Project Submission

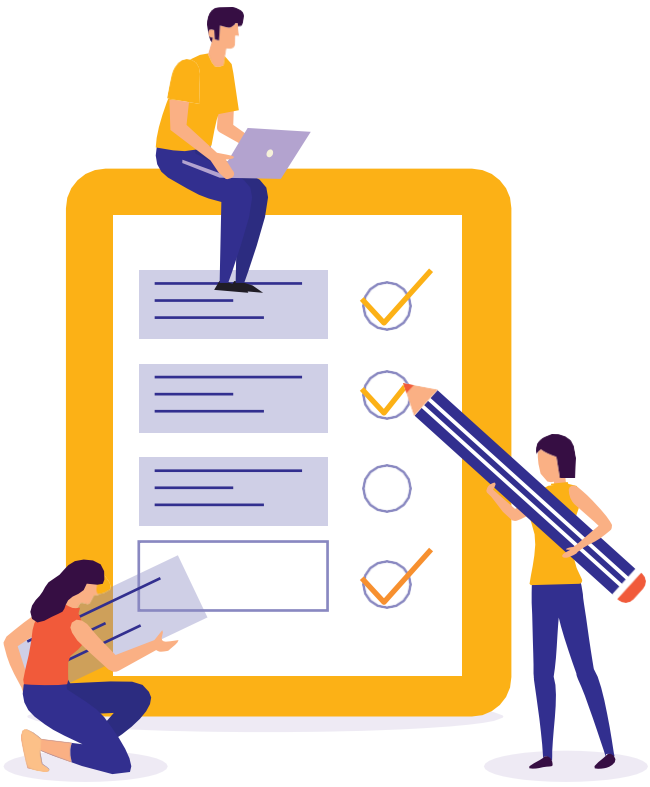
1. Title of the Project

Prediction of Agriculture Crop Production in India

Brief on the Project

This project aims to predict agricultural crop production in India using ML algorithm – Random Forest. The project has shown that technology can play a critical role in improving agricultural productivity and sustainability in India. By providing farmers and policymakers with accurate information about crop yields, production, and area under cultivation, the tool can help improve crop management, reduce crop losses, and increase food security. The use of data analytics and machine learning techniques has also shown promising results in predicting crop yields, which can further improve the accuracy and reliability of the tool. By employing feature selection techniques and evaluating multiple ML algorithm, the project aims to develop a robust and accurate predictive model. The deployment of such a model can significantly benefit farmers, policymakers, and other stakeholders by providing valuable insights for decision-making and optimizing agricultural practices.

Deliverables of the Project



The development of the tool using open-source technologies such as Python, Plotly, Pandas, NumPy, and Scikit-learn has made it cost-effective and easily accessible to users. Python, being a popular programming language, has a large community of developers constantly contributing to its growth and development. The use of Plotly has allowed for the creation of interactive visualizations, making it easier for users to analyze and interpret the data. The Pandas module has facilitated the processing of large datasets, while the NumPy module has enabled the manipulation and computation of large arrays of numerical data. The use of Scikit-learn has allowed for the implementation of various machine learning algorithms, including regression and classification models, to predict crop yields.

Resources

Data set source:

https://drive.google.com/file/d/1zfqvs8-mAO6E0JpgvhBdueNx8Th03pUp/view?usp=sharing

References:

- Agricultural Statistics at a Glance 2018: Ministry of Agriculture and Farmers' Welfare, Government of India. Available at: http://agricoop.nic.in/sites/default/files/asga2018_0.pdf
- 2. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- 3. J. VanderPlas. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, Inc., 2016.

Personal Details

Name – Shruti Shastri
E-mail Id: - shrutishastri13@gmail.com

Milestones

Create Features –

Identifying the most relevant features for predicting crop production can be complex. It requires domain knowledge and expertise to select the appropriate variables from a large pool of potential predictors. Additionally, engineering new features that capture complex relationships between variables can be time-consuming and challenging.

