**ABSTRACT**

The Project entitled **“A STUDY ON AGRICULTURE CROP PRODUCTION IN INDIA USING BIG DATA ANALYTICS AND MACHINE LEARNING MODELS”.** Dataset is taken from kaggle repository it contains agriculture crop production data. Crop production data is in structured format we are using classification models under supervised machine learning.

The agriculture plays a dominant role in the growth of the country’s economy. Area and some seasonal changes has become a major problem in the agriculture field. Machine learning is achieving practical and effective solutions for this problem. Here we are predicting the crop production from available historical data, the available historical data like state, district, crop year, area, crop production. This project focus on predicting the crop production based on the available data by using some machine learning algorithms. The crop production prediction will helps to the farmers to predict the crop before cultivating onto the agriculture field.

To predict the crop production accurately we used Support Vector Machine, Logistic Regression, Random Forest, Ada boost, Gradient decent, xgboost these are the most powerful supervised machine learning algorithms.

**CHAPTER I**

**INTRODUCTION**

* 1. **MACHINE LEARNING**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human interaction or assistance and adjust actions accordingly.

Data visualization

Cleaning data to have homogeneity

Gathering data from various sources

Model Building selecting the right ML algorithm

Gathering Insights from the models results

**Figure 1.1: (Machine learning working procedure)**

* 1. **BIG DATA ANALYTICS**

Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information. Such as hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.

Big Data could be found in three forms

a. Structured Data

b. Unstructured Data

c. Semi-structured Data

**STRUCTURED DATA**

Any data that can be stored, accessed and processed in the form of fixed format is termed as a structured data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zeta bytes.

**UNSTRUCTURED DATA**

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

**SEMI-STRUCTURED DATA**

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.

Example of semi-structured data is a data represented in an XML file.

* 1. **DATA ANALYTICS**

Data Analytics is looking hidden patterns in collected data, drawing conclusions based on patterns, proposing new theories and supporting decision making.

MODEL DEVELOPMENT

EXPLORATORY DATA ANALYTICS

MODEL CLEANING

MODEL EVALUTION

**Figure 1.2 (procedure for data analytics)**

* + 1. **Applications of Data Analytics**

The applications of data analytics are as following:

* Retail Solutions
* Traffic Control
* Agriculture Crop Production
* Business Insights
* Finance
* Healthcare
* Sports Performance Etc,.
  1. **AGRICULTURE**

Agriculture is the backbone of every economy. In a country like India, which has ever increasing demand of food due to rising population, advances in agriculture sector are required to meet the needs. From ancient period, agriculture is considered as the main and the foremost culture practiced in India. Ancient people cultivate the crops in their own land and so they have been accommodated to their needs.

Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life. Nowadays, modern people don’t have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food.

* + 1. **DATA ANALYTICS IN AGRICULTURE**

Data analysis in agriculture used to improve the productivity and quality of the crops in agriculture sector. Production of crop prediction is an important agricultural problem. Each and every farmer is always tries to know, how much production will get from his expectation. In the past, production of crop prediction was calculated by analyzing farmer's previous experience on a particular crop. Accurate information about history of crop production is an important thing for making decisions related to agricultural risk management.

Therefore, this project proposes an idea to predict the production of the crop .The farmer will check the production of the crop as per the acre, before cultivating onto the field.

**1.5. PYTHON**

Python is a popular platform used for research and development of production systems. It is a vast language with number of modules, packages and libraries that provides multiple ways of achieving a task.

Python and its libraries like NumPy, SciPy, Scikit-Learn, Matplotlib are used in data science and data analysis. They are also extensively used for creating scalable machine learning algorithms. Python implements popular machine learning techniques such as Classification, Regression, Recommendation, and Clustering.

Python offers ready-made framework for performing data mining tasks on large volumes of data effectively in lesser time. It includes several implementations achieved through algorithms such as linear regression, logistic regression, Naïve Bayes, k-means, K nearest neighbor, and Random Forest.

**WHY PYTHON**

* Extensive collection of libraries
* Libraries are easy referenced
* Fewer lines of code
* Cross platform (Mac,Linux,Windows)
* Large community
* Massive data libraries
* Easy to learn

**PYTHON LIBRARIES**

Scientific computing Libraries:

* Numpy
* Pandas
* Scipy

Visualization libraries:

* Matplotlib
* Seaborn

Algorithmic Libraries:

* Scikit-Learn
* Stats Models

**CHAPTER II**

**MACHINE LEARNING MODELS**

**2.1.** **SUPPORT VECTOR MACHINE**

Support vector machines so called as SVM is a supervised learning algorithm which can used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for dataset analysis as it takes too long to process.SVM is a linear classification that only classify binary problems (i.e. where the Y variable has only categories).

* 1. **LOGISTIC REGRESSION**

Logistic Regression is used when the dependent variable is categorical. Unlike Linear Regression, where we have to predict the continuous value, in logistic regression, we have to predict a categorical variable. Logistic regression is a linear method of classifying the data and it is not to be confused with Linear Regression, as linear classification means is done by a linear separator (a line/hyper plane)

* 1. **K-NEAREST NEIGHBORS**

K-Nearest Neighbour commonly known as KNN is an instance-based learning algorithm and unlike linear or logistic regression where mathematical equations are used to predict the values, KNN is based on instance and doesn’t have a mathematical equation. As there is no mathematical equation, it doesn’t have to presume anything, such as the distribution of the data being normal etc and thus is a non-parametric way of predicting. It is a supervised non-linear method used to solve classification as well as regression problems.

* 1. **ARTIFICIAL NEURAL NETWORKS**

Neural Networks is one of the most advanced methods to solve Regression and Classification problems. The working of Neural Networks is inspired by the human brain which consists of billions of neurons.

* 1. **NAIVE BAYES**

**N**aive Bayes is a modeling technique used for solving classification problems where the Y variable can have more than two classes. When the independent variable is categorical, frequencies are used while for continuous variables and Gaussian density function is used to compute the probabilities. Naive Bayes is based on the Bayesian Theorem.

* 1. **BAGGING**

Bagging is a type of ensemble method where sampling of the data is done in such a way that that problem of overfitting can be addressed. There are various types of Bagging such as Pasting, Bootstrapping, Random Subspaces and Random patches. However, it is the Bootstrapping method which is often considered as the Bagging method where several subsets of data from the training set are chosen randomly with replacement. This helps us in creating an ensemble of models where the results from all such models are averaged to come up with a single result. The most common algorithm used during the bagging process is decision trees. A popular variant of Bagging is Random Forest which is a combination of bootstrapping and random subspaces where the decision trees algorithm is used.

# RANDOM FORESTS

# Random Forest is a variation of the bagging algorithm which is designed specifically for decision trees where it uses a combination of bootstrapping and random subspaces to form subsets of data. The Random forests can be of two types

Random Forest Regressor used for Regression problems and Random Forest Classifier used for Classification Problems.

It is such a common application of bagging that it is considered as a separate ensemble method; however, it is just a method of implementing bagging using decision trees. Random Forest is better than using Bootstrapping with decision trees because when applying Bagging to Decision trees, the main drawback is that if we have a large dataset then often the decision trees learnt from each of the bags (sub-sets) are not very different, often causing the separate decision trees to have same root note. Thus, to diversify the models we have to introduce some variability in the learner which is done by selecting some features at every step of the decision tree training process (i.e. for every model, different features are selected).

# Thus, in random forests, we create subsets by selecting random samples with replacement along with selecting random features without replacement. Thus, each subset can have duplicate samples (observations) however there are randomly selected features which are unique. Separate, highly complex decision trees are built on these subsets and are consolidated to give us a less complex, less overfitted ensemble decision tree which also has optimum accuracy.

# ****ADA BOOST (ADAPTIVE BOOSTING)****

Bagging is a parallel process while adaptive boosting is sequential where the next model is based on the current model.

