1. INTRODUCTION

Agriculture in India stretches lower back to the Indus Valley Civilization Era, and perhaps a great deal in advance in a few areas of Southern India. In phrases of agriculture output, India stands 2nd withinside the global. While agriculture's share of the Indian economic system has regularly reduced to much less than 15% because of the speedy growth of the economic and carrier sectors, the sector's significance in India's monetary and social material extends a long way past this metric. The cause for this deterioration withinside the agriculture enterprise is due to the fact farmers aren't empowered, and there may be a loss of utility of data era withinside the farming sector. Farmers are much less informed approximately the vegetation they cultivate. We normally triumph over this task with the aid of using suitable deep gaining knowledge of algorithms to forecast crop output and call primarily based totally on numerous parameters like as temperature, rainfall, season, and location. Based at the dataset furnished with the aid of using the Indian government, this takes a look at gives a Neural Network version to forecast agricultural manufacturing and crop fulfillment rate.

The primary task encountered whilst assembling the paintings became the shortage of a unmarried supply dataset to teach the advised version on. To cope with those issues, all dispersed statistics is amassed and applicable characteristic engineering and statistics pre-processing steps are employed. The dataset is massive, comprising statistics for all regions of India that have been filtered to accumulate statistics for Telangana kingdom, ensuing in 12000 entries. The crop cycle statistics for summer, Kharif, Rabi, fall, and the complete year is used. To acquire statistics for the kingdom of Telangana, the dataset is filtered the usage of Python Pandas and Pandas Profiling tools. The crop yield forecast version employs a synthetic neural network's lower back propagation technique. The era of multilayer perceptron's is employed. The proposed paintings have a huge variety of packages in enhancing real-global farming conditions. Every year, a big quantity of crop is broken as a result of a lack of knowledge of climate styles along with temperature, rainfall, and so on, that have a great effect on crop output. This initiative now no longer handiest aids in forecasting those traits in the course of the year, however it additionally aids in projecting agricultural yields in numerous seasons primarily based totally on ancient trends. As a result, it permits farmers to pick out the quality crop to plant on the way to incur the fewest losses. Different regression fashions also are built the usage of device gaining knowledge of, and their performance and accuracy are as compared to the Neural Network version on the way to offer a few tangible results.

1.1 PROBLEM STATEMENT

Crop yield forecasting will surely advantage farmers. The farmer might also additionally make crop choice selections and make a contribution greater to the farm's earnings. There are numerous crop manufacturing prediction fashions available, a number of which might also additionally employ meteorological data. Parameters which might be genuine, rather than parameters which might be static. We commonly conquer this undertaking with the aid of using utilizing suitable deep gaining knowledge of algorithms to forecast crop output and call primarily based totally on plenty of parameters like as temperature, rainfall, season, and location. Based at the dataset furnished with the aid of using the Indian government, this looks at provides a Neural Network version to forecast agricultural manufacturing and crop achievement rate.

1.2 OBJECTIVES

The main challenge encountered when assembling the work was the lack of a single source dataset to train the suggested model on. To address these issues, the following objectives are proposed in order to rectify the problems in the existing system.

- To gather all dispersed data is collected and perform relevant feature engineering and data pre-processing steps are employed.

- Design a neural network model and optimize it with appropriate selection of activation function, epochs, batch size and optimizer in order to increase the success rate.

- Compare the performance of the designed neural network with other classic Machine Learning models.

- To add additional module that would provide smart solutions to better the yield (if low).

- Incorporate a module that would project the profit that one can expect from the predicted yield.

1.3 EXISTING SYSTEM

Many models were earlier designed to resolve the current problem statement. While some focused on working with static data and performing relevant data cleaning process to enhance the model accuracy, others tried to create a model that could process remote sensing data and eradicate the hassle with the dispersed dataset and the ETL process associated with it.

The classic machine learning models were inclined to limit down its capabilities to only one or two features that would impact the crop yield directly. Other features were studied using different approach separately.

On the other hand, research was conducted in order to eliminate the trouble of dealing with static and scattered datasets and shift the attention to image processing while utilizing satellite images to identify and anticipate crop output. However, it was determined that this technique required a lengthy training period with no certainty of consistency.

Considering the following issues, the existing system has the following disadvantages:

- All the features and attributes that affect the crop yield are not taken into account together.

- The issue of scattered data was not considered.

- Remote sensing data cannot be used to provide consistent result.

- Classic machine learning models were not able to detect complex patterns in the data.

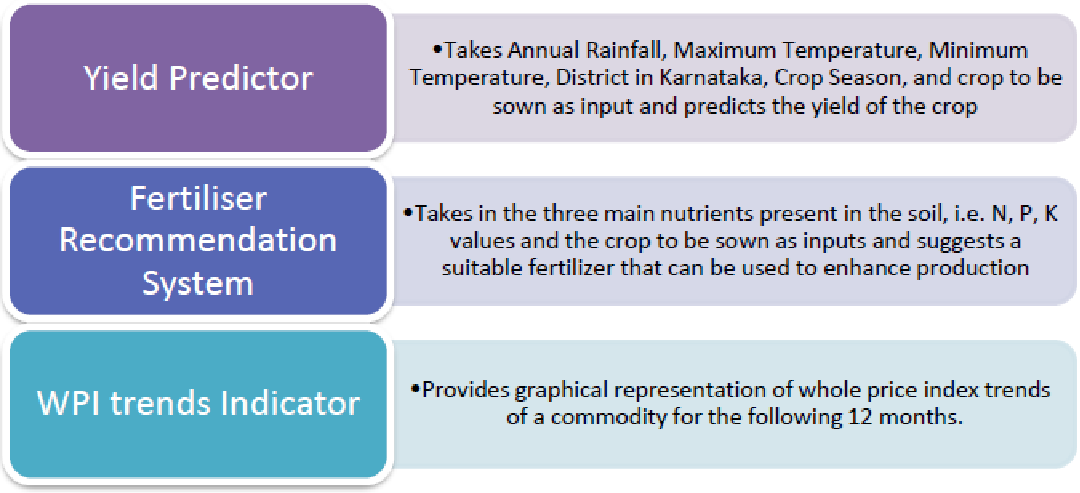
Crop yield prediction may be a very active area of current research interest and has been so since the 1980s. However, within the period the work was mainly concerned with the study of linear systems models and hence was only concerned with the linear relationships among the varied agricultural parameters. Therefore, most of the traditional or traditional models don't seem to be able perform well because they weren't able to effectively accommodate the complexity and non-linear nature of the info. Basically, crop prediction models are divided into two classes; statistical model and crop simulation model. the first stage of the modeling usually involves statistical methods. this can be where the systems use various regression techniques that compute crop yield empirically.

On the opposite hand, simulation models involve the physiologically – based systems of either crops or plant which may affect growth either internally or externally; and is generally involve mathematical analysis so as to predict yield. As an example, allow us to consider one among the wheat yield prediction models as within the case of the Crop Estimation through Resource and Environment Synthesis – Wheat (CERES) model. this kind of simulation model uses a group of information which has the weather, soil attributes and also the detailed management practice which specific farm uses. the following model is representative of the more complex models where a awfully complex set of knowledge is employed to predict wheat yield. ECOSYS and SIRIUS are categorized as complex models that use plenty of differing kinds of information and rely heavily on computer design to simulate the expansion of wheat as within the CERES.

1.4 PROPOSED SYSTEM

Considering the issues discussed above, the proposed system aims to rectify it by gathering the dispersed data from different sources, performing relevant feature engineering to merge those datasets in order to obtain a single source of data, Use Neural Network to make the model consistent and capable to understanding the complex correlations and provide other add on modules in order to aid the better result of the yield.

Figure 1.4 depicts the modules constructed in the proposed work. It is made up of three major components. One module is dedicated to early yield forecast, utilizing characteristics such as crop area, yearly rainfall, temperature data, and Telangana state output history from 1998 to 2014. The second module is the fertilizer recommendation system, which takes into account the quantity of three major nutrients in the soil, namely nitrogen, phosphorus, and potassium, as well as the crop to be sown, and suggests the fertilizer that may be utilized to increase crop output. The third module is a crop-specific WPI trends indicator that depicts the whole index price over the following 12 months graphically.



