Project Report

On

Crop Yield Prediction using Machine Learning

Submitted in the partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

from Know-IT ATC, CDAC ACTS, Pune

Guided by:

Mr. Anay Tamhankar

Mr. Prasad Deshmukh

Submitted By:

Pradnya Bhosale (220343025009)

Neha Ghadage (220343025010)

Shrikant Shingne (220343025045)

Hitesh Upare (220343025052)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Pradnya Bhosale (220343025009)

Neha Ghadage (220343025010)

Shrikant Shingne (220343025045)

Hitesh Upare (220343025052)

have successfully completed their project on

Crop Yield Prediction using Machine Learning

Under the guidance of Mr. Anay Tamhnakar and Mr. Prasad Deshmukh

ACKNOWLEDGEMENT

This project "Crop Yield Prediction using Machine Learning" was agreat learning experience for us and we are submitting this work to Know-IT ATC, CDAC ACTS,Pune.

We are all very glad to mention the names of **Mr. Anay Tamhankar** and **Mr. Prasad Deshmukh** for their valuable guidance on this project. Their continuous guidance and support helped us overcome various obstacles and intricacies during the course of the project work.

We are highly grateful to **Mr. Vaibhav Inamdar,** the Center Coordinator at Know-it in Pune, for his guidance and support whenever necessary while we were pursuing the Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS in Pune.

Our most heartfelt thanks goes to **Mr. Shrinivas Jadhav** (Vice-President, Know-it, Pune) who provided all the required support and his kind co-ordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project, throughout the course and till date, here in Know-IT ATC, CDAC ACTS, Pune.

From:

Pradnya Bhosale (220343025009)
Neha Ghadage (220343025010)
Shrikant Shingne (220343025045)
Hitesh Upare (220343025052)

TABLE OF CONTENTS

ABSTRACT

In this project, we are using machine learning to predict the crop production in India. We are using a dataset that contains information about the area, production, and yield of various crops across different states in India from 2000 to 2014. We are applying different machine learning algorithms to this dataset and comparing their performance. We are also exploring the factors that affect the crop production and how they vary across different crops and regions. Our goal is to develop a reliable and accurate model that can predict the crop production in India for future years. This model can be useful for farmers, policy makers, researchers, and anyone who is interested in the agricultural sector in India.

1. INTRODUCTION

Agricultural sector is a dynamic and complex sector of the economy. Understanding the factors that influence crop production is crucial for farmers, policymakers, and other stakeholders. With the rapid advancement of machine learning techniques, it has become possible to analyze large datasets and develop accurate models for predicting crop yields. By leveraging historical data on crop production, weather conditions, soil characteristics, and other relevant factors, machine learning algorithms can accurately forecast crop yields. This can help farmers make informed decisions about resource allocation and crop selection, ultimately improving the efficiency and productivity of the agricultural sector.

- In this report, we present a study of crop production prediction using machine learning on a crop production and rainfall dataset. The crop production dataset contains information on State name, district name, crop year, season, crop area, Production
- We will use this dataset to develop and evaluate machine learning models that can predict crop yields based on the available features.
- The main objective of this study is to explore the performance of different machine learning algorithms for crop production prediction and identify the most accurate model. We will also investigate the significance of different features in predicting crop yields and explore ways to improve the model’s accuracy.
- Overall, this report aims to provide insights into the use of machine learning techniques for crop production prediction and the factors that influence crop yields in the regions covered by the dataset.

Datasets and features:

- Data used was collected from www.kaggle.com . These dataset provides a huge amount of information on crop production and rainfall in India ranging from several years.
- However, overall the datasets provides a rich source of data for analyzing patterns and trends that affects crop production in India.
- The main goal of the analysis is to build an accurate and robust regression model to predict the outcome of Crop production. This project uses Random Forest, Linear Regression, KNN Regression.

2. SYSTEM REQUIREMENTS

Hardware Requirements:

- Platform – Windows 7 or above
- RAM – Recommended 8 GB of RAM
- Peripheral Devices – Keyboard, Monitor, Mouse
- WiFi connection with minimum 2 Mbps speed

Software Requirements:

- Language: Python 3
- Machine Learning
- Tableau
- OS – Windows

3. FUNCTIONAL REQUIREMENTS

1) Python 3:

- Python is a high-level programming language that is easy to learn and use.
- Python is an interpreted language, which means that code can be executed on the fly, without the need for compilation.
- Python is open source and free to use, with a large and active community of developers contributing to its development and maintenance.
- Python has a vast collection of third-party libraries and packages, such as NumPy, Pandas, Matplotlib, and Scikit-learn, among others, that make it easy to perform data analysis.