Thus, the major difference between bagging and boosting is that with bagging we created m number of bags with each having randomly selected observations while in boosting, the observations of the bag is effected by the performance of the model on the previous bag. Here unlike Bagging, we use weak learners (i.e. models such as very shallow decision trees). The predictions of these weak learners are combined using weighted average (for regression problems) and weighted majority voting (for classification problems), unlike bagging where a simple democratic method was used as here the models that show good performance during training are rewarded by giving them higher weights than the model which had higher error rate.

# ****GRADIENT BOOST****

we scale up the complexity with each iteration. The error in each model is computed which is the gradient of the loss function. In the next stage, the new model compensates for the shortcomings of the previous weak model. These shortcomings are the error or gradients (of the loss function). unlike adaptive boosting, the errors are not identified by providing them weights and using them to create new buckets, rather the training data is used (training and testing (model evaluation) is done not by bootstrapping (as done in bagging/random forest where all the data can be used and out of bag error can then be used to evaluate the model) but is done through k-fold cross validation where each of the fold is regarded as test set and their average is used to measure the final performance) to identify the data points that were difficult to predict (i.e. had large residuals) by the initial weak model and are made to correct the error by making the model more complex. Gradient boosting can be used for Regression as well as Classification problems.

# EXTRA GRADIENT BOOST (XGBOOSTING)

Extra Gradient boosting (XGBoosting) is an advanced form of gradient boosting. It performs 3 types of gradient boosting – Standard Gradient Boosting (discussed above), Stochastic Gradient Boosting (sub-sampling is done at the row and column level) and Regularization Gradient boosting (L1 and L2 Regularization are performed).

The advantages of XGBoost over standard Gradient Boosting is that, it is a parallel learning technique and saves time when compared to Gradient Boosting (It is parallelized by using all of the CPU cores during the training process). It also helps in handling overfitting as it uses both the regularization techniques ( L1 and L2 regularization) and is more effective with missing values (Sparse awareness). Cross-Validation is done at each iteration generally by using k-fold validation (e.g. k=10) where each of the ten folds is regarded as the test set and their average is used to measure the final performance. It is important to note that XGBoosting uses a lot of hyper parameters which are to be optimized especially the number of base models that are to be created as after a point it will end up creating a fairly complex overfitted model.

Generally, XGBoost is used as a Scalable Parallel Tree Boosting Ensemble method and is among the most famous method for performing classification and regression.

Boosting is another great ensemble method which is a variant of Bagging only where the output of previous model effect the next model.

Thus, here the function is learned sequentially. The basic difference between bagging and boosting is that bagging is a parallel ensemble where each model is built independently while boosting is a sequential ensemble method where we add new models to do well where the previous models lack (however, boosting can be done parallelly by using extra gradient boosting). Also, both address the bias-variance problem with bagging having the aim of decreasing variance (very complex base models are used and thus bagging is suitable for reducing the complexity/overfitting of complex models) while the aim of boosting is to decrease bias (very simple base models are used and thus boosting is used to systematically increase the complexity of the model).

**CHAPTER III**

**OBJECTIVE AND METHODOLOGY**

* 1. **OBJECTIVE**

The objective of this project is to analyze the agriculture crop production dataset using supervised classification models.

The specific objective of this project is

* Perform preprocessing and EDA
* Analyze the dataset
* Perform classification models
* Compare the models crop production in agriculture crop production dataset with accuracy score
  1. **METHODOLOGY**

There are five important steps to analyze the dataset. Here is the diagrammatic representation of the methodology. The methodology for analyzing the objective is given below.

Agriculture crop production dataset

EDA

Basic preprocessing

Modeling

Output

**Figure 3 (Methodology)**

* 1. **DATSET DESCRIPTIO N**

The Agriculture crop production in India data set has seven variables with 246090 instances in the years 1997-2015.

**The attributes are**

State Name : States in India

District Name : Districts in India

Crop Year : Year of the crop

Season : Seasons in agriculture

Crop : crops in agriculture

Area : Area occupied for agriculture crop

Production : production for crop

**Variable Data types**

State Name : object

District Name : object

Crop Year : int64

Season : object

Crop : object

Area : float64

Production : object

* + 1. **DATA PREPROCESSING**

**Data preprocessing** is a technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and lacking in certain behaviors or trends, and is likely to contain many errors. **Data preprocessing** is a proven method of resolving such issues.

STEPS IN DATA PREPROCESSING

**Step 1:** Import the libraries

**Step 2:** Import the data-set

**Step 3:** Check out the missing values

**Step 4:** See the Categorical Values

**Step 5:** Splitting the data-set into Training and Test Set

**Step 6:** Feature Scaling

* + 1. **EXPLORATORY DATA ANALYSIS**

Exploratory data analysis (EDA) is an approach to analyzing data set to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Turkey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA) which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed

Typical graphical techniques used in EDA are

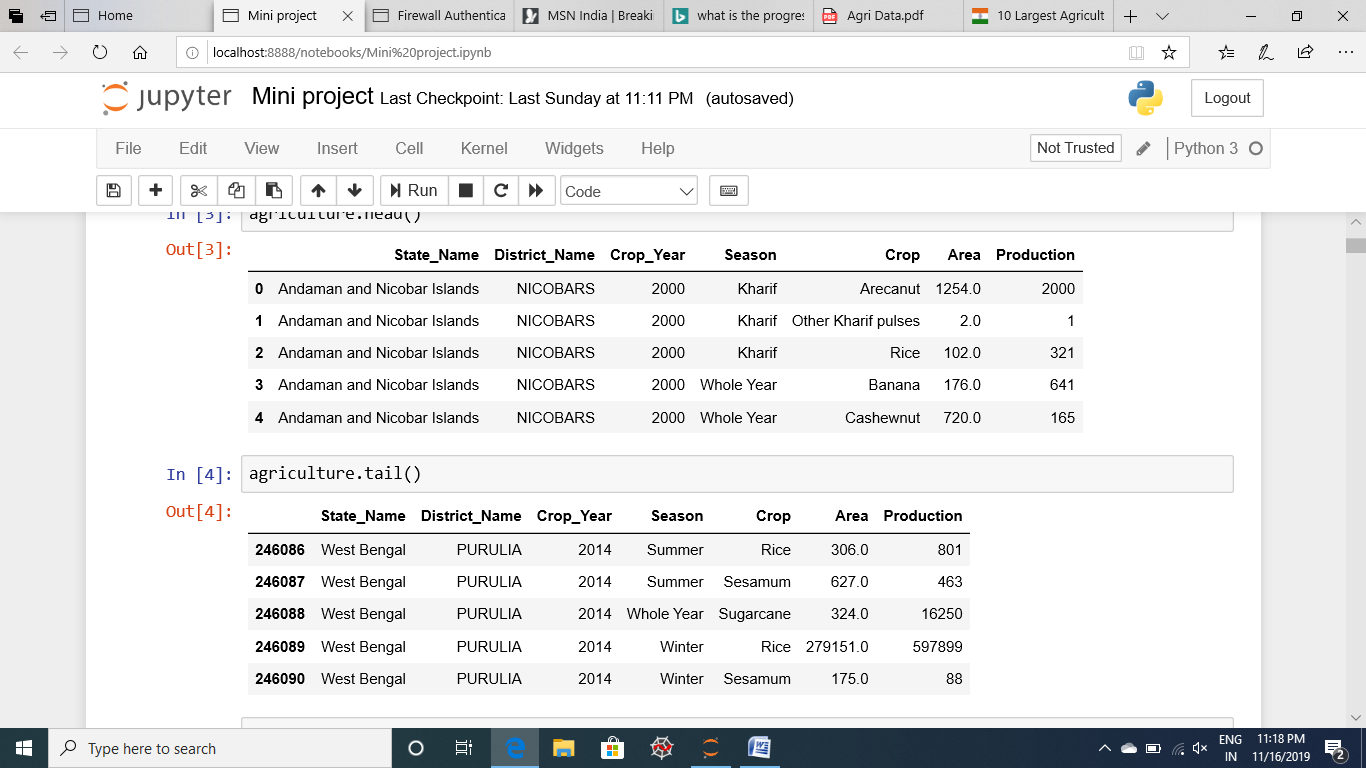
* HISTOGRAM
* BOXPLOT
* SCATTER PLOT
* PIE CHART
* BAR GRAPH etc.,
  + 1. **MODELING**

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data. once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about those data.

