CHAPTER-1

INTRODUCTION

1.1 Introduction:

Python is a high-level programming language designed to be easy to read and simple to implement. Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc). Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick. Python can be treated in a procedural way, an object-orientated way or a functional way.

Artificial Intelligence is an approach to make a computer, a robot, or a product to think how smart human think. AI is a study of how human brain think, learn, decide and work, when it tries to solve problems. And finally this study outputs intelligent software systems. The aim of AI is to improve computer functions which are related to human knowledge, for example, reasoning, learning, and problem-solving.

The intelligence is intangible. It is composed of

- Reasoning
- Learning
- Problem Solving
- Perception
- Linguistic Intelligence

Agriculture is one of the most important occupation practiced in our country. It is the broadest economic sector and plays an important role in overall development of the country. About 60 % of the land in the country is used for agriculture in order to suffice the needs of

1.2 billion people. Thus, modernization of agriculture is very important and thus will lead the farmers of our country towards profit.

1.2 Objective of Research

1. To obtain estimates of aggregate physical production functions for the yields of various crops in specified states, considering various technological factors and a newly developed metrological weather index as inputs.
2. To obtain stochastic yield projections, based upon the estimated production functions and projected inputs with weather as a stochastic input.
3. To derive simple decision models to demonstrate the usefulness of the stochastic yield projections in meeting specified agricultural policy goals.

1.3 Problem Statement:

Every time manual prediction doesn’t reach the accuracy so in this proposed system we apply machine learning and prediction algorithm like Classification to identify the pattern among data and then process it as per input conditions. The system to be designed will be enable farmers with a convenient tool to analyze and select crop to optimize the yield of a particular crop which has potential to increase crop yields and result in greater profits for the farmer.

CHAPTER-2

REVIEW OF LITERATURE

There are different forecasting methodologies developed and evaluated by the researches all over the world in the field of agriculture. Many researchers have been contributed their previous knowledge towards data mining in agriculture. There are many simulations models available for crop productivity predictions. As it depends on economical and environmental parameters so we can not apply these existing models or methods to any other area.

Here we predict the suitability of a crop for a particular climatic conditions and the possibilities of improving the crop quality by using weather dataset that crop yield majorly depends on. Here we used classification and prediction algorithm that is DECISION TREE ALGORITHM which is used take a major decision by predicting the best crop which gives high quality of yield on the basis of that particular climatic conditions on that particular area. here we considered two crops Wheat and Rice which are two major crops cultivated in especially Andhra, and different climatic conditions which are suitable for cultivating Rice and Wheat and predicting the crop as best crop by predicting their probability of getting higher yield and quality.

CHAPTER-3

DATA COLLECTION

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

Data are collected for different weather conditions, temperature, humidity, rainfall and future will be predicted by using machine learning algorithm. Though previous monitoring techniques gathers the crop conditions properly, prediction results have not yet been optimized. First of all, researchers do not have clear idea about crop condition and crop monitoring methods. They should know how to monitor crop condition on different circumstances. So crop characteristics should be well monitored by researchers to deliver good results in prediction methods. Quantitative models can produce quantitative results in crop monitoring which will help to develop crop growth in different conditions. Basically problems in predictions are finding proper algorithm for prediction methods and assuming different location results for predictions. Using these data, prediction results are calculated based on algorithm.

CHAPTER-4

METHODOLOGY

4.1 Exploratory Data Analysis:

Best crop prediction project works under classification model. A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category to which a new data will fall under.

Prediction models for the crop prediction were generated by Support vector machine (SVM), Random Forest Classifier and K nearest neighbours (KNN), Decision Tree Algorithm. A comparison of these models was conducted to determine which method produced the best accuracy. Accuracy is only really useful when there are an even distribution of values in a data set.

Above discussed algorithms, we get more accuracy in Decision Tree Algorithm. So, by using Decision tree algorithm we can predict which crop is best based on the parameters.

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes.