2) Tableau:

- Tableau is a data visualization and business intelligence software that allows users to connect, analyse, and share data in a visual and interactive way.
- It offers a user-friendly drag-and-drop interface that enables users to create interactive dashboards, reports, and charts without the need for complex coding or programming.
- Tableau supports various data sources, including spreadsheets, databases, cloud services, and bigdata platforms, such as Hadoop and Spark.

Data Cleaning:

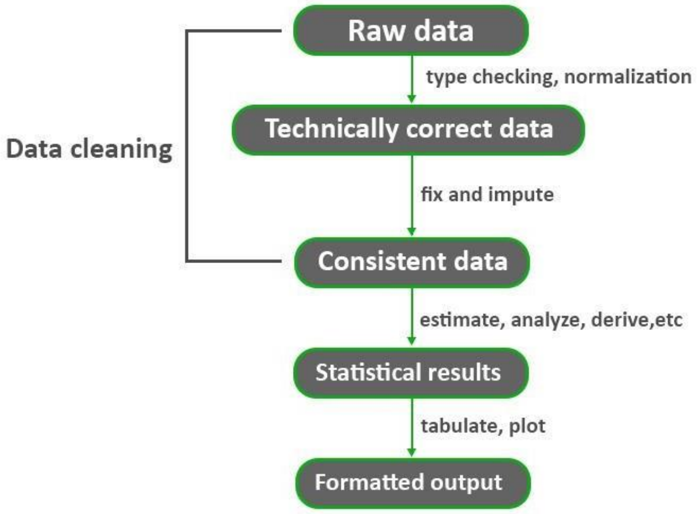


Fig: Data Cleaning Process

- Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywherein information systems. These problems are solved by data cleaning.

- Without proper data cleaning, data analysis and modelling can lead to erroneous or biased results,which can have serious consequences for businesses and organizations.

- Hence, it is a critical step in the data preparation process, as it can significantly impact the accuracyand reliability of the insights and decisions that are derived from the data. By improving the quality of data, organizations can gain a better understanding of their operations, customers, and market trends, and make more informed and effective decisions.