**CHAPTER IV**

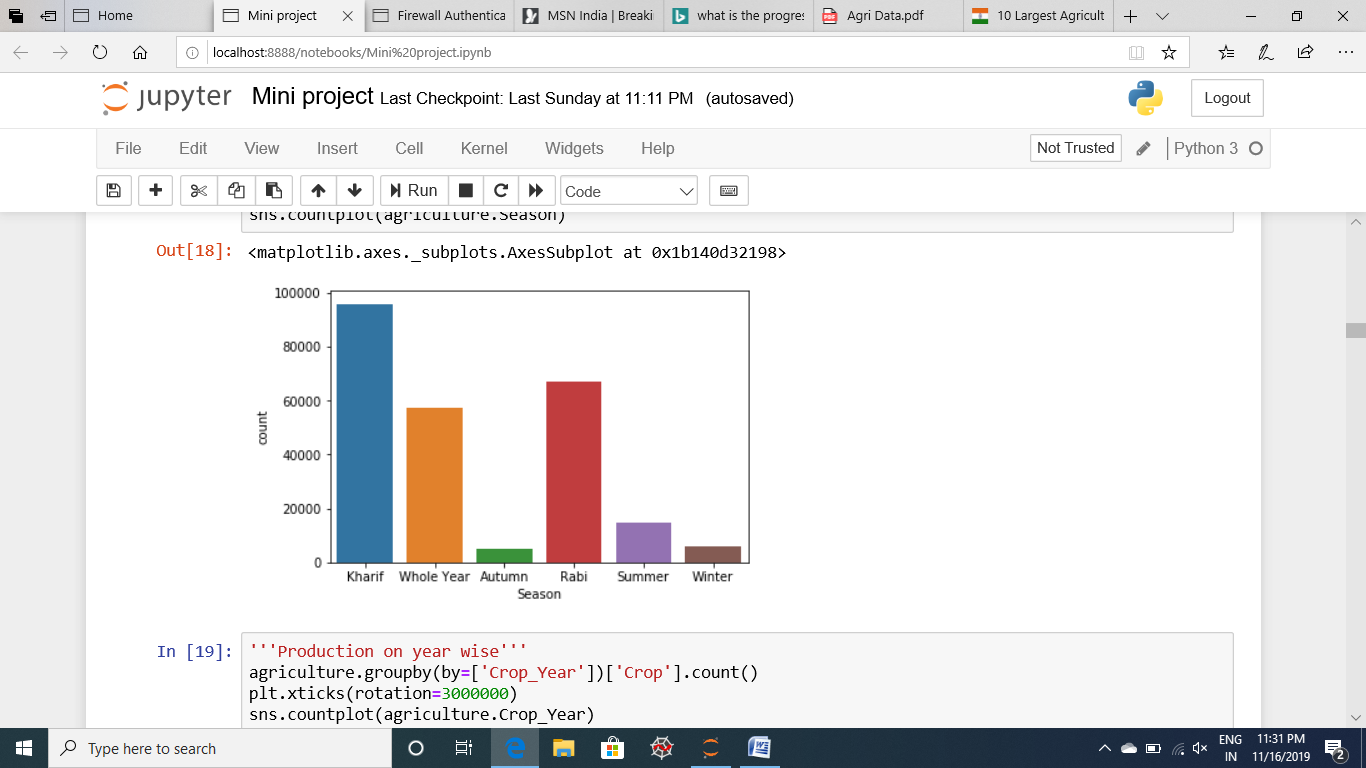
1. **RESULT**

The following picture showing the some data set instances and variables.



**Figure 4.1**

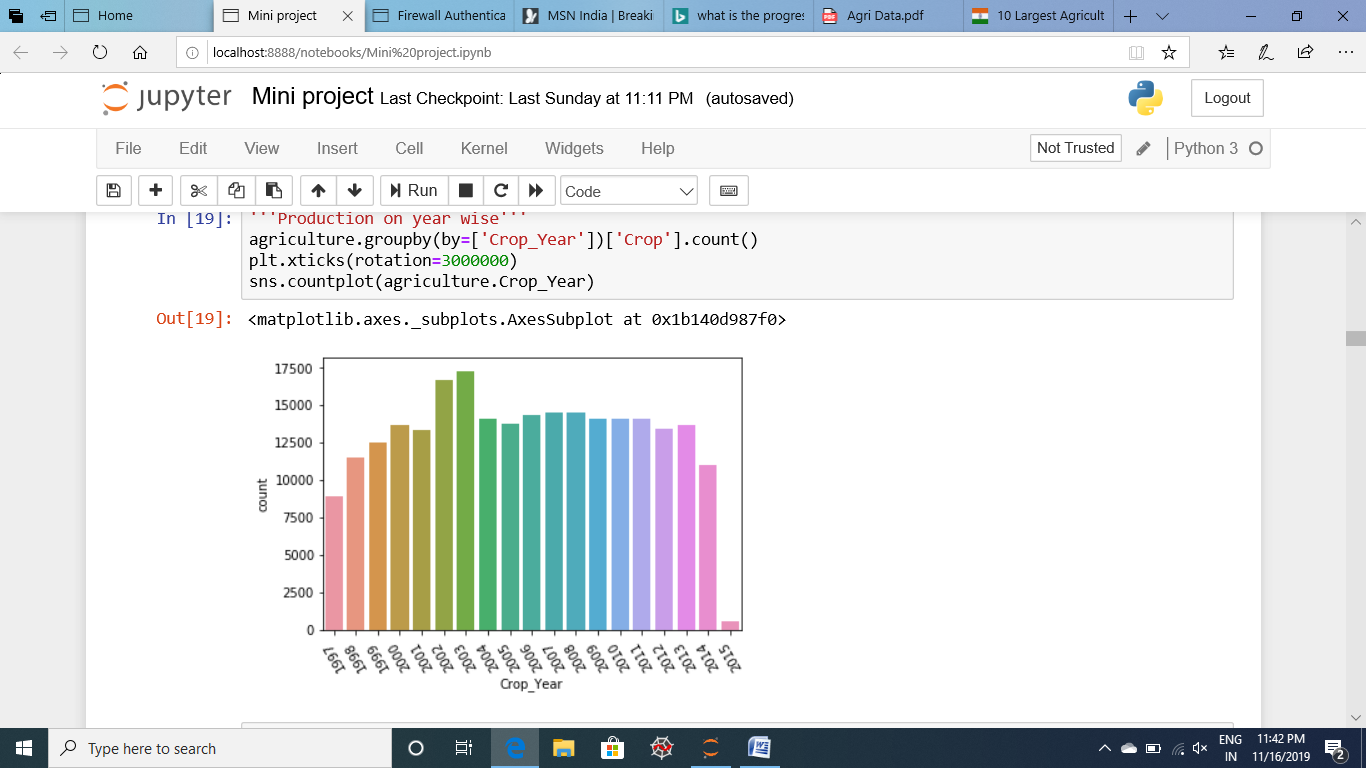
* 1. **VISUALIZATION FOR INDIA’S AGRICULTURE CROP PRODUCTION**