- Takes Annual Rainfall, Maximum Temperature, Minimum Temperature, District in Karnataka, Crop Season, and crop to be sown as input and predicts the yield of the crop

- Takes in the three main nutrients present in the soil, i.e. N, P, K values and the crop to be sown as inputs and suggests a suitable fertilizer that can be used to enhance production

- Provides graphical representation of whole price index trends of a commodity for the following 12 months.

Figure 1. System Modules

Figure 1.1: System Modules

1.5 DATASETS:

The data for this study was obtained from the Indian government’s website. The datasets are freely accessible for research and scholarly purposes. The collection contains information spanning the years 1997 to 2014. All the required datasets are collected and pre-processed to obtain a final dataset which will be used to train the model. For the experiment in this investigation, the following parameters are used.

- **Crop** – The dataset contains a number of crops such as Sunflower, Bajra, Jowar, Season, groundnut, rice, cottonseed, tur etc.

- **State** – TELANGANA

- **District** – ‘ADILABAD’, ‘KARIMNAGAR’, ‘KHAMMAM’, ‘MAHBUBNAGAR’, ‘MEDAK’, ‘NALGONDA’, ‘NIZAMBAD’, ‘RANGA REDDY’, ‘WARANGAL’, ‘HYDERABAD’ etc.

- **Season** – Kharif, Rabi, Zaid, Whole year

- **Year –** 1998 to 2014

- **Rainfall –** Monthly rainfall data (mm) for each district of Telangana State, whose sum is taken to evaluate annual rainfall and concatenated to the final dataset.

- **Temperature –** District wise maximum and minimum temperature (˚C), who’s mean is calculated and appended to the final dataset

- **Production –** It is given in tons per hector in lakh

- **Fertilizers –** Describes the amount of N, P, K required in the soil in order to grow a specific crop in a region.

The data in the dataset provided by the government has been examined for outliers and noise. The variables were also transformed to category and numerical formats as needed by the model. Figure 1.5 displays a bar graph from 1998 to 2014 that correlates the season and produce of all Telangana districts. According to the graph below, the bulk of the crops cultivated in Telangana are year-round crops.

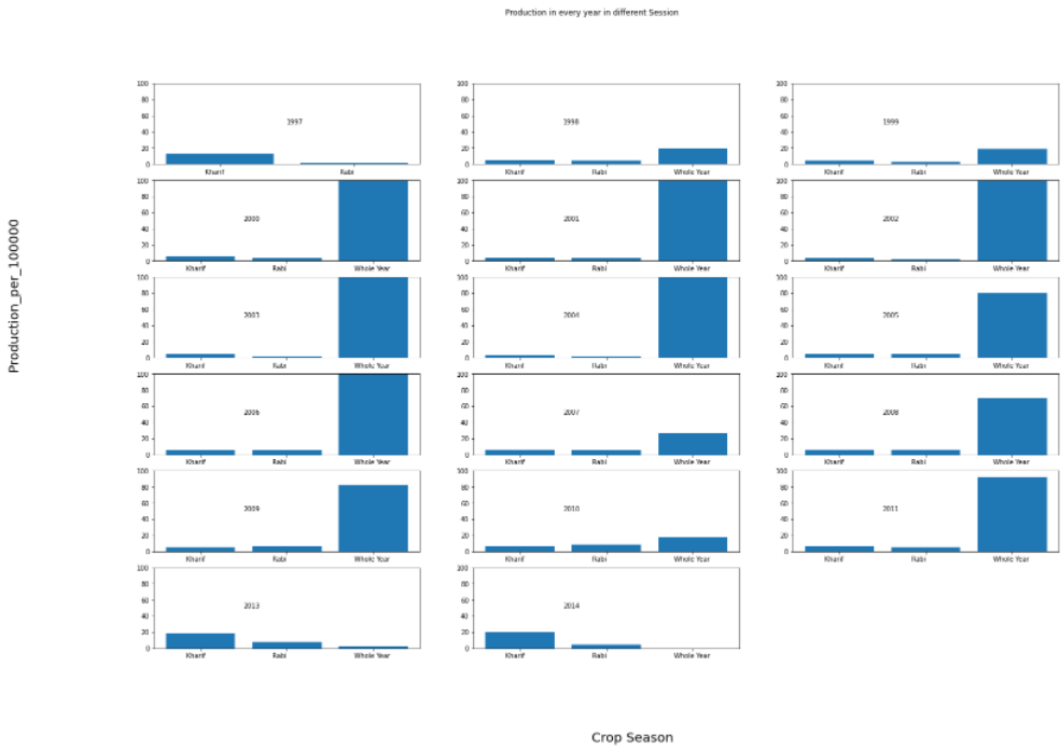


Figure 1.2: Year-specific Bar graph to depict the correlation of season vs. Production

2. LITERATURE SURVEY

[1] Crop Yield Prediction to Maximize Profit using Machine Learning: [3]

Authors: Gabhane Srushti, Shaikh Naushinnaaz, Sadavarte Shivani, Khan Huda, A.I. Waghmare

Abstract: The proposed system applies device gaining knowledge of and prediction algorithms to advise the fine appropriate plants for the farmers. The purpose of the gadget is to lessen the losses because of drastic climatic modifications and growth the yield fees of plants. The gadget integrates the records received from the beyond prediction, modern climate and soil circumstance because of these farmers receives the concept and listing of plants that may be cultivated. Machine Learning techniques are widely utilized in prediction strategies like SVM (Support Vector Machine), linear regression. This in go back offers the fine crop for cultivation primarily based totally at the modern surroundings circumstance. The proposed gadget considers the rainfall quantity of beyond, modern and destiny and additionally the sort of soil the farmer has. Based in these parameters the right plants for the given circumstance are anticipated the use of the device gaining knowledge of algorithms greater correct prediction consequences are produced.

Disadvantages: The following System incorporated diverse parameters however could fail to detect complicated styles if new records are provided. This problem can be resolved through the use of appropriate Deep Learning Algorithm that could offer correct, green and steady consequences.

[2] Correlation of Climatic Factors with Cereal Crops Yield. A Study from Historical Data of Morang District, Nepal: [4]

Authors: Badri Khanal

Abstract: The present study is based on the secondary sources of information on temperature, rainfall and productivity of four major cereals (Rice, Maize, Finger Millet and Wheat) in Morang district of Nepal. An overall of 17 years data (1995-2011) on yield of plants, annual overall rainfall, annual imply most temperature and annual imply minimal temperature is analyzed. The suitability evaluation of plants suggests that everyone the 4 cereals discovered to be appropriate for cultivation in temperature variety of Morang district, while irrigation is needed further to recorded rainfall in case of rice and wheat. The manufacturing of 3 cereals besides millet (which is nearly stable) has improved in the course of the observe period. The evaluation of correlation coefficient suggests that maize yield and minimal temperature have robust tremendous correlation (0.7755). The linear regression evaluation confirmed that the yield of maize became sizable and rather touchy to mixed impact of all 3 climatic factors (R2 0.7414).

Disadvantages: The study focuses on a large geographical area, thereby hindering the model accuracy. Also, it takes into account only two commodities i.e., rice and wheat and deduces a correlation of the same with climatic factors.

[3] Agro based Crop and Fertilizer Recommendation system using Machine Learning: [5]
Authors: Preethi G, Rathi Priya V, Sanjula S M, Lalitha S D, Vijaya Bindhu B

Abstract: The project explains how the quantity of soil nutrients and environmental factors observed via way of means of the guidelines for cropping and unique fertilization of the web page may be established. The choice of the fine crop for the soil and the sowing of it to offer the entire yield is one of the key troubles in agriculture. The proposed technique takes the soil and PH samples because the enter and enables to are expecting the vegetation that may be advocated appropriate for the soil and fertilizer that may be used as the answer withinside the shape of the webpage. So, the soil statistics is accumulated via sensors and the statistics transmitted from the Arduino via Zigbee and WSN (Wireless Sensor Network) to MATLAB and reading the soil statistics and processing is achieved with assist of ANN (Artificial Neural Network) and crop tips is achieved the usage of SVM (Support Vector Machine).