4. SYSTEM ARCHITECTURE

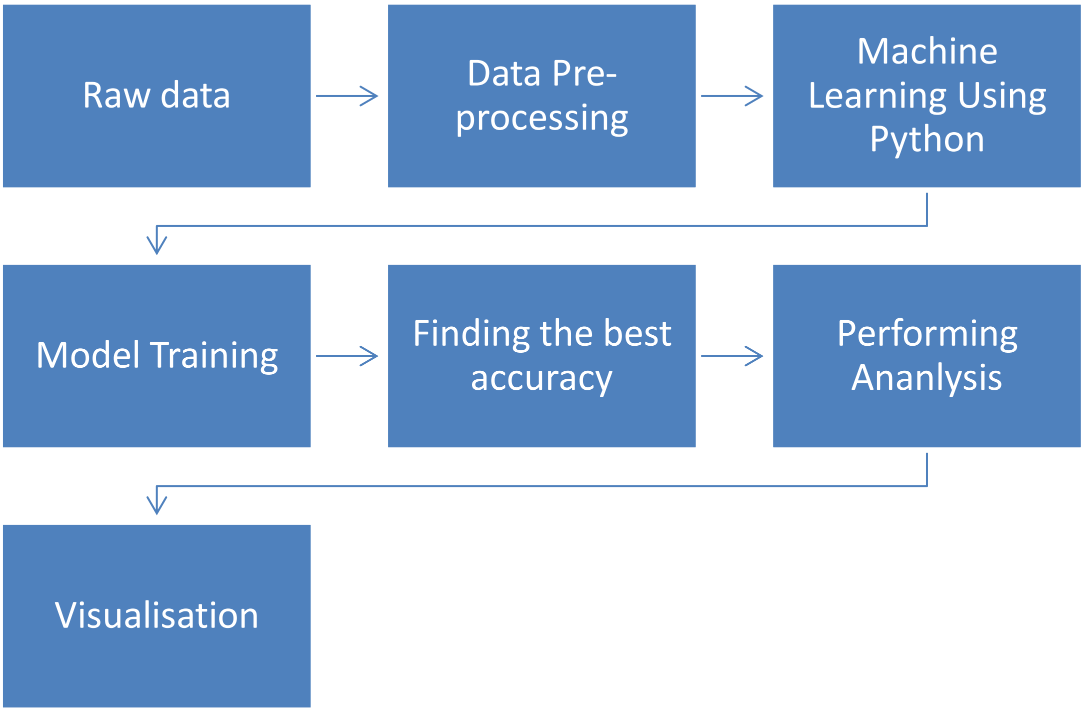


Fig: System Architecture of Crop Yield Prediction

5. METHODOLOGY

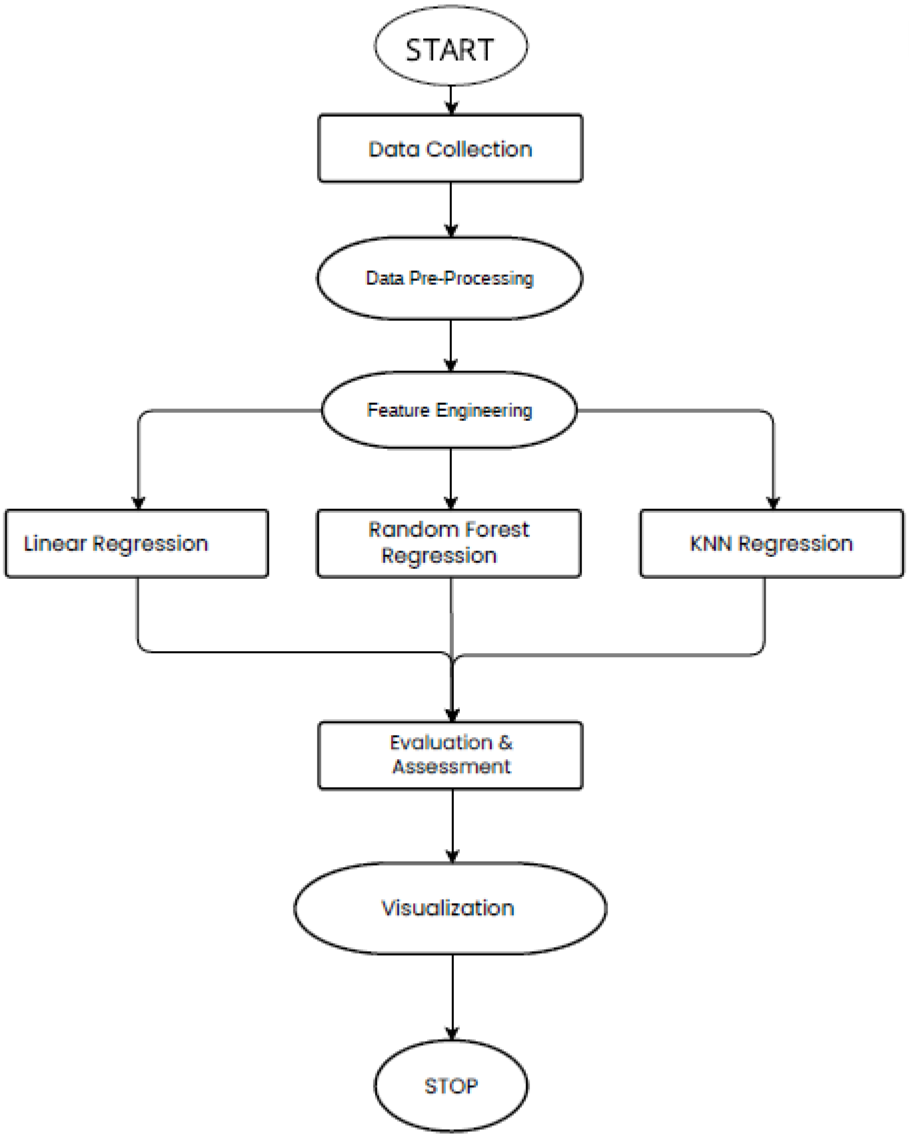


Fig: Methodology of Crop Yield Prediction

6. MACHINE LEARNING ALGORITHMS

- Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable computers to improve their performance over time by learning from experience and feedback.
- In our project, we applied various Regression Algorithms such as Random Forest, Decision Tree, Linear Regression, Polynomial Regression, and Gradient-Boosting.. After the implementation, were able to analyze the accuracy of the algorithms on our data.
- Accuracy was one of the major factors that helped to decide which model has the accurate predictions.

1. Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fit line or hyper plane that minimizes the distance between the predicted values and the actual values.

Pros:

- Linear regression is a simple and easy-to-understand method that requires little technical knowledge.
- It can be used to identify the strength and direction of the relationship between variables.
- It is useful for predicting the value of a dependent variable when the values of theindependent variables are known. .

Cons:

- It assumes that the error terms are normally distributed and have equal variances, which may not be true in some cases.
- It can be sensitive to outliers, which can affect the accuracy of the predictions.
- It can be affected by multi-collinearity, which occurs when two or more independentvariables are highly correlated with each other.

2. Random Forest:

Random forest is a machine learning algorithm that is used for classification, regression, and feature selection tasks. It is an ensemble method that combines multiple decision trees, where each tree is trained on a subset of the training data and a subset of the input features.

Pros:

- It is a highly accurate and powerful machine learning algorithm that can perform well on a wide range of classification and regression tasks.
- It can handle both categorical and continuous input variables, and it can detect and handle interactions between variables.

Cons:

- It may not perform well on small datasets or with rare or unseen classes, which may require more specialized techniques or models.
- It may not be suitable for online or real-time prediction tasks, which require faster and more lightweight models or techniques.

KNN Regressor:

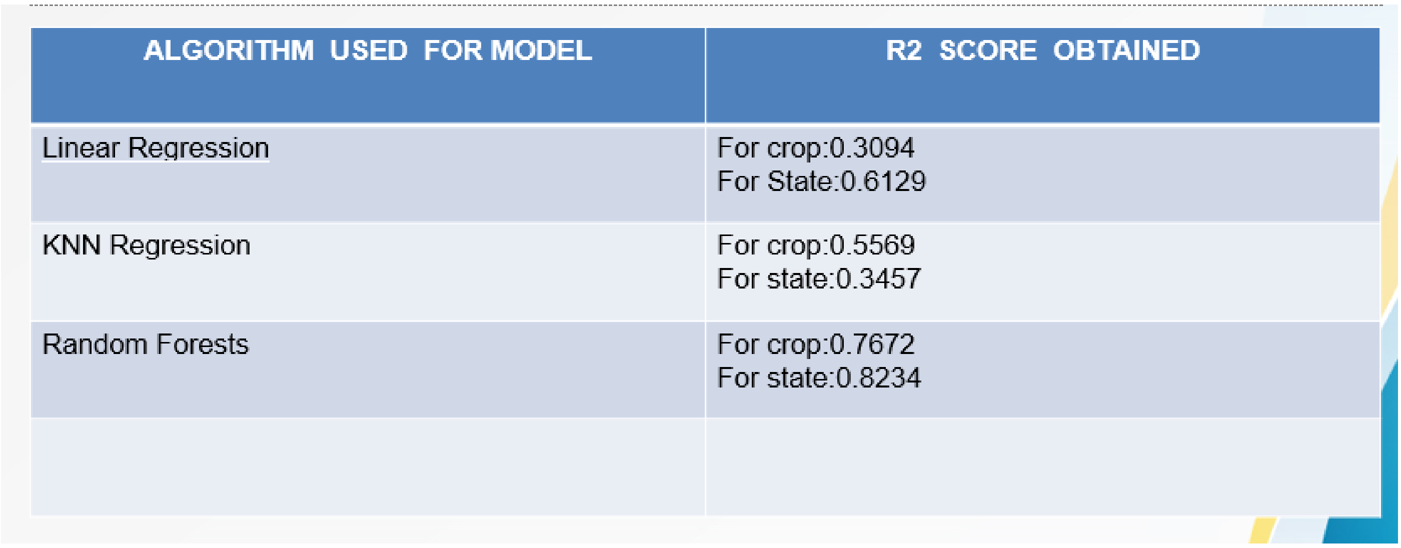
The K-Nearest Neighbors (KNN) algorithm is a simple yet powerful classification algorithm that classifies based on a similarity measure. This supervised ML algorithm can be used for classifications and predictive regression problems1. KNN groups the data into coherent clusters or subsets and classifies the newly inputted data based on its similarity with previously trained data.