**Figure 4.2 : Season wise crop production**

**Insights:**

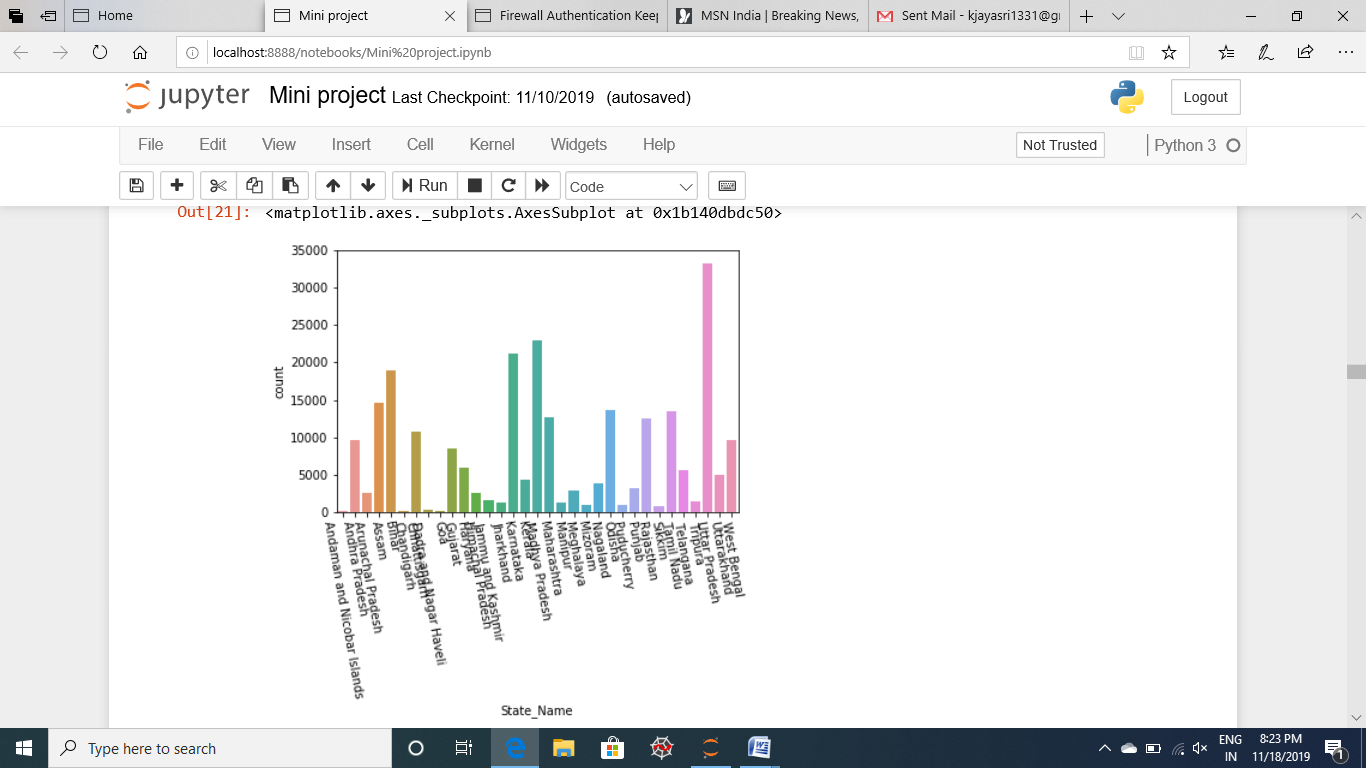
* From the graph we can clearly see that, In agriculture the Kharif season is showing more height in India because most of the crops are cultivating in Kharif season so this is showing more height.
* After Kharif season Rabi season is height season because most of the pulses are producing in Rabi Season.
* The remaining seasons are very less by comparing to the two seasons.

****

**Figure 4.3 Year wise crop production**

**Insights:**

* When we are comparing the all the years of agriculture production in India maximum in every year the crop production is maximum
* But when we are seeing the 2002 and 2003 the agriculture crop production in India is more comparing to other years.

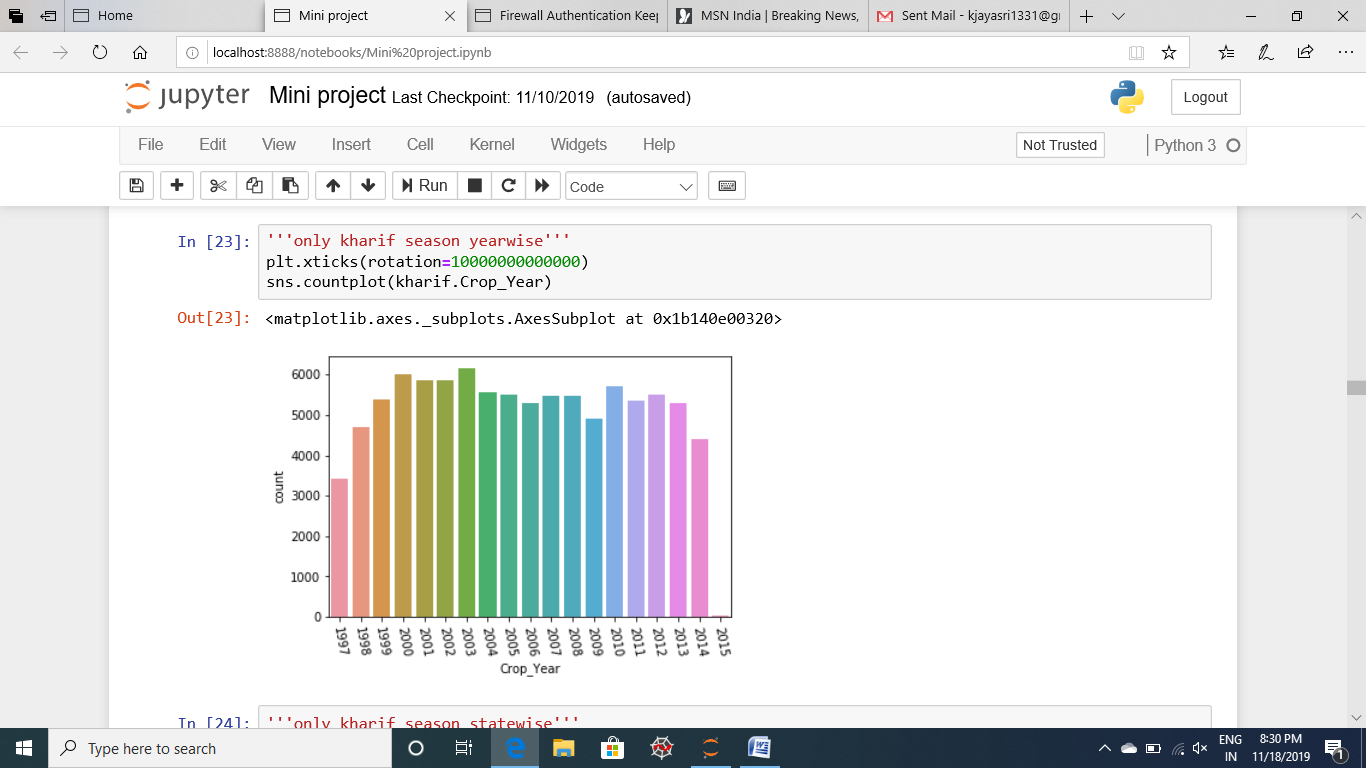
****

**Figure 4.4 State wise crop production**

**Insights:**

When we are comparing to year wise and state wise in India. Years are gradually increasing and decreasing because agriculture in India sometimes have more production and sometimes it have less production. When we are comparing to state wise utter Pradesh is the highest production in India and next Madhya Pradesh is the highest production in India