Disadvantages: The following version takes account handiest the soil parameter and recommends appropriate fertilizer to decorate the soil quality. This looks at shows the approaches to higher the yield of the crop via way of means of oblique relation to soil and its nutrient values. However, this module may be utilized in affiliation to the primary version.

[4] Rice Crop Yield Prediction using Artificial Neural Networks: [6]

Authors: Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong

Abstract: This project aimed to apply neural networks to are expecting rice manufacturing yield and look at the elements affecting the rice crop yield for numerous districts of Maharashtra nation in India. Data had been sourced from publicly to be had Indian Government's information for 27 districts of Maharashtra nation, India. The parameters taken into consideration for the existing observe had been precipitation, minimal temperature, common temperature, most temperature and reference crop evapotranspiration, area, manufacturing and yield for the Kharif season (June to November) for the years 1998 to 2002. The dataset changed into processed the use of WEKA tool.

Disadvantages: This method has been confirmed with the aid of using forecasting of rice crop yield prediction for Kharif season from year 1998 to 2002 for Maharashtra nation of India, on the idea of various predictor variables inclusive of precipitation, minimal temperature, common temperature, most temperature, reference crop evapotranspiration and yield. Artificial Neural Networks with Multilayer Perceptron had been taken into consideration for the existing research.

[5] Rice Crop Yield Prediction in India using Support Vector Machines: [7]

Authors: Niketa Gandhi, Owaiz Petkar, Leisa J. Armstrong, Amiya Kumar Tripathy

Abstract: This project gives the evaluate on use of such device getting to know approach for Indian rice cropping areas. This paper discusses the experimental effects received by making use of SMO classifier the use of the WEKA device at the dataset of 27 districts of Maharashtra state, India. The dataset taken into consideration for the rice crop yield prediction became sourced from publicly to be had Indian Government records. The parameters taken into consideration for the look at have been precipitation, minimal temperature, common temperature, most temperature and reference crop evapotranspiration, area, manufacturing and yield for the Kharif season (June to November) for the years 1998 to 2002. For the prevailing look at the imply absolute error (MAE), root imply squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) have been calculated. The experimental effects confirmed that the overall performance of other strategies at the equal dataset became a lot higher in comparison to SMO.

Disadvantages: The following look at demonstrates using Support Vector device which is a traditional device getting to know set of rules to depict the rice manufacturing the use of Weather as properly as historic data. However, the version can emerge as inconsistent because of the static nature of dataset and the incapacity of the version to examine from newly introduced data.

[6] Wheat Yield Prediction: Artificial Neural Network based Approach: [8]

Authors: Muhd Khairulzaman Abdul Kadir, Mohd Zaki Ayob, Nadaraj Miniappan

Abstract: In this project, our wheat yield prediction version is designed the use of a Multi-Layer Perceptron (MLP) backpropagation-primarily based totally- feed ahead synthetic neural network (ANN). The facts used become climate facts including: sun, frost, rain and temperature because the enter parameters from 12 months 1997-2007. The output parameter of the version is the use of the wheat yield facts for the years 1997 – 2007. The facts are split into 3 separate sets; – for training, validation and testing. Our MLP become capable of predict, wheat yield with an accuracy of 98 %. Hence our MLP primarily based totally wheat yield prediction version indicates wonderful promise as a device for you to be capable of offer fantastically correct wheat yield prediction and can be implemented to different plants.

Disadvantages: The version is flawlessly described to match for one commodity. However, a comparable version may be used to encompass numerous kinds of plants with comparable accuracy.

3. REQUIREMENTS

SOFTWARE REQUIREMENT SPECIFICATION

A Software Requirements Specification (SRS) -a requirements specification for a software system – is a complete description of the behavior of a system to be developed. In addition to a description of the software functions, the SRS also contains non-functional requirements. Software requirements are a sub-field of software engineering that deals with the elicitation, analysis, specification, and validation of requirements for software.

REQUIREMENTS

HARDWARE REQUIREMENTS

System	:	Pentium IV 2.4 GHz or more
Hard Disk	:	40 GB.
Monitor	:	15 VGA Color.
Ram	:	512 Mb
Processor	:	Dual Core

SOFTWARE REQUIREMENTS

Operating System	:	Windows
IDE	:	Jupyter Notebook, HTML
Datasets	:	".csv" files
Modules	:	TensorFlow, Pandas, Matplotlib, NumPy, Sklearn, Flask, Keras
Software	:	Anaconda Navigator, Visual Studio Code

LIBRARIES AND PACKAGES:

NumPy:

NumPy (pronounced /ˈnʌmpaɪ/ (NUM-py) or sometimes /ˈnʌmpi/ (NUM-pee)) is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors. NumPy targets the CPython reference implementation of Python, which is a non-optimizing bytecode interpreter. Mathematical algorithms written for this version of Python often run much slower than compiled equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly inner loops, using NumPy.

Using NumPy in Python gives functionality comparable to MATLAB since they are both interpreted, and they both allow the user to write fast programs as long as most operations work on arrays or matrices instead of scalars. In comparison, MATLAB boasts a large number of additional toolboxes, notably Simulink, whereas NumPy is intrinsically integrated with Python, a more modern and complete programming language. Moreover, complementary Python packages are available; SciPy is a library that adds more MATLAB-like functionality and Matplotlib is a plotting package that provides MATLAB-like plotting functionality. Internally, both MATLAB and NumPy rely on BLAS and LAPACK for efficient linear algebra computations.

Pandas:

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself.

Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007. Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features

Developer Wes McKinney started working on pandas in 2008 while at AQR Capital Management out of the need for a high performance, flexible tool to perform quantitative analysis on financial data. Before leaving AQR he was able to convince management to allow him to open source the library.

Another AQR employee, Chang She, joined the effort in 2012 as the second major contributor to the library. In 2015, pandas signed on as a fiscally sponsored project of NumFOCUS, in the United States.

Matplotlib:

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.

SciPy makes use of Matplotlib. Matplotlib was originally written by John D. Hunter. Since then, it has an active development community and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in August 2012 and was further joined by Thomas Caswell.

Matplotlib 2.0.x supports Python versions 2.7 through 3.10. Python 3 support started with Matplotlib 1.2. Matplotlib 1.4 is the last version to support Python 2.6. Matplotlib has pledged not to support Python 2 past 2020 by signing the Python 3 Statement.

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross platform library for making 2D plots from data in arrays. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, xPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

Seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library, you can read the introductory notes. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see what you can do with seaborn, and then check out the tutorial and API reference to find out how.

Data visualization has been one of the most important driving forces in the field of Data Analytics. It powers millions of businesses across the globe and provides in-depth insights into the data at hand. This makes it vital that you know about the latest and the most popular data visualization libraries out there: Seaborn in Python is one of these

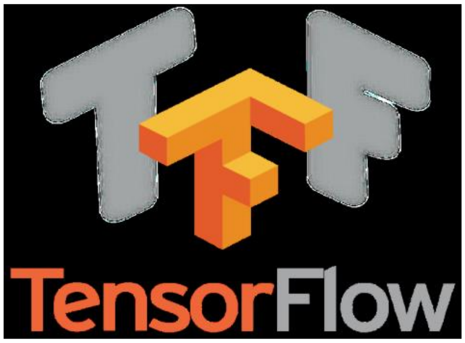
Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

TensorFlow:

TensorFlow is a free and open-source software library for machine learning. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. Tensorflow is a symbolic math library based on dataflow and differentiable programming. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. Its flexible architecture allows for the easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices.

TensorFlow computations are expressed as stateful dataflow graphs. The name TensorFlow derives from the operations that such neural networks perform on multidimensional data arrays, which are referred to as tensors. During the Google I/O Conference in June 2016, Jeff Dean stated that 1,500 repositories on GitHub mentioned TensorFlow, of which only 5 were from Google.In December 2017, developers from Google, Cisco, RedHat, CoreOS, and Cai Cloud introduced Kubeflow at a conference. Kubeflow allows operation and deployment of TensorFlow on Kubernetes.