Pros:

- It is very simple algorithm to understand and interpret.
- It is very useful for nonlinear data because there is no assumption about data in this algorithm.
- It is a versatile algorithm as we can use it for classification as well as regression.

Cons:

- K-NN slow algorithm: K-NN might be very easy to implement but as dataset grows efficiency or speed of algorithm declines very fast.
- Curse of Dimensionality: KNN works well with small number of input variables but as the numbers of variables grow K-NN algorithm struggles to predict the output of new data point.



ALGORITHM USED FOR MODEL	R2 SCORE OBTAINED
Linear Regression	For crop:0.3094 For State:0.6129
KNN Regression	For crop:0.5569 For state:0.3457
Random Forests	For crop:0.7672 For state:0.8234

Fig. R2 Score of different ML model

7. DATA VISUALIZATION AND REPRESENTATION

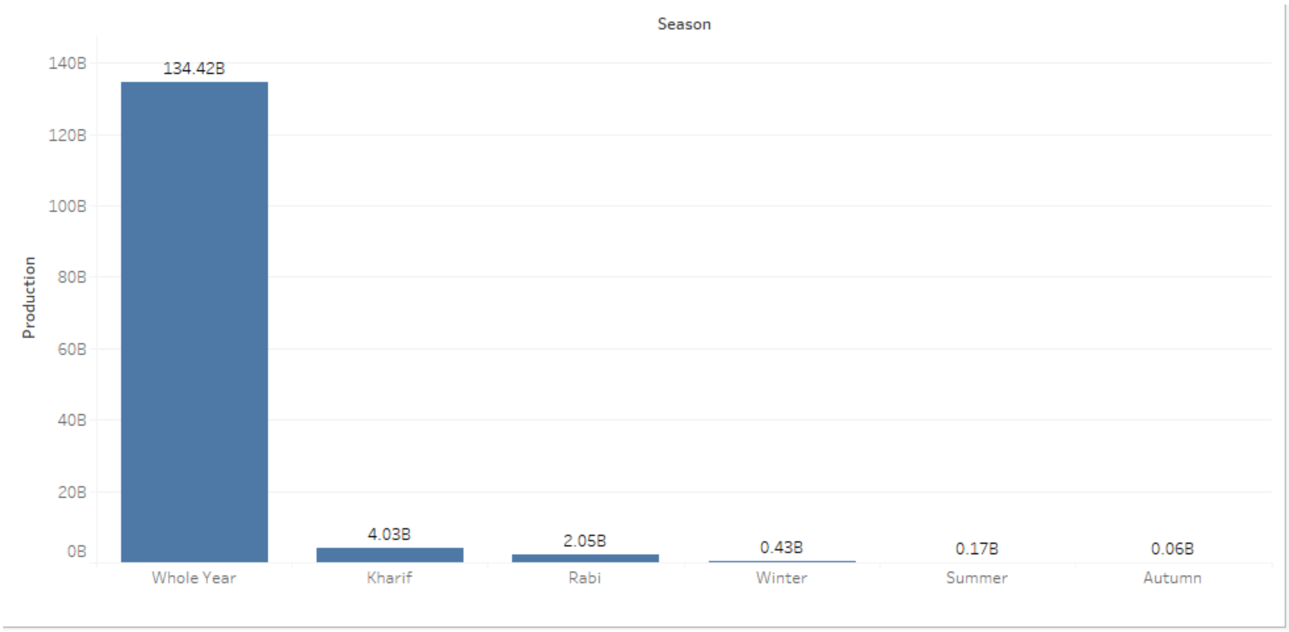


Fig: Production based on Rainfall Seasons (Side-by-Side Bar Chart)

