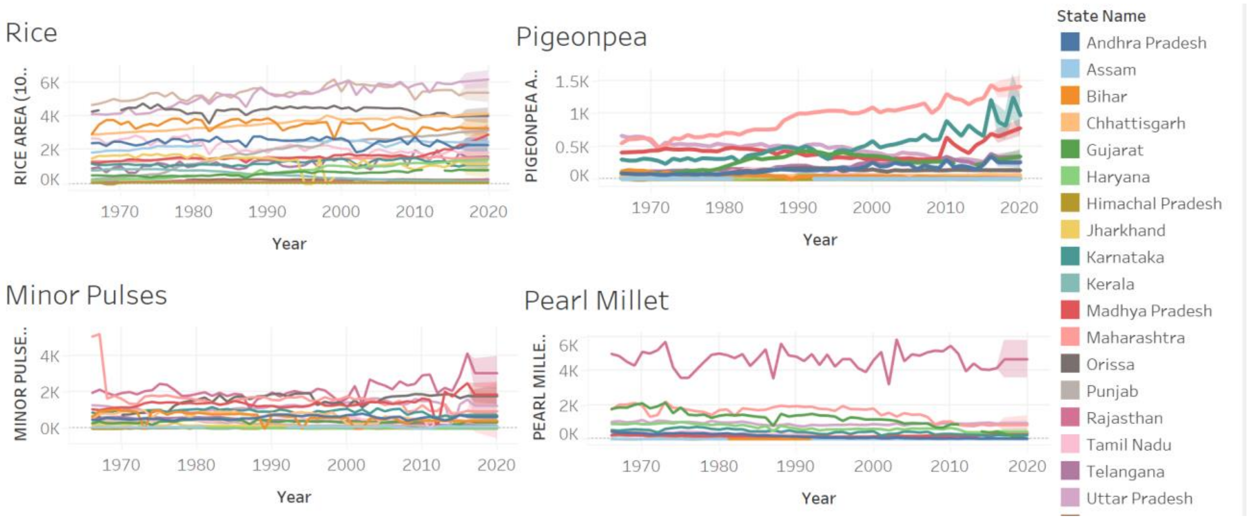


Fig 10b

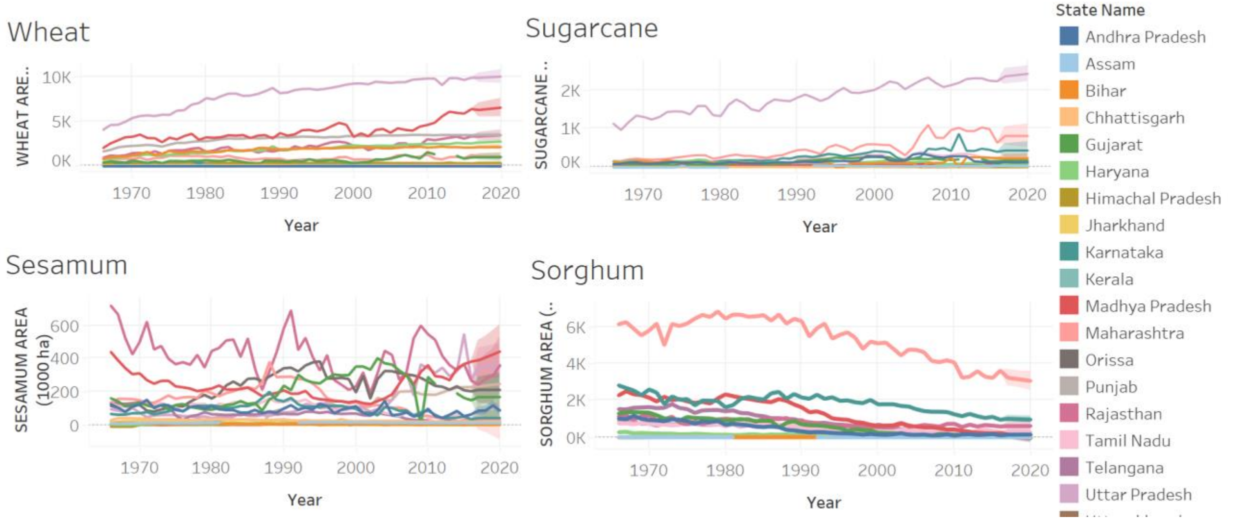


Fig 10c

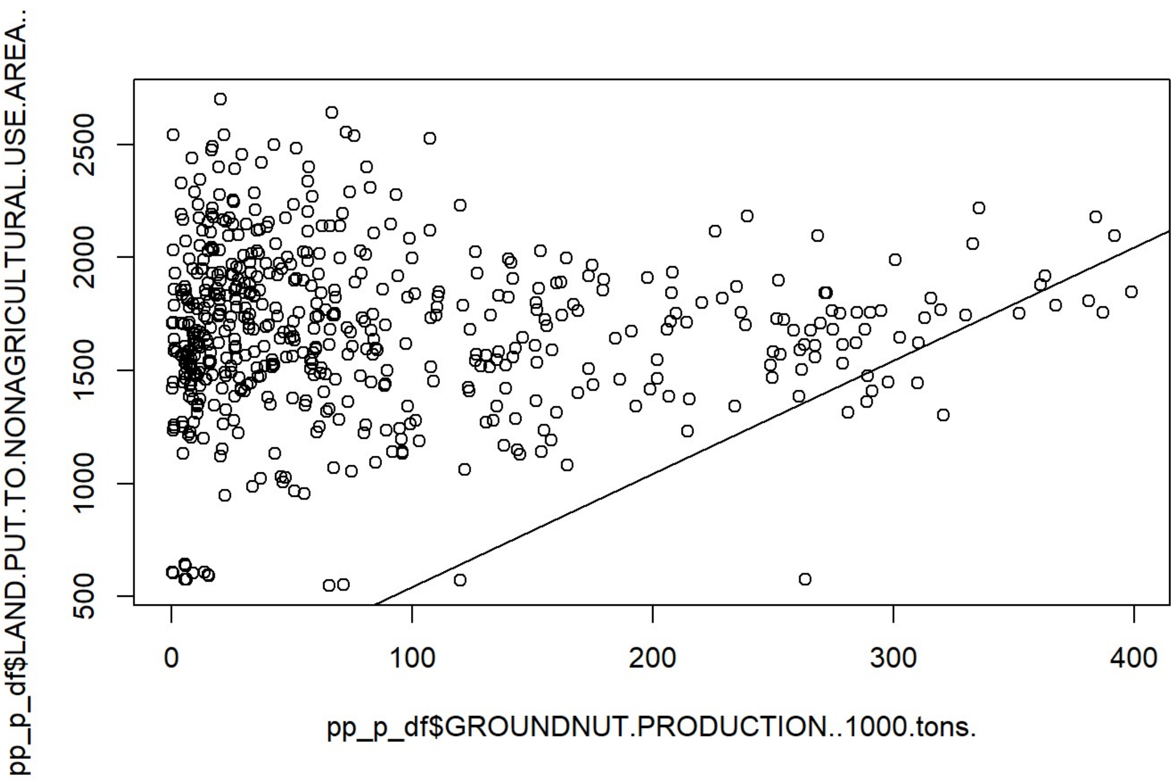


Fig 11a