In May 2016, Google announced its Tensor processing unit (TPU), an application-specific integrated circuit (ASIC, a hardware chip) built specifically for machine learning and tailored for TensorFlow. A TPU is a programmable AI accelerator designed to provide high throughput of low-precision arithmetic (e.g., 8-bit), and oriented toward using or running models rather than training them. Google announced they had been running TPUs inside their data centers for more than a year, and had found them to deliver an order of magnitude better optimized performance per watt for machine learning.



Keras:

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Up until version 2.3, Keras supported multiple backends, including TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML. As of version 2.4, only TensorFlow is supported. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro- Electronic Intelligent Robot Operating System), and its primary author and maintainer is François Chollet, a Google engineer. Chollet is also the author of the XCeption deep neural network model.

Keras contains numerous implementations of commonly used neural-network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code. The code is hosted on GitHub, and community support forums include the GitHub issues page, and a Slack channel.

In addition to standard neural networks, Keras has support for convolutional and recurrent neural networks. It supports other common utility layers like dropout, batch normalization, and pooling. Keras allows users to productize deep models on smartphones (iOS and Android), on the web, or on the Java Virtual Machine. It also allows use of distributed training of deep-learning models on clusters of Graphics processing units (GPU) and tensor processing units.

4. PROJECT DETAILS

4.1 SYSTEM DESIGN

System design is the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements. System design could see it as the application of systems theory to product development. Theory is some overlap with the disciplines of system analysis, systems architecture and systems engineering. If the broader topic development ″blends the perspective of marketing, design, and manufacturing into a single approach to product development,″ then design the act of talking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

Until the 1990s systems design had crucial and respected role in the data processing industry. In the 1990s standardization of hardware and software resulted in the ability to build modular systems. The increasing importance of software running on generic platforms has enhanced the discipline of software engineering.

Object-oriented analysis and design methods are becoming the most widely used methods for computer systems design. The UML has become the standard language in object-oriented analysis and design. It is widely used for modelling software systems and is increasingly used for high designing non- software systems and organizations. System design is one of the most important phases of software development process. The purpose of the design is to plan the solution of a problem specified by the requirement documentation. In other words, the first step in solution is the design of the project.

The design of the system is perhaps the most critical factor affecting the quality of the software. The objective of the design phase is to produce overall design of the software. It aims to figure out the modules that should be in the system to fulfil all the system requirements in efficient manner. The design will contain the specification of all the modules, their interaction with other modules and the desired output from each module.

4.2 DATA FLOW DIAGRAM

A data flow diagram (DFD) is a graphical representation of the flow of the visualization of data processing. On a DFD, data items flow from an external data source or internal data source to internal data source or external data sink via an internal process. DFD provides no information about the timing of process or about whether process will operate in sequence or in parallel.

The diagram below depicts the flow of data through the system. The flow of all modules stays constant, with the only variation being the final result. Inputs for the relevant modules, such as yearly rainfall, temperature, district, crop name, season, and fertiliser data, are obtained via a web-based application by the user. A JSON data object is returned, which has been scaled with the sklearn package. The categorical data, such as district, season, and crop name, is again one hot encoded, and the data object is ultimately transformed to a NumPy array. This information is subsequently put into the Neural Network model.

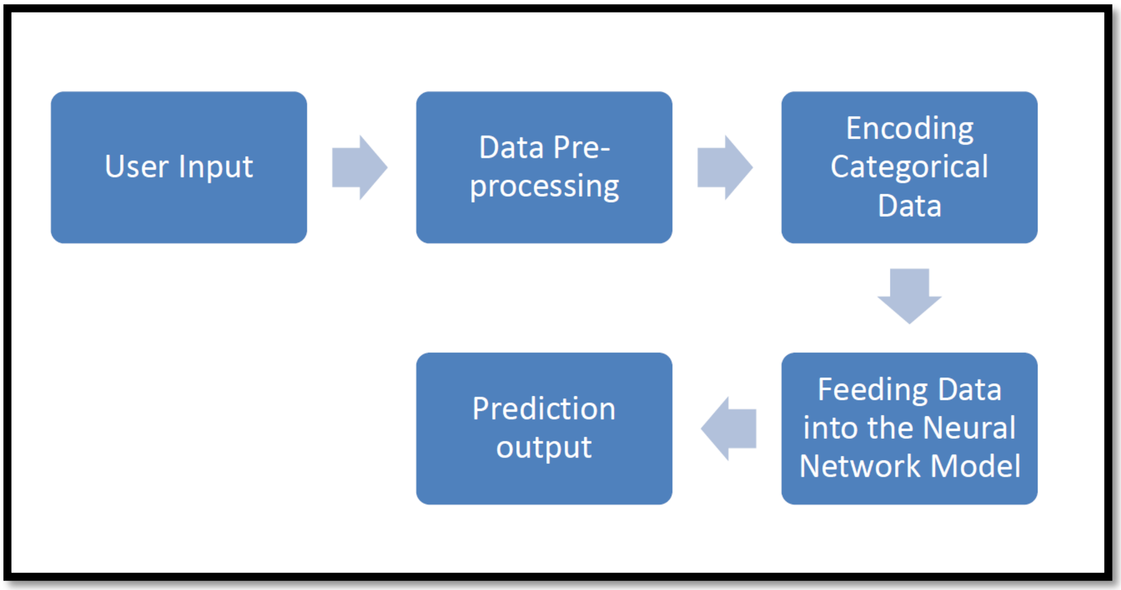

User Input
Data Pre-processing
Encoding Categorical Data
Feeding Data into the Neural Network Model
Prediction output

Fig 4.1: Data Flow Diagram of the Proposed System

5. SYSTEM IMPLEMENTATION

Implementation is the realization of an application, or execution of a plan, idea, model, design, specification, standard, algorithm, or policy. In other words, an implementation is a realization of a technical specification or algorithm as a program, software component, or other computer system through programming and deployment. Many implementations may exist for a given specification or standard.

Implementation is one of the most important phases of the Software Development Life Cycle (SDLC). It encompasses all the processes involved in getting new software or hardware operating properly in its environment, including installation, configuration, and running, testing, and making necessary changes. Specifically, it involves coding the system using a particular programming language and transferring the design into an actual working system. This phase of the system is conducted with the idea that whatever is designed should be implemented; keeping in mind that it fulfils user requirements, objective and scope of the system. The implementation phase produces the solution to the user problem.

5.1 PSEUDOCODE

Pseudo code is an informal high-level description of the operating principle of a computer program or other algorithm. It uses the structural conventions of a programming language, but is intended for human reading rather than machine reading. Pseudo code typically omits details that are not essential for human understanding of the algorithm, such as variable declarations, system-specific code and some subroutines. The programming language is augmented with natural language description details, were convenient, or with compact mathematical notations. The purpose of using pseudo code is that is easier for people to understand than conventional programming language code, and that it is an efficient and environment independent description of the key principles of an algorithm.

It is commonly used in textbooks and scientific publications that are documenting various algorithms, and also in planning of computer program development, for sketching out the structure of the program before the actual coding takes place. No standard for pseudo code syntax exists, as a program in pseudo code is not an executable program. Pseudo code resembles, but should not be confused with skeleton programs, including dummy code, which can be compiled without errors.

Flowcharts and Unified Modelling Language (UML) charts can be thought of as a graphical alternative to pseudo code, but are more spacious on paper.

The Project is divided into 3 different modules:

1. Crop Yield Predictor (Base Module)
2. Fertilizer Recommendation system
3. Whole price index trend analysis

All the module follows the same Data pre-processing steps. The processed data is then fed to the respective Deep Learning models in order to obtain the required results.

5.2 DATA PRE-PROCESSING

[1]. The raw data set was then collated in single sheet which consisted of the following columns in Microsoft Excel: sr. no, name of the state, name of the district, year, precipitation, minimum temperature, average temperature, maximum temperature, soil type, area, production and yield.

[2]. For some of the districts particular year’s climatic parameters or production data was not available hence those records were omitted. That particular year’s data was not used for the current research. Record number was added for each record.

[3]. For preparing the data set for applying multilayer perceptron technique, unrequited columns were removed. They were sr. no, name of the district and year.

[4]. The data set was then sorted on the basis of area. Area less than 100 hectares were not considered for the present research. So those records were omitted.