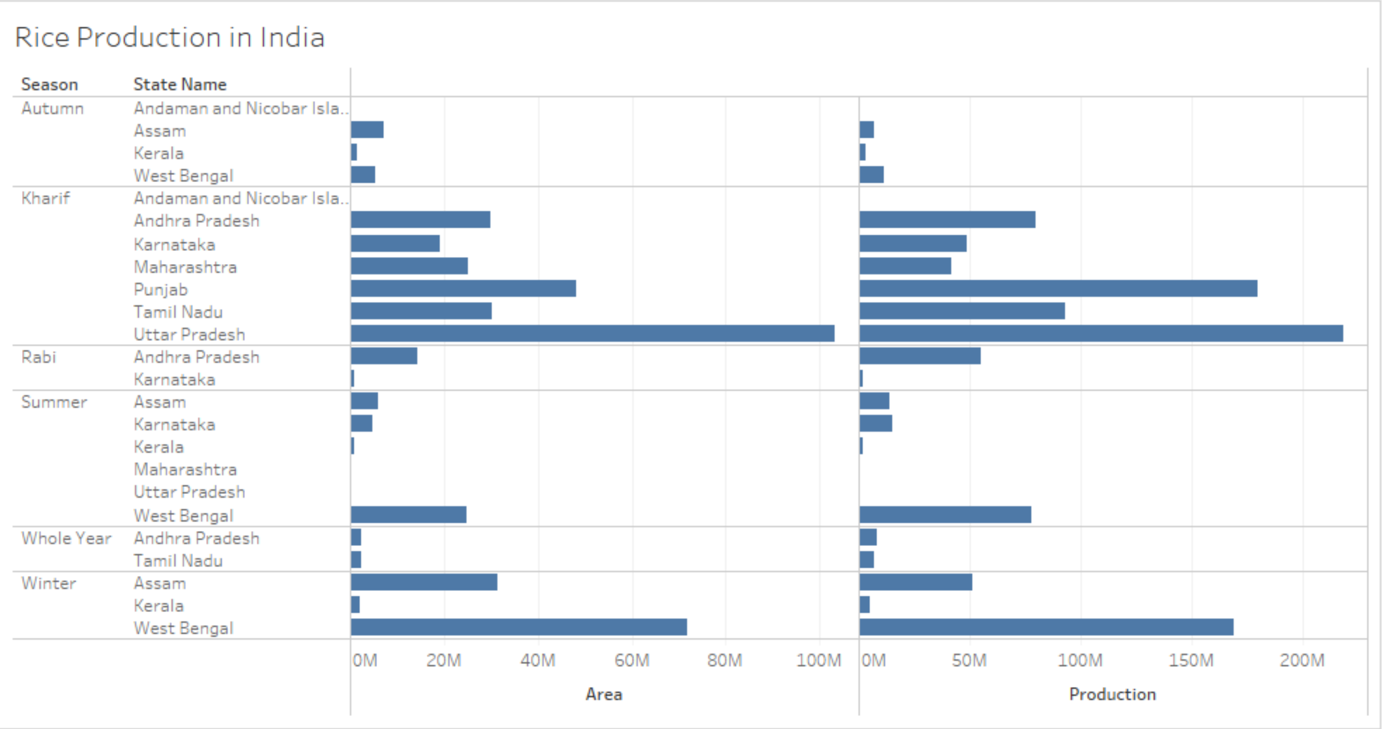


Fig. Rice production in India Season wise