There are two main types of Decision Trees:

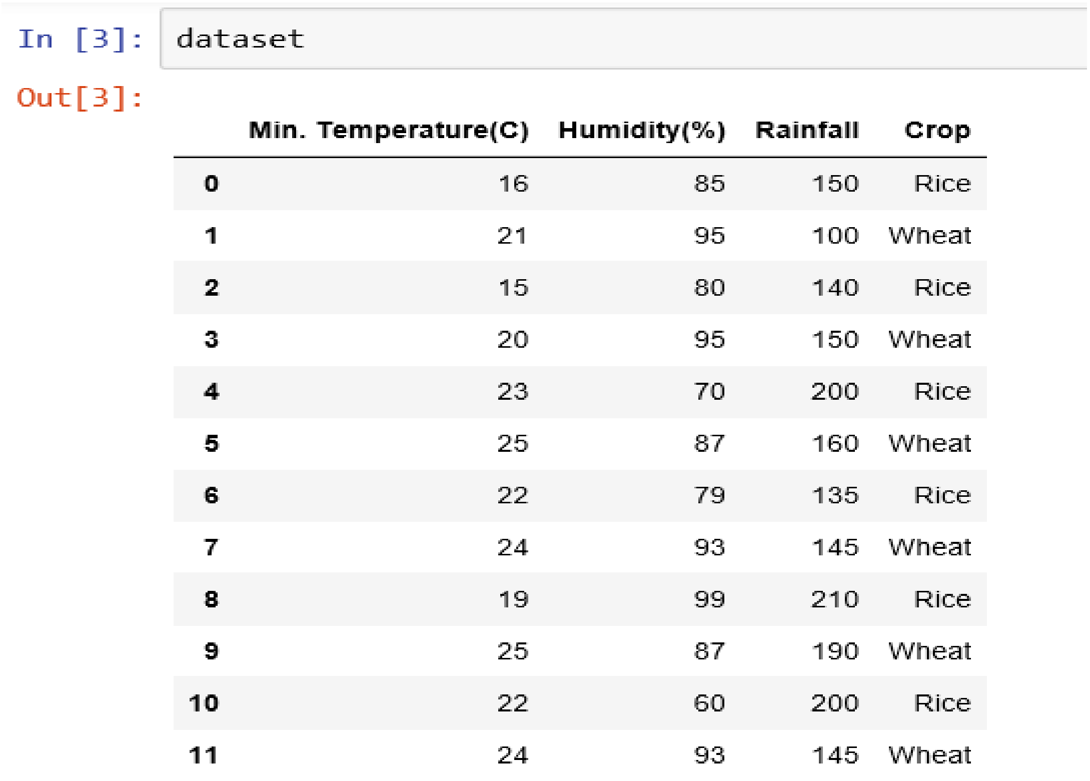
CLASSIFICATION TREES (YES/NO TYPES):

What we’ve seen above is an example of classification tree, where the outcome was a variable like ‘fit’ or ‘unfit’. Here the decision variable is Categorical.

REGRESSION TREES (CONTINUOUS DATA TYPES):

Here the decision or the outcome variable is Continuous, e.g. a number like 123.

4.1.1 FIGURES AND TABLES:



```
In [3]: dataset
Out[3]:
```

	Min. Temperature(C)	Humidity(%)	Rainfall	Crop
0	16	85	150	Rice
1	21	95	100	Wheat
2	15	80	140	Rice
3	20	95	150	Wheat
4	23	70	200	Rice
5	25	87	160	Wheat
6	22	79	135	Rice
7	24	93	145	Wheat
8	19	99	210	Rice
9	25	87	190	Wheat
10	22	60	200	Rice
11	24	93	145	Wheat