[5]. the data which is present in label from converted to encoding using sklearn.

[6]. The dataset was then sorted on the basis production.

[7]. we considered production as output parameter and features like: crop, area, district, season.

[8]. This data set was then saved in .csv format for further application of the multilayer perceptron technique in Python TensorFlow.

5.3 CROP YIELD PREDICTOR

This is the base module, which includes comparison to other classic Machine Learning Algorithm as well. This module deals with scattered datasets which is cleaned and process using the data pre-processing step discussed above. The final dataset is then fed to the neural network model, with the following training parameters.

- ➢ batch size=100,
- ➢ epochs=50
- ➢ Layer : 3
- ➢ Neuron at each layer : Layer 1, Layer 2 = 20
- ➢ Layer 3 = 1
- ➢ Optimizer = Adam
- ➢ Activation : ReLu
- ➢ kernel initializer='uniform
- ➢ lr rate : 0.01

The algorithm used for the following module is the Artificial Neural Network. An Artificial Neuron is basically an engineering approach of biological neuron. It has device with many inputs and one output. ANN is consisting of large number of simple processing elements that are interconnected with each other and layered also. An ANN begins with a training phase in which it learns to detect patterns in data, whether visually, audibly, or textually. During this supervised phase, the network compares its actual output to what it was supposed to produce—the expected output.

Backpropagation is used to correct the discrepancy between the two results. This implies that the network works backward, from the output unit to the input units, adjusting the weight of its connections between the units until the discrepancy between the actual and planned outcome generates the smallest mistake possible.

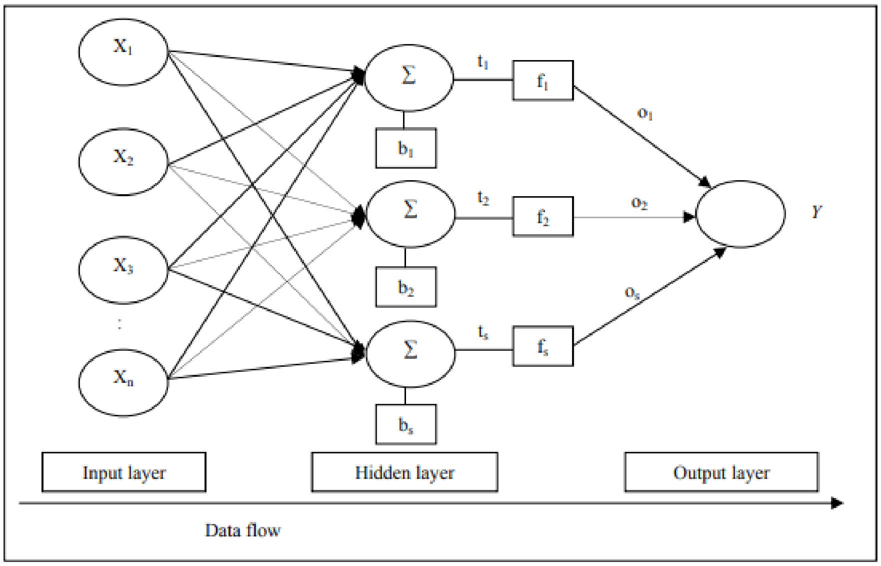


Fig 5.1: Layers and connections of ANN Model

5.4 FERTILIZER RECOMMENDATION SYSTEM

The following module uses three essential nutrients required for a healthy soil i.e., Nitrogen, Phosphorous, and Potassium. The amount of these three nutrients is taken from the user in order to determine, the soil health. After analyzing the soil nutrient contents the module suggest an appropriate fertilizer to balance the soil quality for a better yield.

The following module uses

Logistic Regression with Gradient Descent whose Pseudo code is given below. Logistic regression is a traditional and classic statistical model, which has been widely used in the academy and industry. Unlike linear regression, which is used to make a prediction on the numeric response, logistic regression is used to solve a classification problem.

```
1.Initialize the parameters
2. Repeat
2.2. Make a prediction on y
2.3 Calculate cost function
2.4. Get gradient for cost function
2.5. Update parameters
```

5.5 WHOLE PRICE INDEX ANALYSIS

The following module uses Decision tree to analyze the ongoing Price of different commodity grown through the breath and length of Telangana. It provides a 12 month analysis of the Whole price trends, thereby aiding the agronomic workers have a rough idea on the profit one must expect by cultivating a certain crop.

The pseudo code for the algorithm used is given below:

1. It begins with the original set S as the root node.
2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute.
3. It then selects the attribute which has the smallest Entropy or Largest Information gain.
4. The set S is then split by the selected attribute to produce a subset of the data.
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

6. PERFORMANCE EVALUATION

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out of sample) data. Methods for evaluating a model's performance are divided into 2 categories: namely, holdout and Cross-validation. Both methods use a test set (i.e. data not seen by the model) to evaluate model performance. It's not recommended to use the data we used to build the model to evaluate it. This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This is known as overfitting.

We utilise scatter plots to compare the actual test data output to the predictions generated by the model on the test data. The graph below illustrates a linear connection between the actual and predicted results. A positive slope with a strong correlation between the actual and anticipated results indicates a greater success rate. It can also be stated that for the majority of the test inputs, the model was able to forecast a yield with extremely low error to the actual yield output.

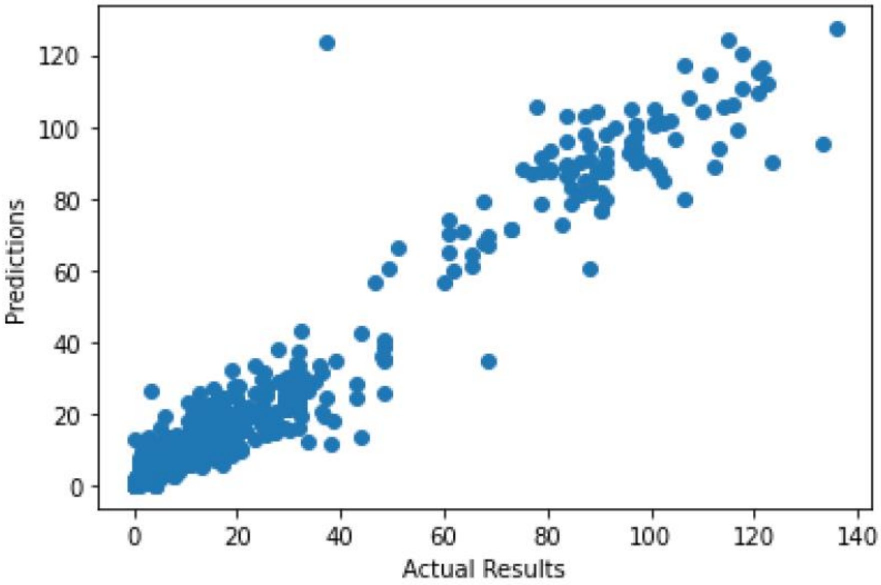


Fig 6.1: Linear correlation between Actual and Predicted Results

The algorithm's performance is measured using the two metrics listed below.

6.1 Mean Absolute Error

In the context of machine learning, absolute error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group. MAE can also be referred as L1 loss function.

As one of the most commonly used loss functions for regression problems, MAE helps users to formulate learning problems into optimization problems. It also serves as an easy-to-understand quantifiable measurement of errors for regression problems. MAE measures the average magnitude of absolute differences between N predicted vectors S ={x1, x2, ..., xN} and S_ = {y1, y2, ..., yN}, the corresponding loss function is defined as:

$$L_{MAE}(S, S_{*}) = 1N\Sigma \|_{Ni=1} x_i - y_i \|$$

where || ・ || denotes L1 norm.

6.2 R-Squared:

R-squared (R2) is a statistical metric that indicates the proportion of the variation explained by an independent variable or variables in a regression model for a dependent variable. Whereas correlation describes the strength of the link between an independent and dependent variable, Rsquared explains how well one variable's variation explains the variance of the other. R-squared is commonly defined as the percentage of a fund's or security's movements that can be explained by changes in a benchmark index. The R-Squared score for the proposed work was able to reach approximately 0.9645 within 50 epochs. The score can be calculated using the formula below:

$$R^2 = 1 - (RSS\ /\ TSS)$$

Where,

R^2 is coefficient of determination

RSS is Sum of Squares of residual

TSS is total sum of squares

The graph below depicts the metric scores discussed above for our neural network model. It plots the Accuracy calculated using R-Square metrics and the mean absolute error with the numbers of epochs take to train the model.