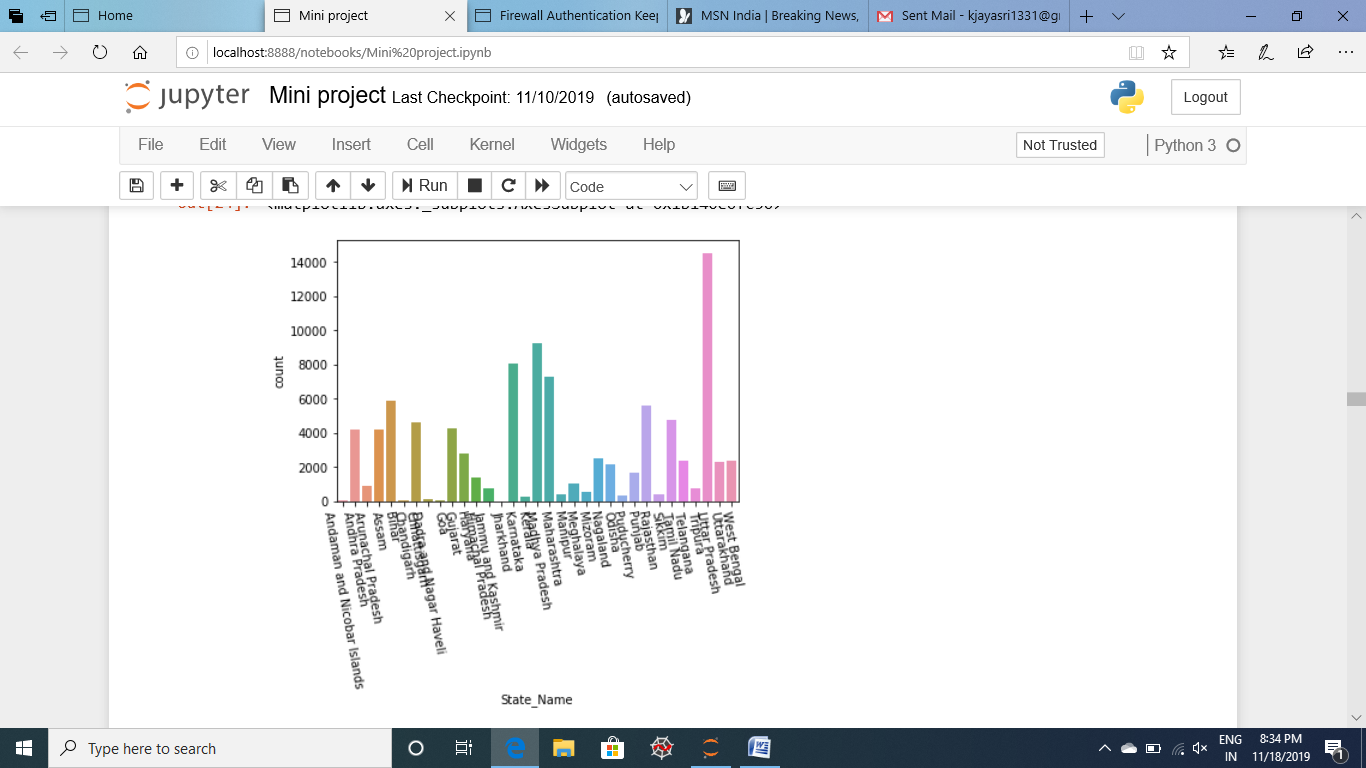
**SUB SETTING ONLY KHARIF SEASON**

****

**Figure 4.5 Year wise kharif season production**

**Insights:**

* When we are seeing the year wise kharif season in India year the agriculture crop production is gradually increasing and gradually decreasing.
* Maximum in every the crop production is highest.
* Comparing to all years the 2003 year is highest crop production in kharif season

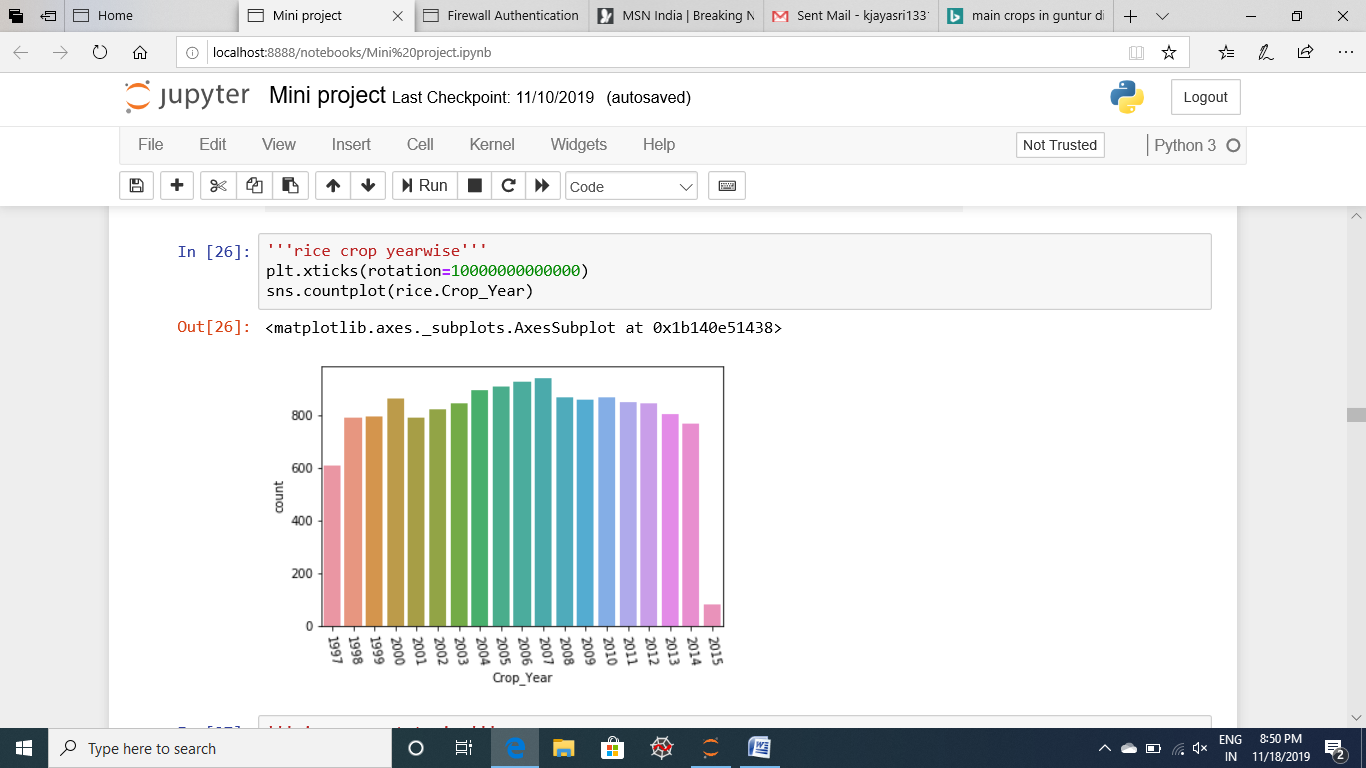
****

**Figure4.6 State wise kharif season**

**Insights:**

When we are seeing the only Kharif season it similar to total India production because maximum most of the crops in India are cultivating in kharif season only. So, that’s why kharif season and Agriculture production in India are similar.