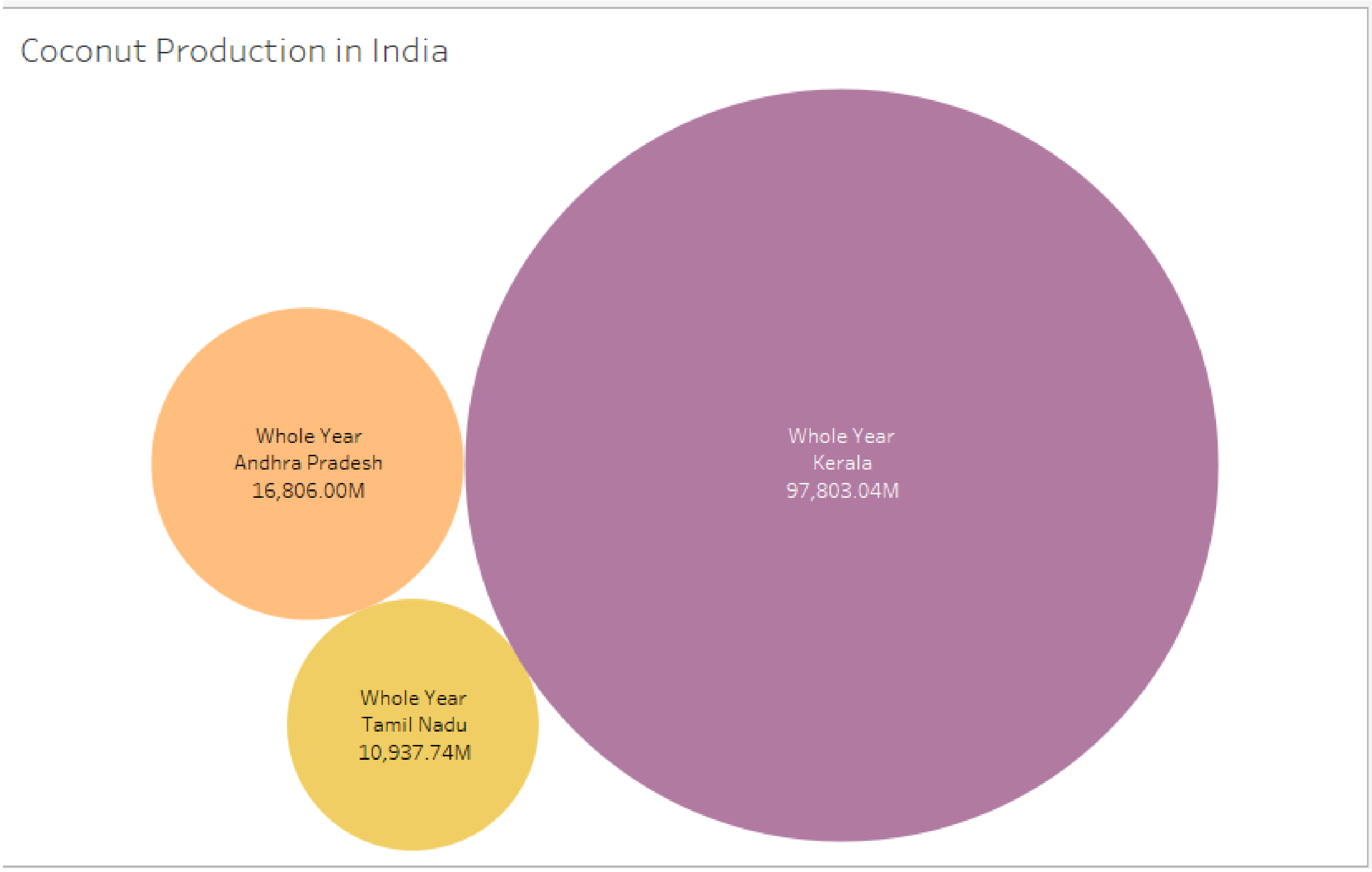


Fig. Coconut Production in India (Bubble Chart)