Table: 1.1

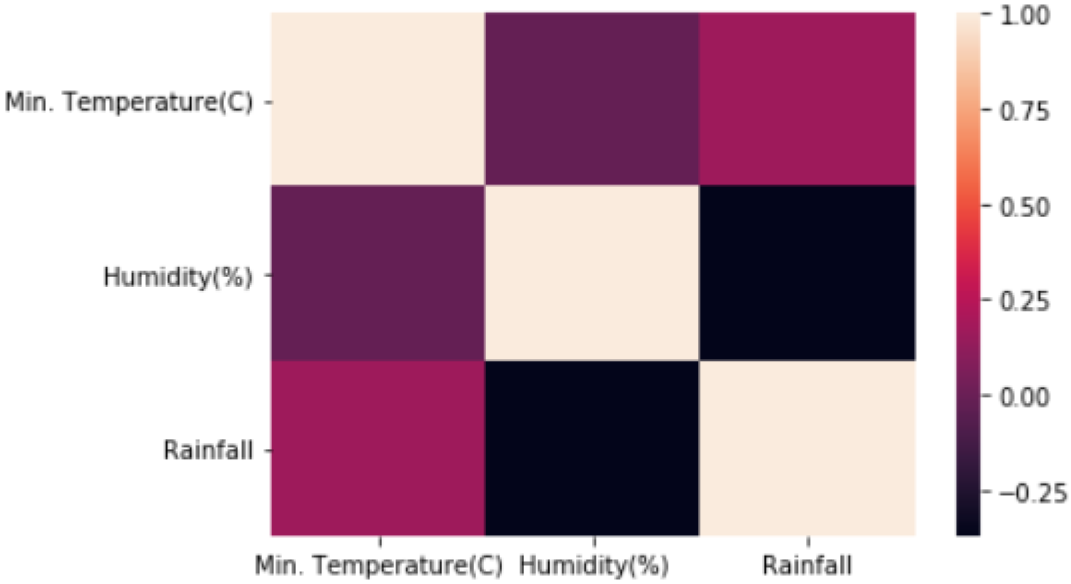
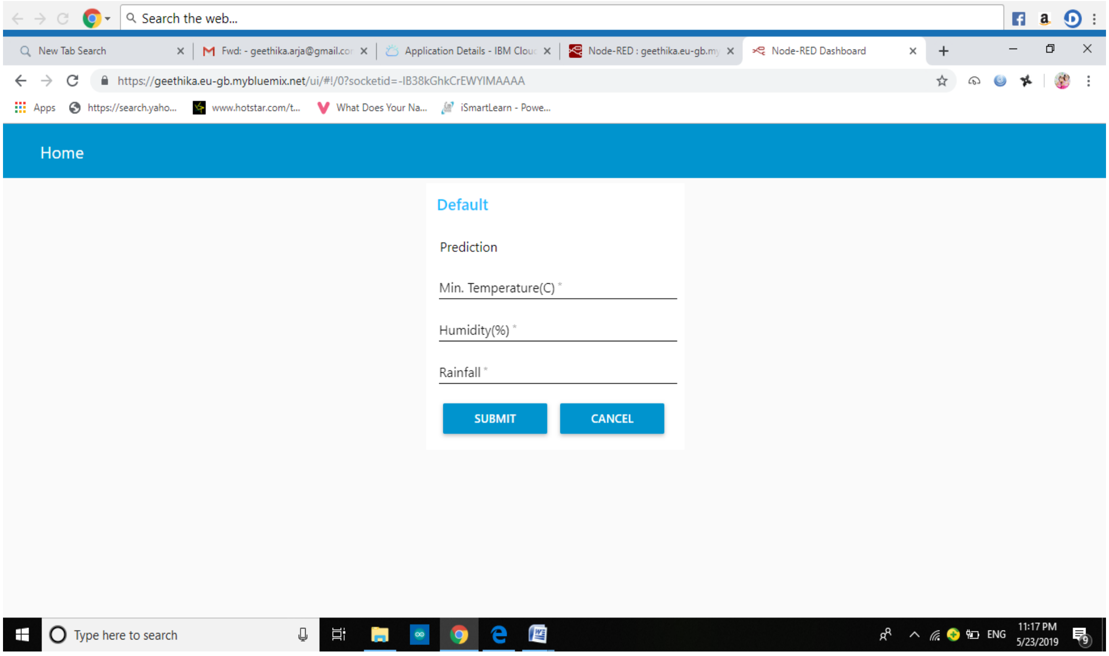