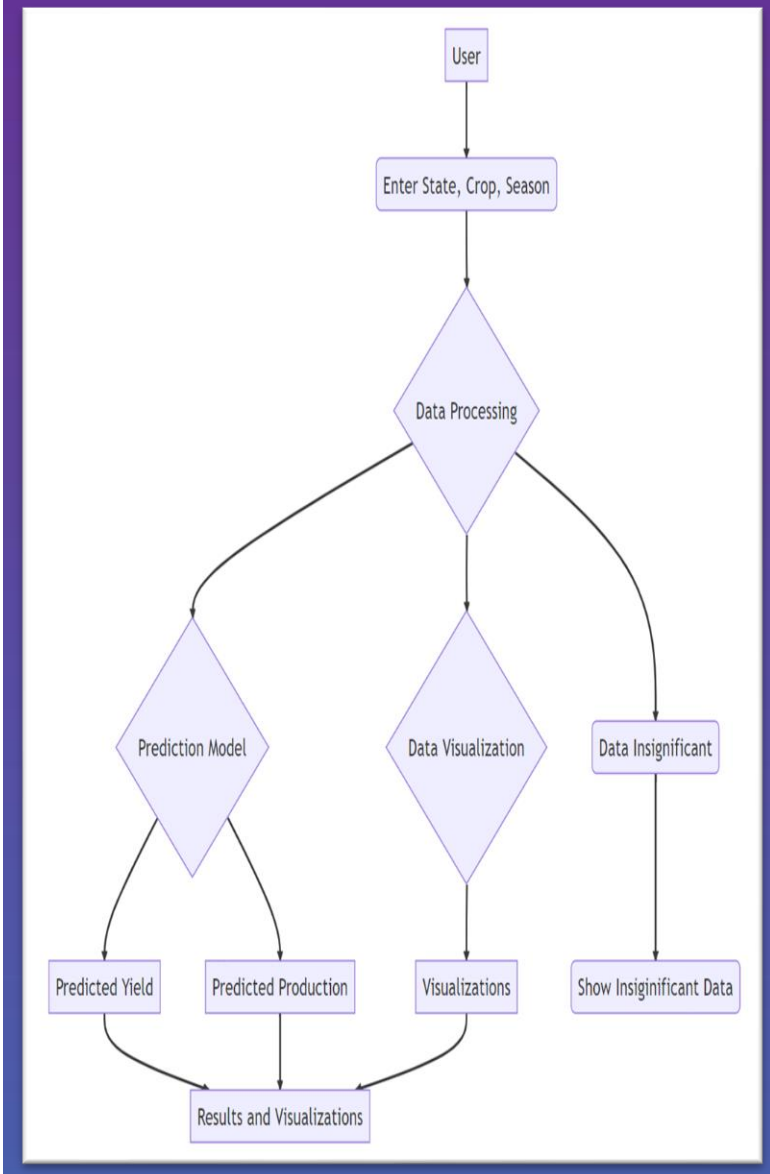
EDA –

- Handle outliers: Identify outliers by the box-plot technique and them decide to remove them or replace them with the mean of the data.
- Perform hypothesis testing: Conducted statistical tests to validate differences between groups.
- Explore data subsets or segments: Divide the dataset into meaningful subsets based on relevant variables and perform EDA on each subset individually. Compare the findings and identify any differences or patterns specific to each segment.
- Document insights: Analysed the findings, insights, and observations throughout the EDA process. This includes visualizations and key insights.

Feature Engineering –

- Feature extraction: Create new features from the existing data that can provide additional information and improve the performance of the model. This can involve techniques like dimensionality reduction (e.g., PCA) or creating new features based on domain knowledge.

Flowchart –



- Feature transformation: Transform the features to make them suitable for the machine learning algorithms. This can involve techniques like scaling, normalization, or encoding categorical variables.