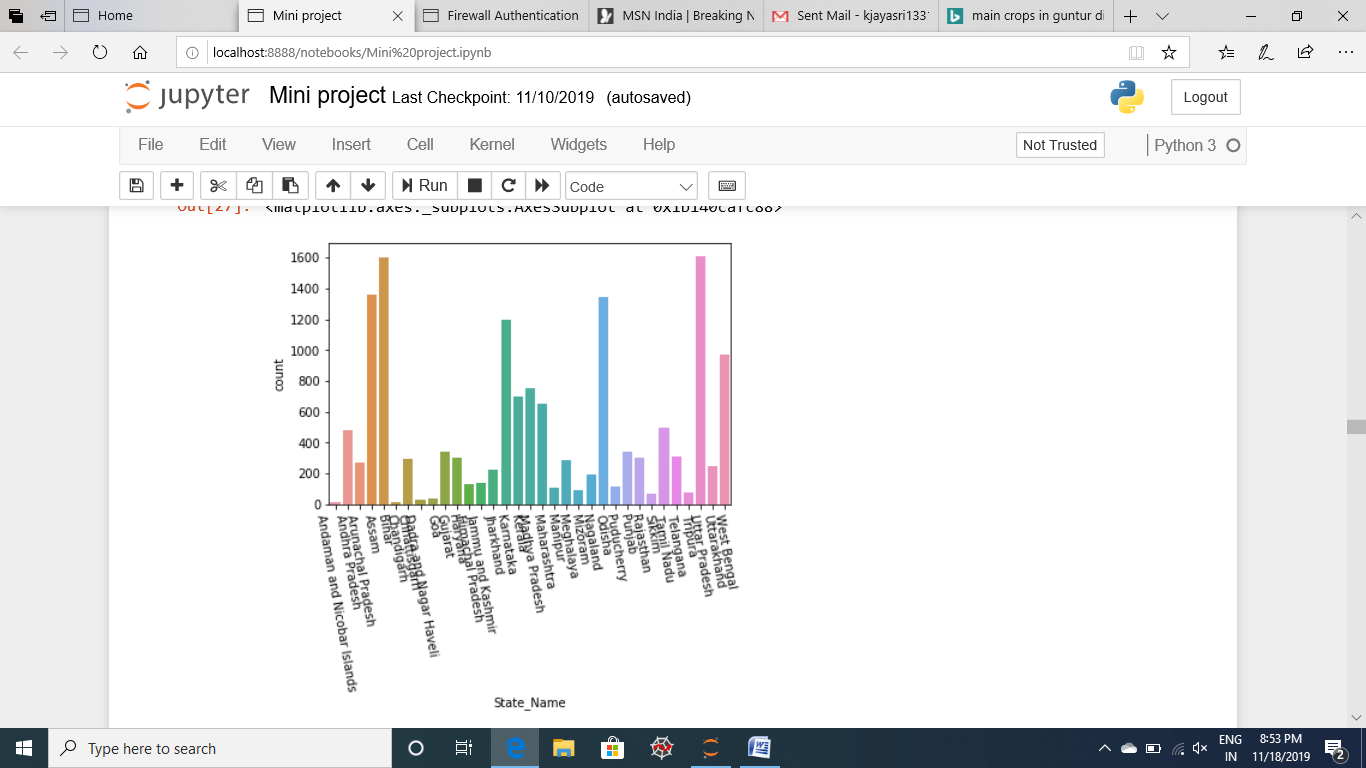
**RICE PRODUCTION IN INDIA**

****

**Figure 4.7 Year wise rice crop production**

**Insights:**

* When we are seeing the year wise kharif season rice crop in India year the agriculture crop production is gradually increasing and gradually decreasing.
* Maximum in every the crop production is highest
* Comparing to all years every year rice production is increasing only because rice is very important in our daily life.

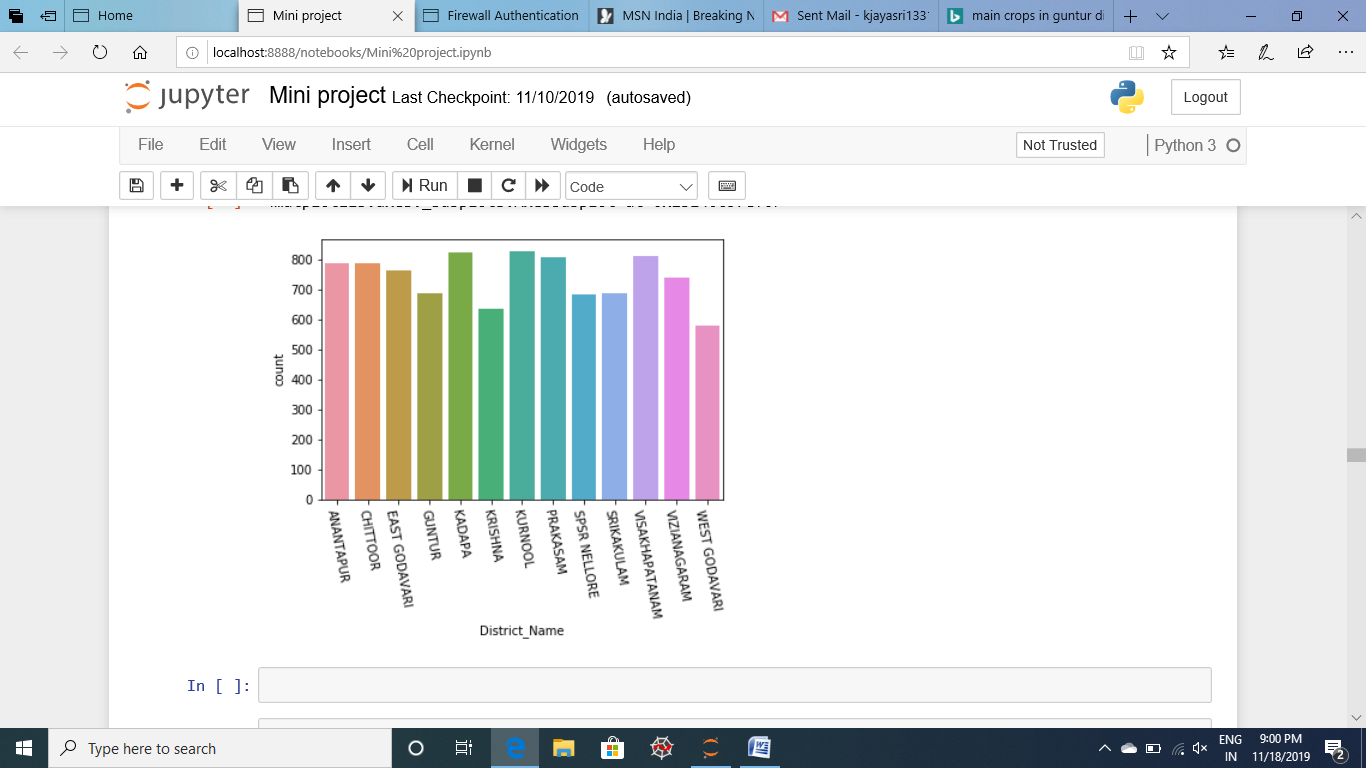
****

**Figure4.8 State wise rice crop production**

**Insights:**

Maximum every state is producing rice crop in agriculture, when we are seeing the statistics for agriculture rice crop the state Uttar Pradesh is giving more rice production in India.