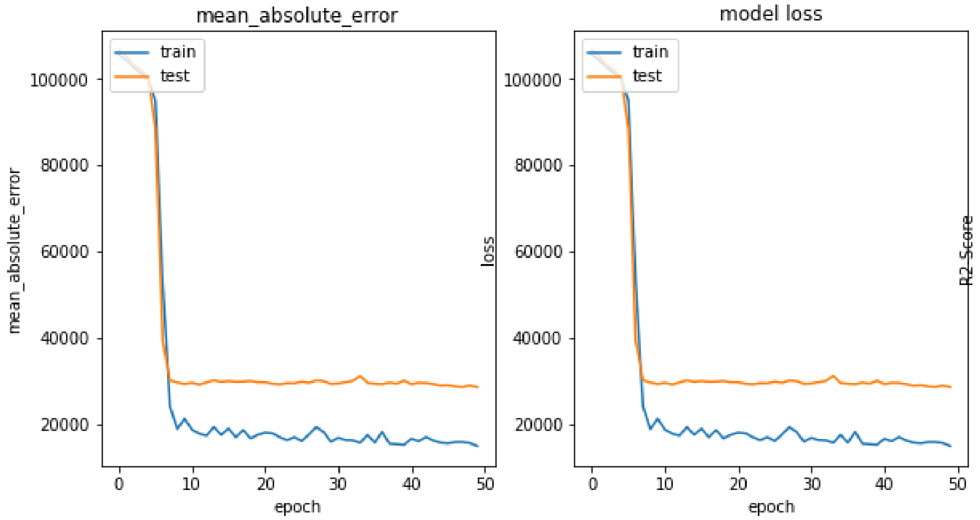


Fig 6.2: Performance graphs of the Model.

Various other Machine Learning model is also trained and evaluated using the metrics discussed above. A bar graph is plotted for the performance of all the models trained to predict the yield results. A side by side comparison clearly tells that Neural Network has outperformed classic Machine Learning models in terms of Accuracy and Minimum error.

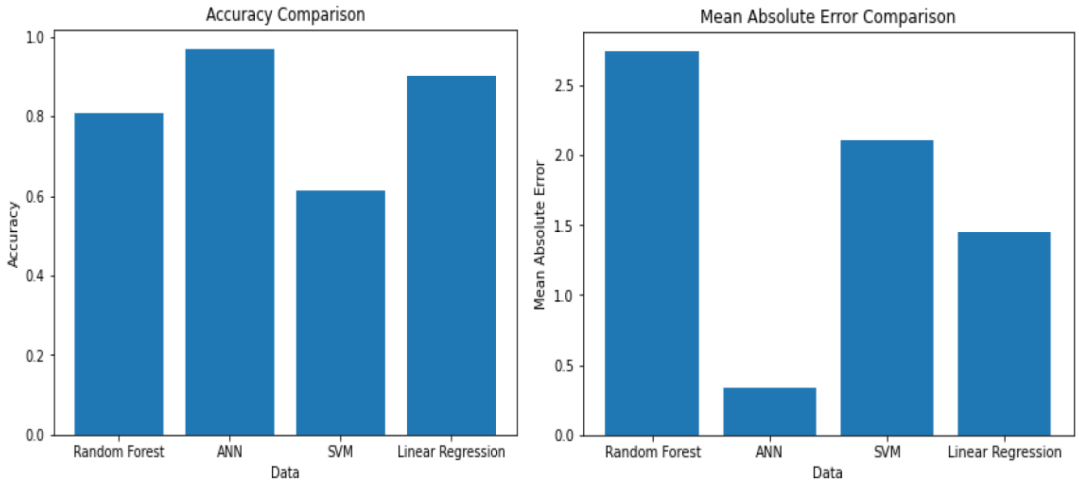
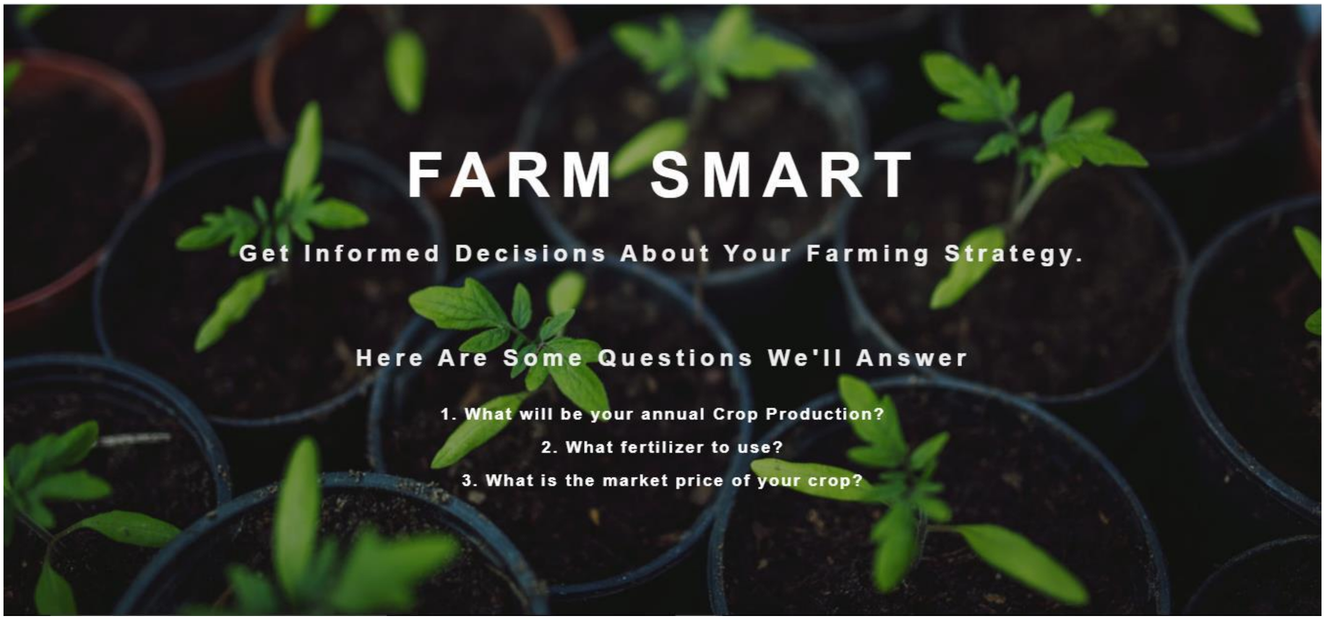


Fig 6.3: Comparative Analysis of the performance of all regression models.

7. OBSERVATION & RESULTS

The project uses several frontend libraries and packages such as chartJS, MaterialiseJS, Bootstrap and JQuery to design a web based application to ease out the access to the users. The figure below shows the landing page of the project. The project is named as Farm Smart.



FARM SMART

Get Informed Decisions About Your Farming Strategy.