- Feature selection: Select the most relevant and informative features for the model. This step helps to reduce dimensionality and improve model performance. Techniques like correlation analysis, feature importance, or recursive feature elimination can be used.

Create Model –

Model development must strike a balance between underfitting and overfitting. Underfitting occurs when a model is too simple to capture the underlying patterns in the data, resulting in poor predictive performance. Overfitting, on the other hand, happens when a model becomes too complex and memorizes noise in the training data, leading to poor generalization on unseen data. Finding the optimal model complexity is a challenge that requires careful model selection, regularization techniques, and hyperparameter tuning.

Model Evaluation –

In this step, the trained model is evaluated to determine its accuracy, precision, recall, and other performance metrics.

Model Evaluation Metrics:
Choosing appropriate evaluation metrics for crop production prediction can be challenging. The selection of metrics depends on the specific goals of the project, such as minimizing prediction errors or maximizing yield accuracy. Additionally, the interpretation of metrics may vary across stakeholders, and trade-offs between different metrics may need to be considered.

Accuracy testing and continuous improvement to the model –

The creation and evaluation of models for predicting agricultural crop production must be flexible and constantly improve. The agricultural industry is dynamic, and a number of variables, including climatic trends, technical development, and market variations, are always changing. To maintain the prediction models' relevance and accuracy over time, it is crucial to constantly improve and update them.

To supplement the current dataset, continuous improvement entails gathering more recent and pertinent data. The models can capture the most recent trends and patterns in crop output by including the most recent data. This can involve including current meteorological information, satellite images, or details on pests and diseases that have a big impact on agricultural output.

The models must be adaptable enough to include modifications as new agricultural practices, methods, and problems arise. To capture complicated relationships, this may need changing the model architecture, adding additional features, or utilizing cutting-edge machine learning techniques.

To spot problem areas and correct any biases or flaws, the model's performance must be regularly monitored and evaluated. Potential flaws or biases can be found and fixed by comparing the model's predictions to real crop production data.

To guarantee that the models satisfy the requirements of farmers, policymakers, and other end users, stakeholder input and cooperation are crucial. Engaging with subject matter experts, agricultural practitioners, and important stakeholders can yield insightful information about particular difficulties and needs, guiding continual improvement and adaptation.

Conclusion –

By this project we have developed a Machine learning Model that leverages technology and data analytics to provide insights into crop yields, production, and area under cultivation in India. The tool has been developed using a comprehensive dataset comprising information on crop yields, production, and area under cultivation in different states and regions of India. The study has also evaluated the effectiveness of different algorithms and techniques for predicting crop yields, with the aim of providing accurate information to farmers and policymakers.

The project has shown that technology can play a critical role in improving agricultural productivity and sustainability in India. By providing farmers and policymakers with accurate information about crop yields, production, and area under cultivation, the tool can help improve crop management, reduce crop losses, and increase food security. The use of data analytics and machine learning techniques has also shown promising results in predicting crop yields, which can further improve the accuracy and reliability of the tool.

The development of the tool using open-source technologies such as Python, Plotly, Matplotlib Pandas, NumPy, and Scikit-learn has made it cost-effective and easily accessible to users. Python, being a popular programming language, has a large community of developers constantly contributing to its growth and development. The use of Plotly has allowed for the creation of interactive visualizations, making it easier for users to analyze and interpret the data. The Pandas module has facilitated the processing of large datasets, while the NumPy module has enabled the manipulation and computation of large arrays of numerical data. The use of Scikit-learn has allowed for the implementation of various machine learning algorithms, including regression and classification models, to predict crop yields.

The project has also highlighted the importance of collaboration between different stakeholders, including researchers, farmers, policymakers, and technology experts, in addressing the challenges facing the agricultural sector in India. By working together and sharing knowledge and resources, these stakeholders can help improve the efficiency and sustainability of the agricultural sector, ultimately improving the livelihoods of farmers and ensuring food security for the growing population.