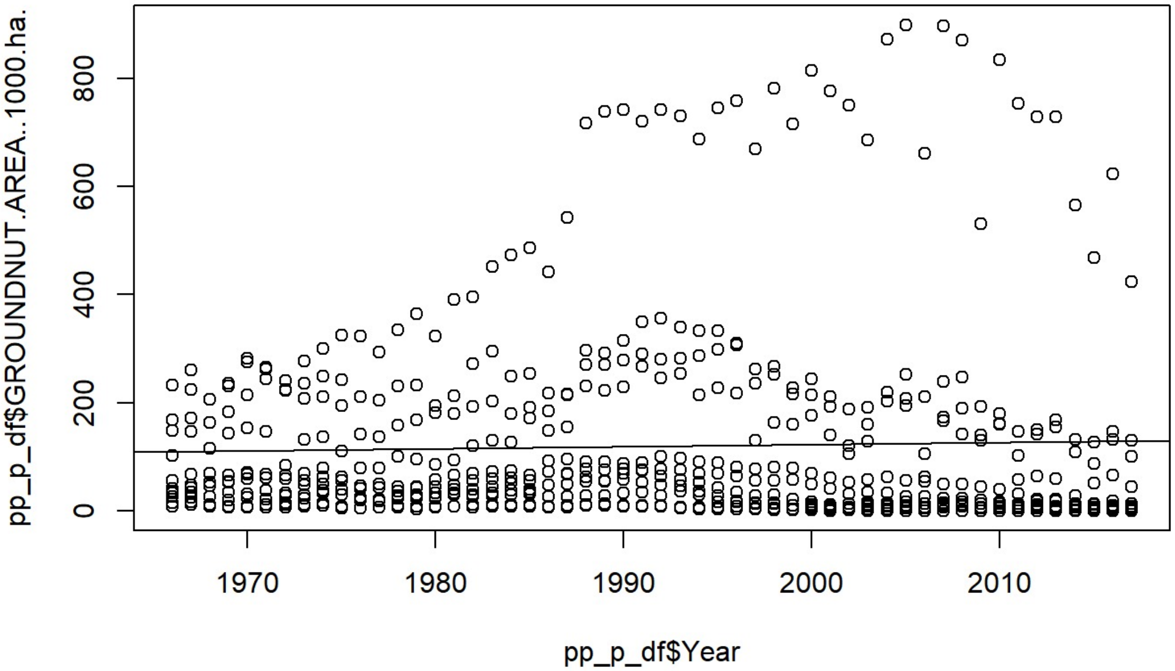


Fig 11b

Explanation

In Fig 1, We used 3 different classifiers to build our voting system, and each had varying levels of accuracy.