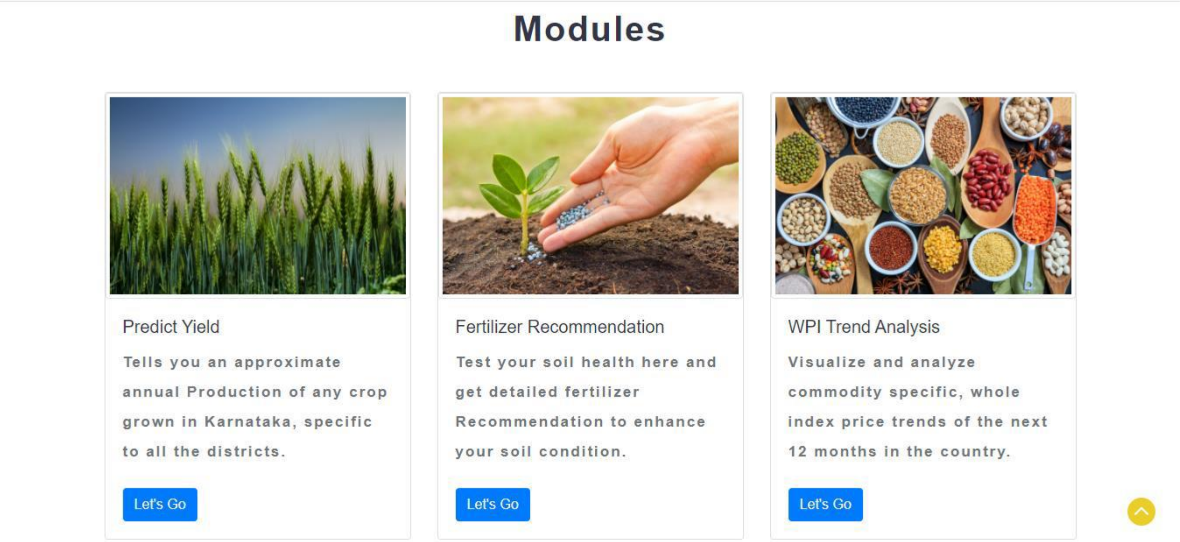
Here Are Some Questions We'll Answer

1. What will be your annual Crop Production?
2. What fertilizer to use?
3. What is the market price of your crop?

Fig 7.1: Landing Page of the Project

The landing page has many navigation points from which the user can easily navigate to different modules without the break of flow. One such navigation element is the sticky Navigation bar with 3 links to respective 3 module. However, the main focus on navigation is the card elements used to navigate to different module as shown below.

As it is clearly seen that the landing page provides access to three main services of focus. Yield Predictor is the base module whereas the fertilizer recommender and Whole Price Index Analysis are the add on modules. The buttons on the cards navigates a user to the respective page of the module wherein the user can provide their inputs and obtain the required results. Bootstrap and JQuery is used to design the page and its elements like the cards and navigation bar.



Modules

Predict Yield

Tells you an approximate annual Production of any crop grown in Karnataka, specific to all the districts.

Fertilizer Recommendation

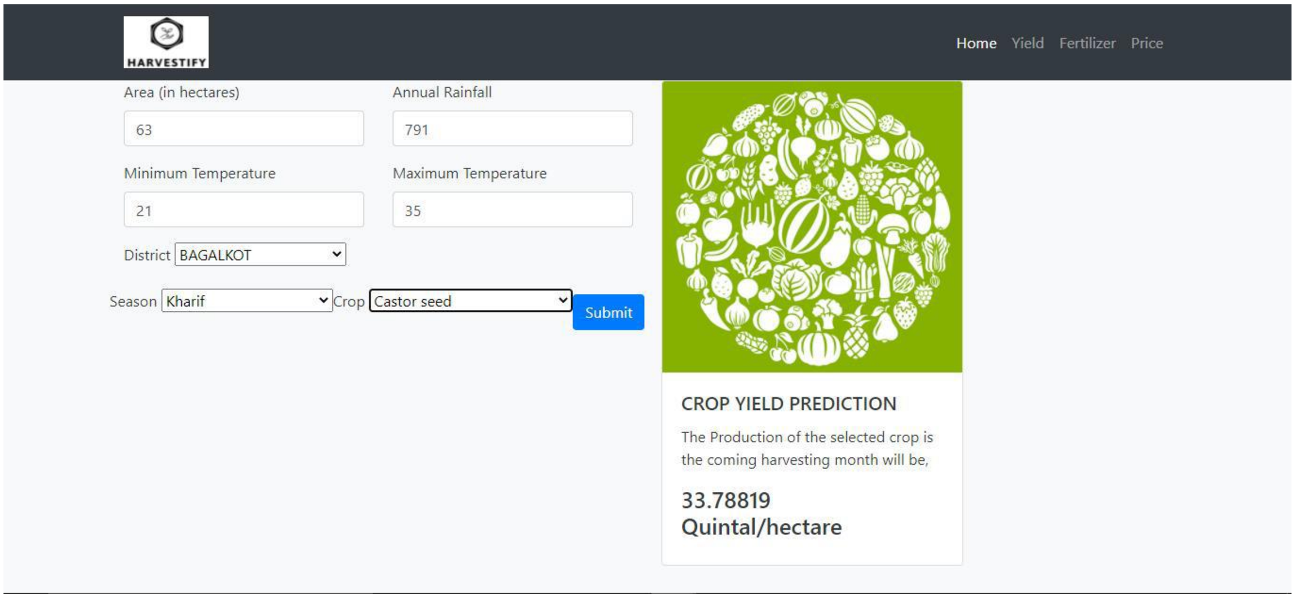
Test your soil health here and get detailed fertilizer Recommendation to enhance your soil condition.

WPI Trend Analysis

Visualize and analyze commodity specific, whole index price trends of the next 12 months in the country.

Fig 7.2: Module Navigation Links

Each of the link addresses to each of the module. The figure below shows the page for the yield predictor. It has a form that takes inputs required by the neural network to predict a yield. All the form inputs must be filled in order to obtain a result.



Area (in hectares): 63

Annual Rainfall: 791

Minimum Temperature: 21

Maximum Temperature: 35

District: BAGALKOT

Season: Kharif

Crop: Castor seed

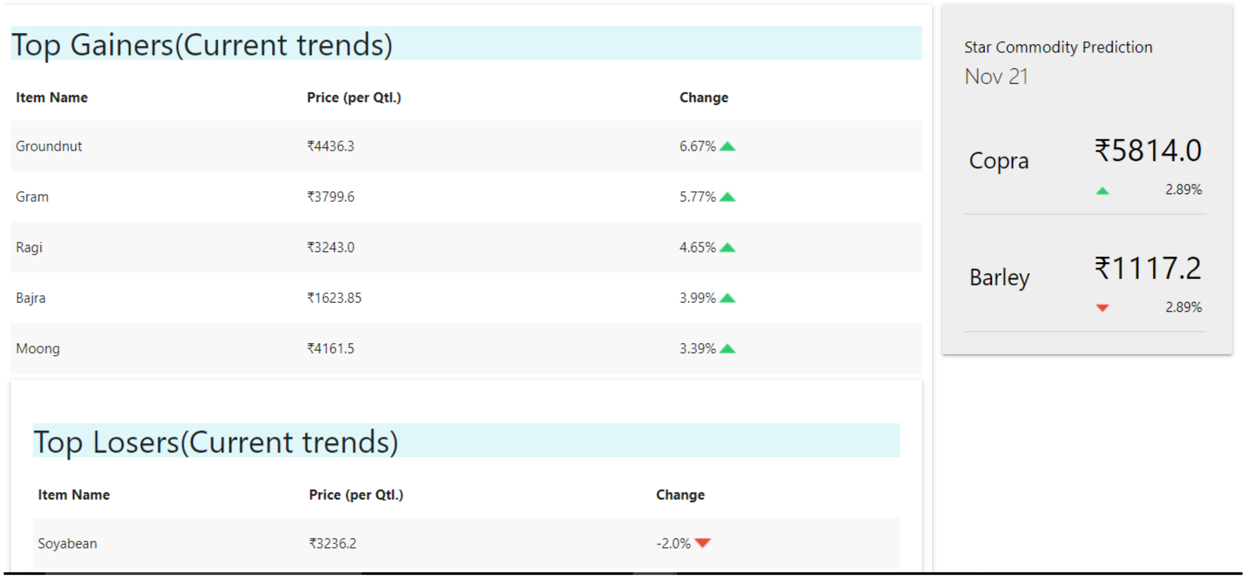
CROP YIELD PREDICTION

The Production of the selected crop is the coming harvesting month will be,

33.78819 Quintal/hectare

Fig 7.3: Yield Predictor Page

The next three figures show the Whole Price index analysis. The whole price trend analysis welcomes the user by first showing the top gainers and bottom most gainers on the next 6 months, and provides a slide show card that rotates to show the rise and fall of the WPI in the next 6 months.