**SUB SETTING ANDHRA PRADESH STATE**

****

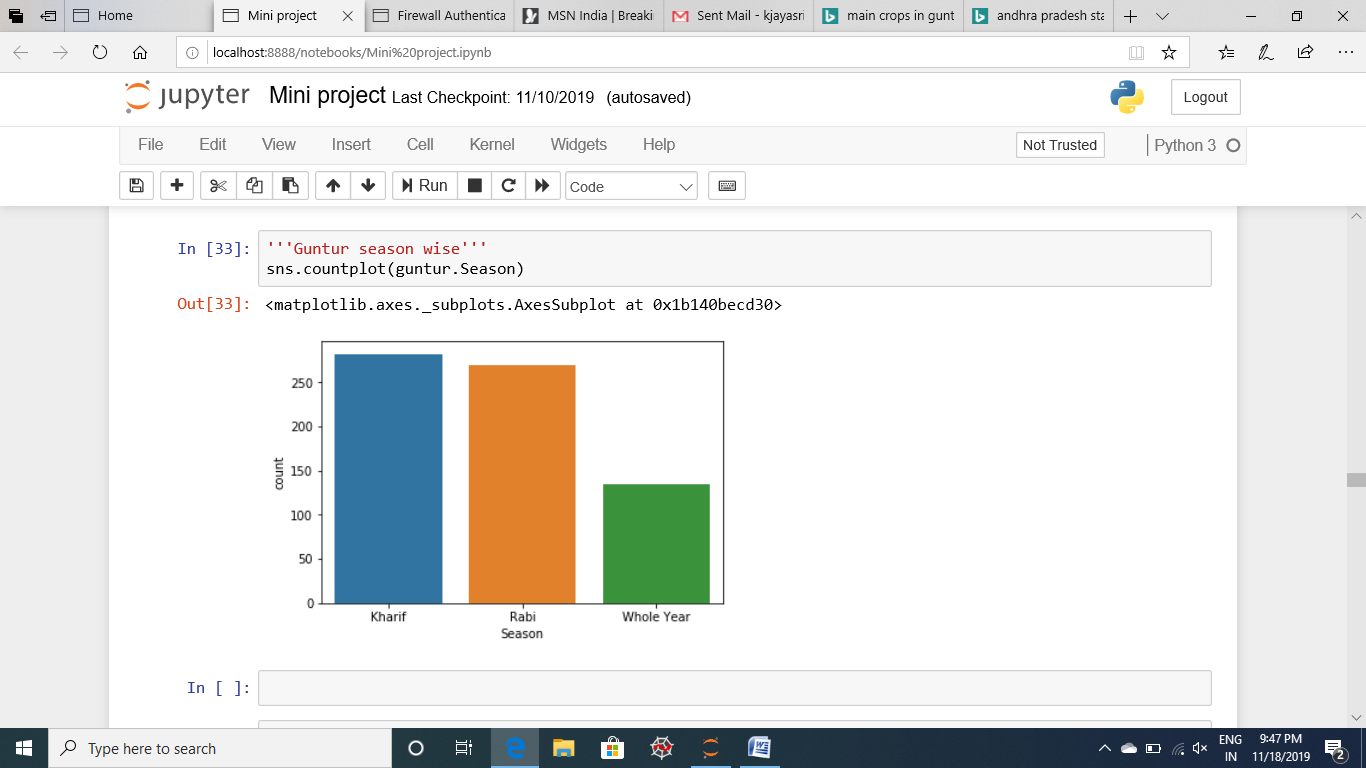
**Figure 4.9 Andhra Pradesh district wise crop production**

Before dividing the states of Andhra Pradesh and Telangana the total districts in Andhra Pradesh were 29 districts. After dividing the states like Andhra Pradesh and Telangana the count of Andhra Pradesh districts was 13 they are anantapur, chittor, east Godavari, Guntur, Kadapa, Krishna, Kurnool, Prakasam, SPSR Nellore, Visakhapatnam, vizianagaram, west Godavari. The count of Telangana districts are 16.

In Andhra Pradesh Agriculture plays a important role in the livelihoods of people as 63% of the population in Andhra Pradesh live in rural areas and depend on Agriculture and related livelihood opportunities.

When we are seeing the every district in Andhra Pradesh the maximum crop productions are producing each every district. When we are sub setting the every state we can clearly know the crop production in every state. Here we are sub setting the only Guntur district.

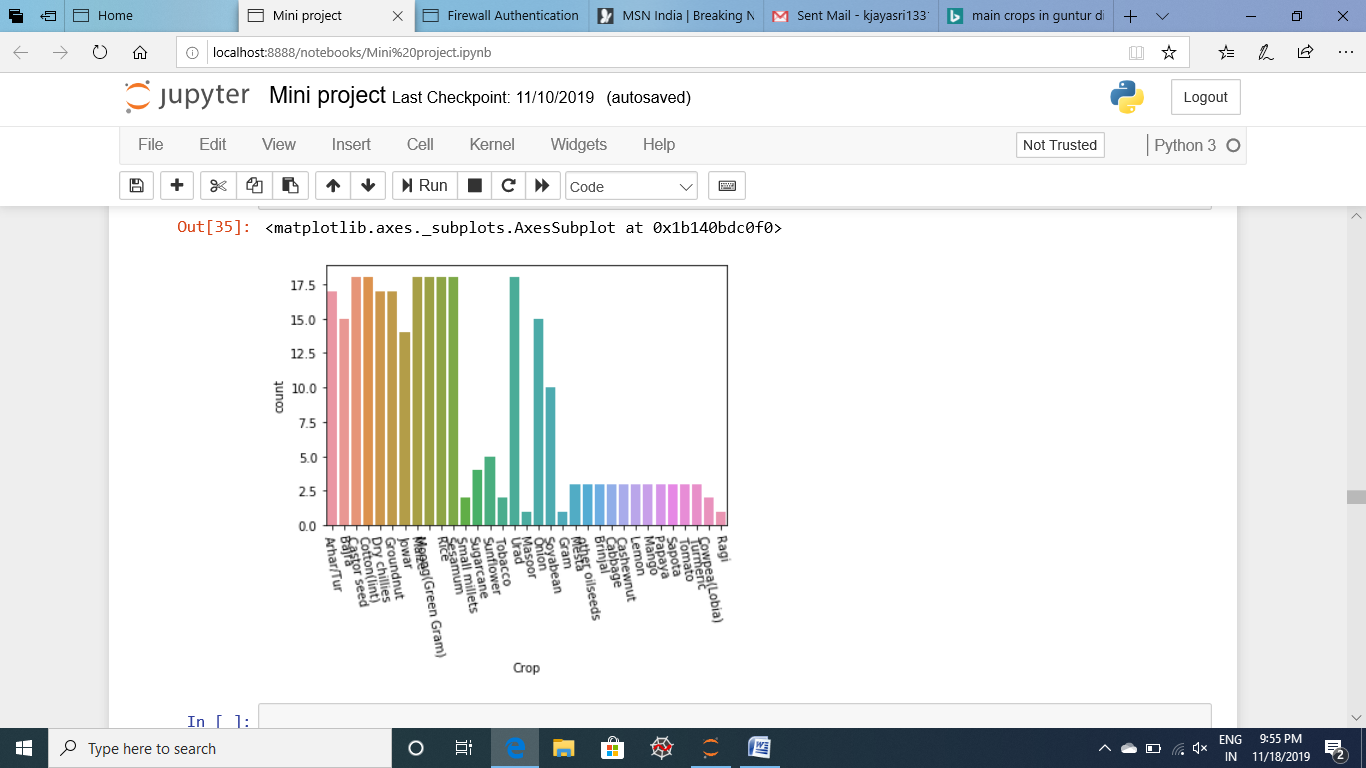
**SUB SETTING ANDHRA PRADESH STATE GUNTUR DISTRICT**

****

**Figure 4.10 Andhra Pradesh Guntur district season wise productions**

**Insights:**

* When we are seeing the seasons in Andhra Pradesh both Kharif and Rabi seasons are giving the similar crop productions
* The kharif season crops are Rice, Maize, cotton, groundnut, Jowar, bajra etc.,
* The Rabi season crops are Wheat, peas, gram, oilseeds, barley etc.,
* Kharif crops can be describes as the crops which are sown with beginning of the rainy season.
* Rabi crops are the crops that are sown after the end of the monsoon, i.e., during the winter season