In Fig 2, We used 3 different kernels for SVC. Rbf, Linear and Poly. They demonstrated different accuracies.

In Fig 3, XGBoot seemed to be the best fit model, achieving about 99% accuracy consistently.

In Fig 4, Accuracy of the ensemble model was found to be 98.182%.

In Fig 5, When using “soft” and “hard” ensemble voting, the accuracy was different

In Fig 6, Using the deep learning framework Tensorflow, and running 30 epochs, we achieved 97.5% accuracy.

In Fig. 7, the map of India is shown as a heat map, where darker colours depict higher amount if rainfall.

In Fig. 8, the number of banks and post offices are depicted using a line chart.

In Fig. 9a, and 9b, tree maps are used for the different Biophysical attributes.

Fig. 10a,10b and 10c shows the time series chart of the various crops along with forecasting values between one year and 3.

Fig 11a and 11b shows the groundnut production with respect to land put to non agricultural use and groundnut production over the years for Andhra Pradesh

Conclusion

In conclusion, the data visualisation project on analysing crops grown in different states in India using NPK values, rainfall, pH, and humidity, in combination with machine learning algorithms such as linear regression, ensemble learning, KNN, SVC, and random forest, has yielded valuable insights and actionable information for stakeholders in the agriculture sector. These algorithms enabled the identification of key factors that significantly impact crop growth and yield, and the development of predictive models to estimate crop performance under different conditions. The project involved extensive data cleaning and preprocessing to ensure the accuracy and reliability of the results.

Through the use of various visualisations in Tableau and Python, including bar charts, line charts, heatmaps, and geographical maps, the project provided a comprehensive analysis of crop data, highlighting patterns, trends, and correlations between the different factors considered. The visualisations facilitated the identification of optimal crop-growing conditions, as well as the prediction of crop yields based on the input factors.

The project findings and visualisations can serve as a valuable resource for farmers, agronomists, policymakers, and other stakeholders involved in crop production and management.

References

[1] M. Kalimuthu, P. Vaishnavi and M. Kishore, "Crop Prediction using Machine Learning," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 926-932, doi: 10.1109/ICSSIT48917.2020.9214190.

[2] R.L Semwal, S Nautiyal, K.K Sen, U Rana, R.K Maikhuri, K.S Rao, K.G Saxena, Patterns and ecological implications of agricultural land-use changes: a case study from central Himalaya, India, Agriculture, Ecosystems & Environment, Volume 102, Issue 1, 2004, Pages 81-92, ISSN 0167-8809, https://doi.org/10.1016/S0167-8809(03)00228-7.

[3] P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, "Crop Yield Prediction based on Indian Agriculture using Machine Learning," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154036.

[4] L. Breiman. Random forests. Machine Learning, 45(1):5–32, Oct. 2001

[5] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[6] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction", IEEE Access, vol. 9, pp. 63406-63439, 2021.

[7] A. X. Wang, C. Tran, N. Desai, D. Lobell and S. Ermon, "Deep transfer learning for crop yield prediction with remote sensing data", Proc. 1st ACM SIGCAS Conf. Comput. Sustain. Soc. COMPASS 2018, 2018.

[8] E. G. Moung, C. C. Wooi, M. M. Sufian, C. K. On and J. A. Dargham, "Ensemble-based face expression recognition approach for image sentiment analysis", Int. J. Electr. Comput. Eng., vol. 12, no. 3, pp. 2588-2600, 2022.

[9] A. Nagaraju, M. Ajith, Kumar Reddy, CH. Venugopal reddy and R. Mohandas, "Multifactor Analysis to Predict Best Crop using Xg-Boost Algorithm", IEEE Xplore, Jun. 01, 2021, [online] Available: https://ieeexplore.ieee.org/abstract/document/9452918.

[10] V. Sellam and E. Poovammal, "Prediction of Crop Yield using Regression Analysis", Indian J. Sci. Technol., vol. 9, no. 38, Oct. 2016.

Appendix

1. GitHub link

https://github.com/vishalnandagopal/crop-analysis-and-prediction. This contains the code to run the prediction part (file: crop-analysis-and-prediction.ipynb), visualisations part (visualizatoins.twb) and other documents.