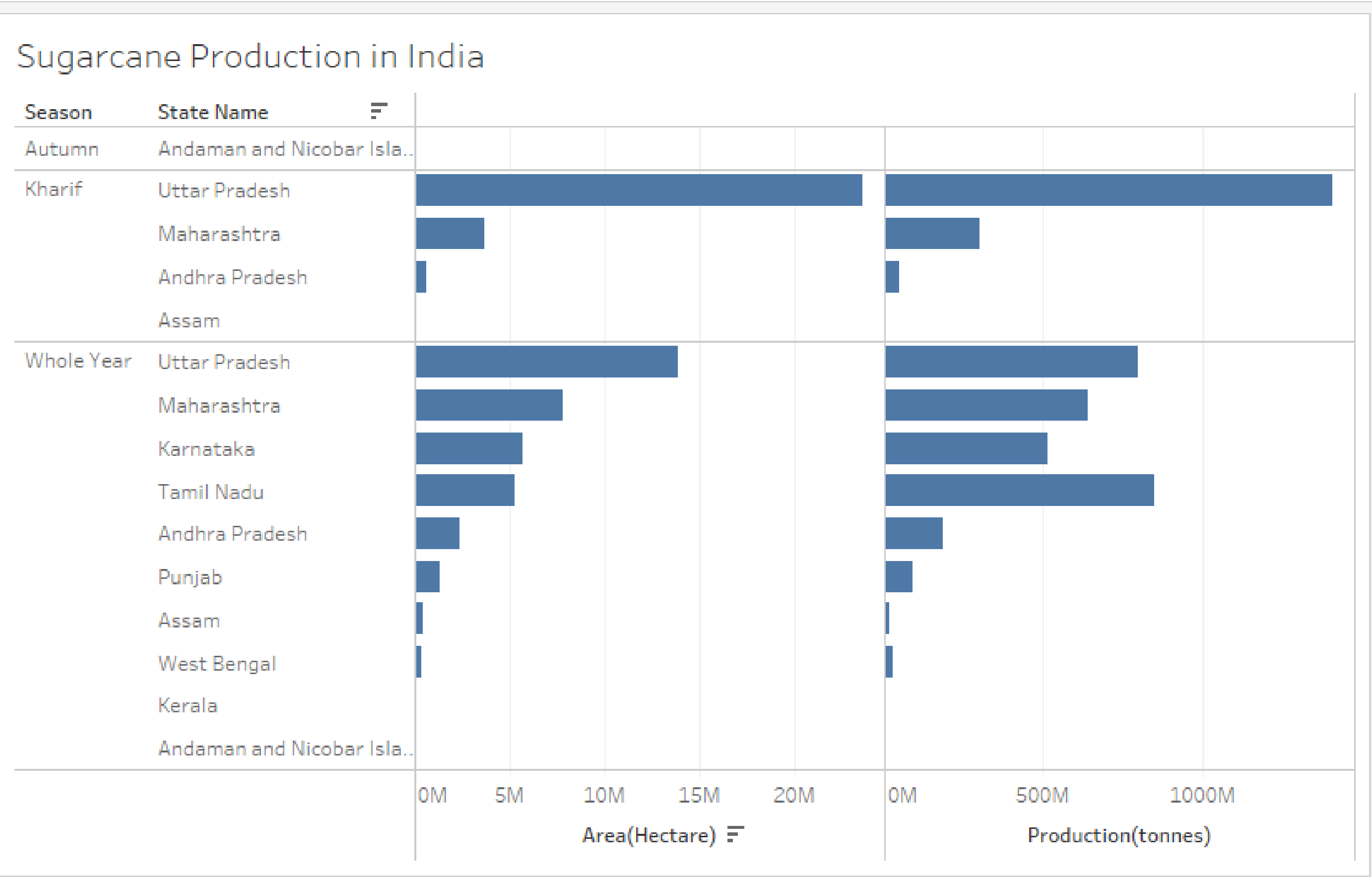


Fig. Sugarcane Production in India

8. CONCLUSION AND FUTURE SCOPE

In conclusion, predicting crop production using machine learning algorithms and datasets containing information on crop production and rainfall can provide valuable insights for farmers, governments, and other stakeholders in the agricultural industry. By leveraging historical data, it is possible to create accurate models that can help improve decision-making and increase efficiency in the agricultural sector.

The insights we get from analysis are as follows:-

1. **It's very inaccurate to predict production for different crops. But it's very precise to predict production for a single crop.**

2. **Insight from Rice Production**
   - Rice production is mostly depends on Season, Area, State (place).

3. **Insight from Coconut Production**
   - Coconut production is directly proportional to area
   - Its production is also gradually increasing over a time of period
   - Production is high in kerala state
   - It does not depends on season

4. **Insights from Sugarcane Production**
   - Sugarcane production is directly proportional to area
   - And the production is high in some state only.

5. **Bar plot of crop production by Seasons**
   - A bar chart effectively visualizes crop production across different seasons, offering insights into crucial seasonal patterns that significantly influence crop yields.

Future work on this study could focus on several areas to improve the accuracy of the predictions even further. Some potential areas for future research include:

1. Experimenting with different combinations of pre-processing methods to achieve better prediction accuracy.
2. Using Hyper-Parameter-Tuning for current existing algorithms for enhanced accuracy.
3. Exploring the challenges and best practices for deploying machine learning models in production environments, including different deployment options such as containerization, server less computing, and API-based deployment.
4. Incorporating additional data sources such as satellite imagery or soil moisture data to improve the accuracy of the predictions.
5. Developing more advanced machine learning algorithms that can better capture the complex relationships between the different variables in the dataset.
6. Conducting more extensive validation studies to assess the performance of the models in different regions and under different conditions.
7. Collaborating with farmers and other stakeholders to ensure that the models are practical and useful for decision-making in the agricultural sector.

Overall, there is significant potential for further research in this area, and continued efforts to improve the accuracy of crop production predictions using machine learning algorithms can have a significant impact on the agricultural industry.

References

1. Crop Production in India

   Crop Production in India | Kaggle

   Rainfall in India

   Rainfall in India | Kaggle

2. Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, J., Lapis, G., & Brown, R. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill.
3. Python documentation: https://docs.python.org/3/
4. "Machine Learning using Python" by Prof. U Dinesh Kumar, IIM Bangalore.
5. . Annina S, Mahima SD, Ramesh B. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 2015 January; I(1): 22-24.