Fig: 1.1



Default

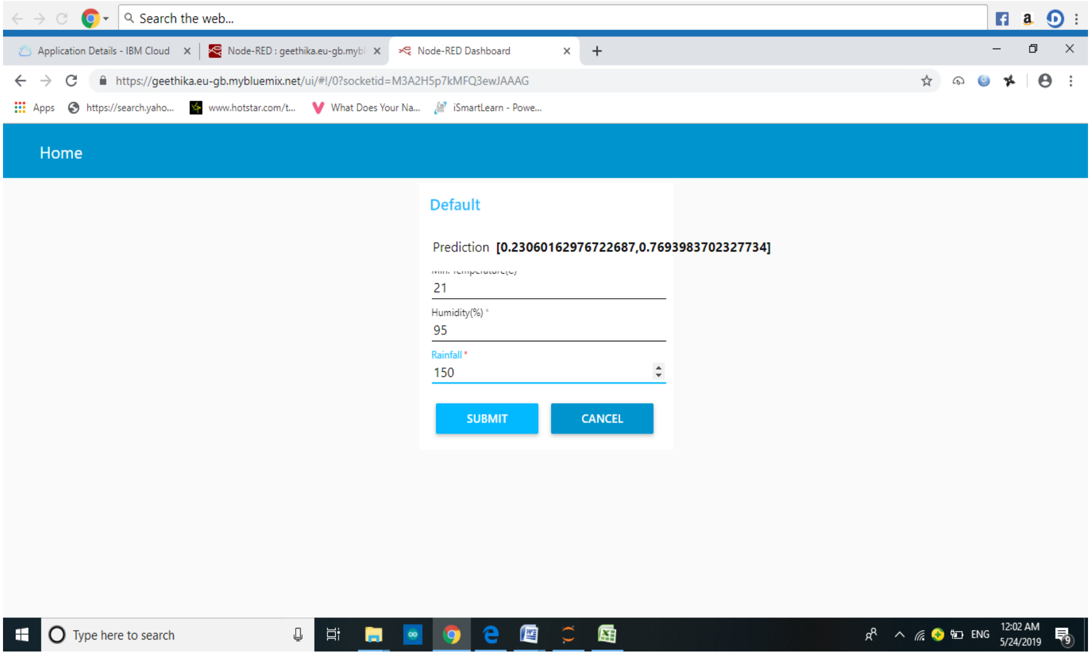
Prediction

Min. Temperature(C) *

Humidity(%) *

Rainfall *

Fig: 1.2



Default

Prediction **[0.23060162976722687,0.7693983702327734]**

Min. Temperature(C)
21

Humidity(%) *
95

Rainfall *
150

Fig: 1.3

In the Fig: 1.3, the Zeroth index indicates Rice and First index indicates Wheat prediction values.

4.2 Statistical techniques and visualization:

NUMPY:

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using import numpy as np.

NumPy’s main object is the homogeneous multidimensional array. It is a table with same type elements, i.e, integers or string or characters (homogeneous), usually integers. In NumPy, dimensions are called axes. The number of axes is called the rank.

PANDAS:

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels. Hence, with 2d tables, pandas is capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using import pandas as pd.

New columns and rows can be easily added to the dataframe. In addition to the basic functionalities, pandas dataframe can be sorted by a particular column. Dataframes can also be easily exported and imported from CSV, Excel, JSON, HTML and SQL database.

MATPLOTLIB:

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB analogs with similar arguments.

On the X array below we saying... include all items in the array from 0 to 2. On the y array below we are saying... just use the column in the array mapped to the **3rd row**. The **Crop** column. We are using group by to view the distribution of values in our **Crop** column. Recall that this column is our **target variable**. It's that thing we are trying to predict.

Best Crop prediction is a classification problem, we will import the DecisionTreeClassifier function from the sklearn library. Next, we will set the 'criterion' to 'entropy', which sets the measure for splitting the attribute to information gain. Accuracy is only really useful when there are an even distribution of values in a data set. This module for Node-RED contains a set of nodes which offer machine learning functionalities. Such nodes have a python core that takes advantage of common ML libraries such as SciKit-Learn and Tensorflow. Classification and outlier detection can be performed through the use of this package. These flows create a dataset, train a model and then evaluate it. Models, after training, can be use in real scenarios to make predictions. Flows and test datasets are available in the 'test' folder. Make sure that the paths specified inside nodes' configurations are correct before trying to execute the program.

4.3 Data modelling and visualization:

Imported libraries are numpy, pandas, matplotlib. NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a list for most popular Python libraries for Data Science. A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program) and read it using a method called read_csv which can be found in the library.

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.

Now we need to split our dataset into two sets a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import **test_train_split** from **model_selection** library of scikit. The final step of data preprocessing is to apply the very important feature scaling. It is a method used to standardize the range of independent variables or features of data.

CHAPTER-5

REFERENCES

https://www.kaggle.com/

https://www.irjet.net/

http://www.ijaerd.com/

CHAPTER-6

CONCLUSION

6.1 Conclusion:

The aim of this research is to propose and implement a system to predict the best crop from the collection of past data. This has been achieved by applying decision tree algorithm on previous dataset.

6.2 Future Work:

This system can be further extended by applying on the various crops to predict in various regions using different algorithms.