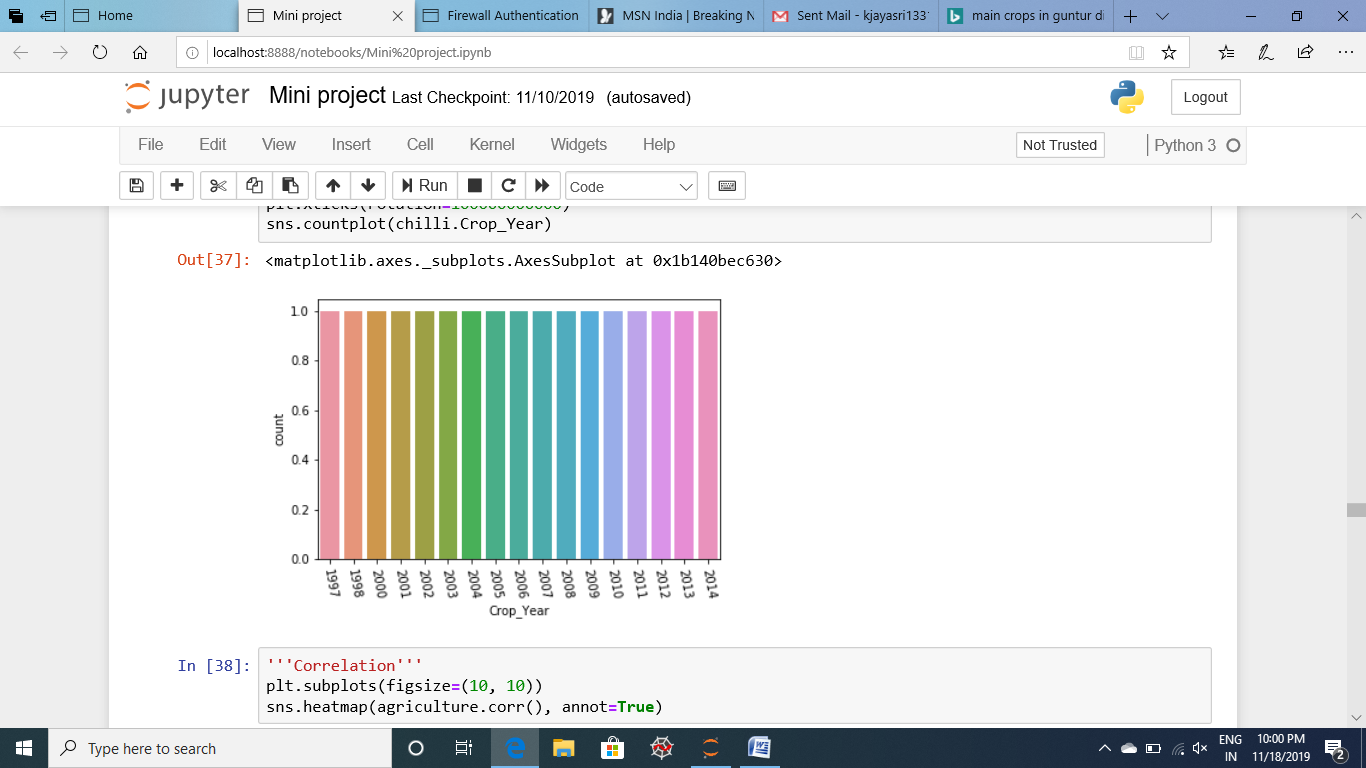
****

**Figure 4.11 Guntur district kharif season crop production**

**Insights:**

* When we are seeing the every crop in Kharif season most of the crops are giving maximum count
* The kharif season crops are Rice, Maize, cotton, groundnut, Jowar, bajra etc.,
* Kharif crops can be describes as the crops which are sown with beginning of the rainy season.
* It requires huge amount of water and hot weather to grow
* For these crops it requires shorter day length.

**SUB SETTING GUNTUR DISTRICT KHARIF SEASON CHILI CROP**

****

**Figure 4.12 Guntur district year wise chili crop production**

Here first we sub setting the Andhra Pradesh Guntur district and next we are sub setting Guntur seasons.

In this season we are seeing the three crop seasons in Guntur they are

* Kharif
* Rabi and
* Whole year

After that we are sub setting the only kharif season because most of the crops are producing in kharif season only comparing to other seasons and the main important point in Guntur district is production of chilies. Sometimes people will call the name of Guntur district is referred to as the **Land of Chilies**. The crop chilies are producing in kharif season so that’s why we are sub setting the only chili crop in Guntur. When we are seeing the year wise production of chili crop all the years are producing equal chili crop production in Guntur district.

* 1. **MODEL PERFORMANCE**

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | ALGORITHM USESD | CODE AND ACCURACY |  |
| 1 | SUPPORT VECTOR MACHINE |  |
| 2 | LOGISTIC REGRESSION |  |
| 3 | K NEAREST NEIGBOURS |  |
| 4 | ARTIFICIAL NEURAL NETWORKS |  |
| 5 | NAÏVE BAYES |  |
| 6 | BAGGING |  |
| 7 | RANDOM FOREST |  |

**OPTIMIZATION ALGORITHMS**

|  |  |  |
| --- | --- | --- |
| 8 | **ADAPTIVE BOOSTING**  **(ADA BOOST)** |  |
| 9 | GRADIENT BOOSTING |  |
| 10 | EXTRA GRADIENT BOOSTING  (XGBOOST) |  |

* 1. **AGRICULTURE FACTS**
* India is the largest producer of Millet in the world and second largest producer of Wheat and Rice/Paddy.
* India is the largest producer of Chickpea and Pulses in the world.
* India is the largest producer of Ginger and Okra in the world and second for Dry Bean, Onion, Cabbage and other breccias.
* India is third largest producer of Cauliflowers and Broccoli and Lettuce and chicory.
* India is the largest producer of Wood fuel and forest products.
* India is the largest producer of Buffalo Milk and Cow Milk in the world.
* Andhra Pradesh is the largest producer of Tobacco in India, followed by Assam, Bihar and Chhattisgarh.
* Andhra Pradesh Guntur is the largest producer of chili.
* Maharashtra is the largest producer of Sugar in India as well as Cashew Nuts, grapes and onion.
* Tamil Nadu is the largest producer of cassava in India as well as largest producer of bananas, followed by Gujarat and Maharashtra.

**CHAPTER V**

**CONCLUSION AND FUTURE ENHANCEMENT**

**5.1. CONCLUSION**

The study deals with prediction of Agriculture crop production in India Agriculture dataset. Agriculture crop production in India dataset is structured data using prediction techniques to get the information from the Agriculture dataset.

The Results shows that we can predict an accurate crop production by using Extra Gradient Boosting algorithm. Extra Gradient Boosting algorithm achieves the largest number of crop production compared to other models. It is suitable for massive crop production prediction in agriculture planning. The accurate score is 42%, the best models for the crop prediction in agriculture is Gradient boost and Extra gradient boost.

This makes the farmers to take the right decision in selecting the right crop for farming such that the agricultural sector will be developed by innovative ideas.

**5.2. SCOPE FOR FUTURE ENHANCEMENT**

Applying Machine learning technology in the field of agriculture, farmers can determine the best crop choice. By comparing the information about crop types in certain area and the probability of production data and year by year outcomes farmers can make decisions on how to maximize crop returns. Machine learning technologies and Prescriptive analytics combined together helps greatly in the farming industry by helping the farmers not only in analyzing the reasons behind problems but also in providing an optimal solution for them to improvise their yield.

**REFERENCE**

* <https://www.kaggle.com/prasadkevin/crop-prediction>
* <https://www.bing.com/search?q=agriculture&form=EDGTCT&qs=PF&cvid=4d1aea2121c049ccbe9fd88680716f8e&refig=3eb567552c1b4c56a8c02f75fc49e879&cc=US&setlang=en-US&plvar=0>
* <https://www.mapsofindia.com/andhra-pradesh/economy/agriculture.html>
* <https://guntur.ap.gov.in/>
* <https://learnnaturalfarming.com/cropping-seasons-of-india-kharif-rabi-and-zaid/>
* <https://www.fb.org/newsroom/fast-facts>
* [www.arthapedia.in/index.php?title=Cropping\_seasons\_of\_India-\_Kharif\_%26\_Rabi](http://www.arthapedia.in/index.php?title=Cropping_seasons_of_India-_Kharif_%26_Rabi)
* <https://www.quora.com/How-is-machine-learning-used-in-agriculture>