Top Gainers(Current trends)

Item Name	Price (per Qtl.)	Change
Groundnut	₹4436.3	6.67% ▲
Gram	₹3799.6	5.77% ▲
Ragi	₹3243.0	4.65% ▲
Bajra	₹1623.85	3.99% ▲
Moong	₹4161.5	3.39% ▲

Top Losers(Current trends)

Item Name	Price (per Qtl.)	Change
Soyabean	₹3236.2	-2.0% ▼

Star Commodity Prediction

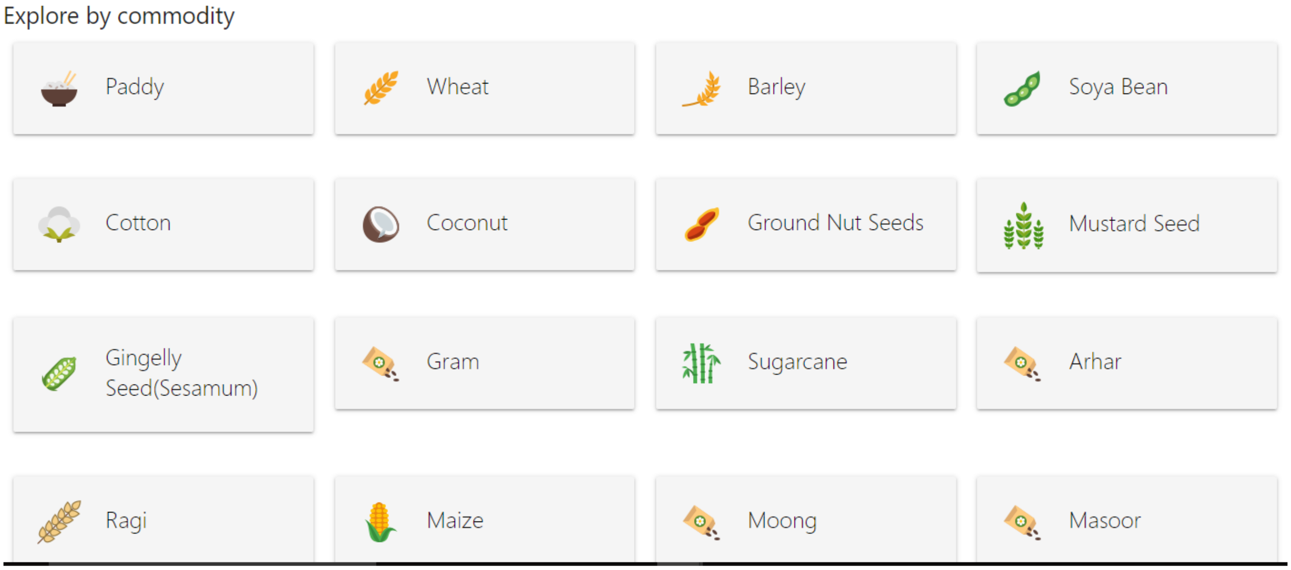
Nov 21

Copra ₹5814.0 ▲ 2.89%

Barley ₹1117.2 ▼ 2.89%

Fig 7.4: Whole Price Index Analysis Landing Page

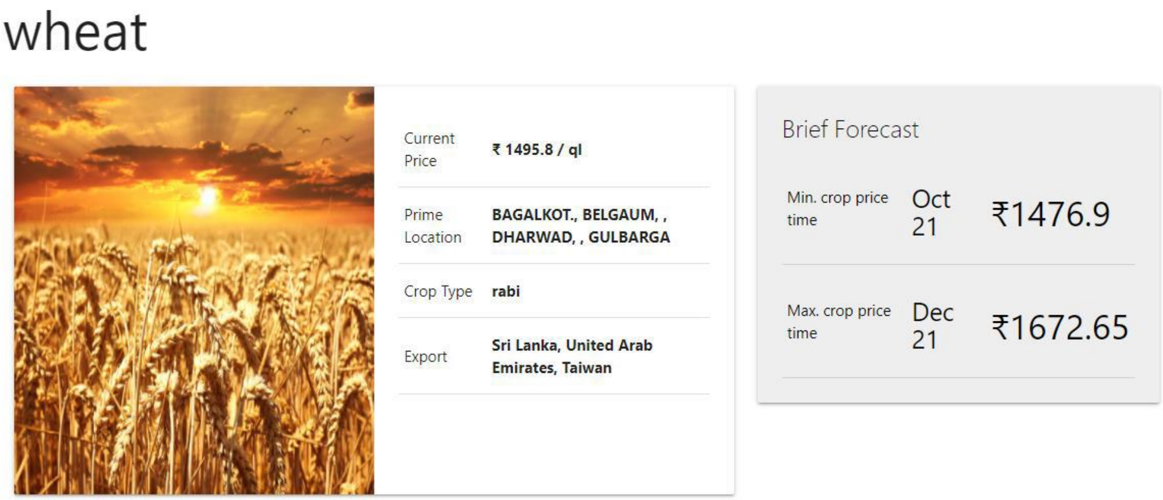
The user can also view the rise or fall of the WPI for each commodity. The top commodities or crops grown in Telangana are listed in the following sub-section. Each of the link takes the user to a detailed information of the profit analysis of a crop.



Explore by commodity

Fig 7.5: Commodity Wise Navigation of WPI Analysis

As said the user gets a detailed information on any crop via the "explore by commodity" tab. The detailed page shows the import and export details of a crop and what was their price is the previous years, as shown below.



wheat

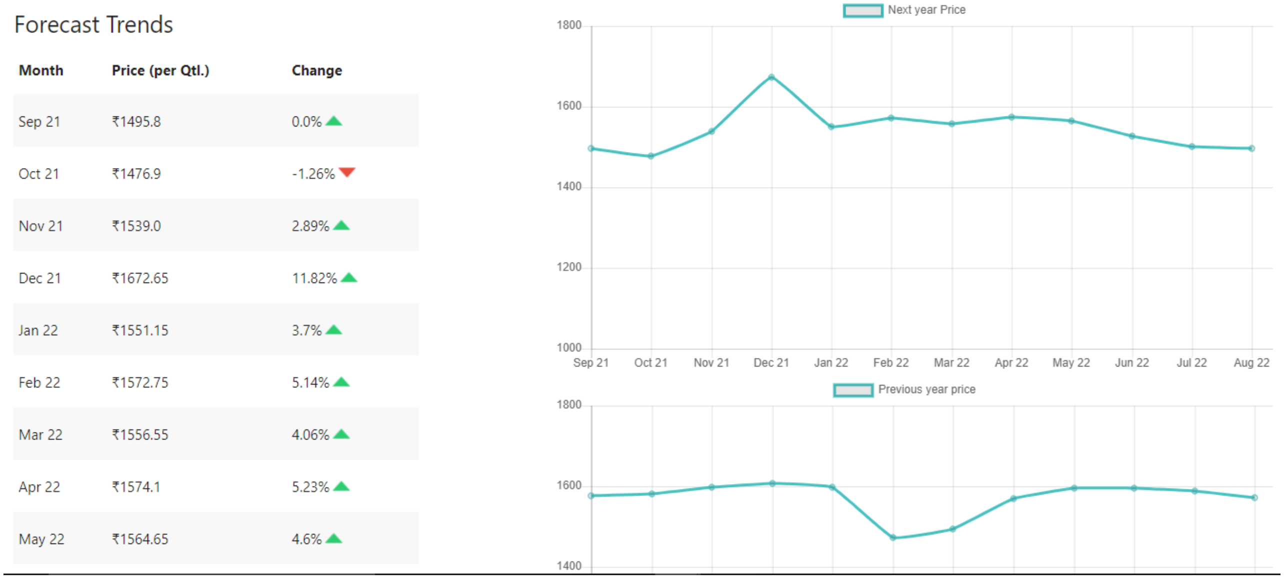
Current Price	**₹ 1495.8 / ql**
Prime Location	**BAGALKOT., BELGAUM, , DHARWAD, , GULBARGA**
Crop Type	**rabi**
Export	**Sri Lanka, United Arab Emirates, Taiwan**

Brief Forecast

Min. crop price time	Oct 21	₹1476.9
Max. crop price time	Dec 21	₹1672.65

Fig 7.6: Detailed Crop WPI Analysis Section

The final page shows the WPI Analysis graphically using ChartJS. It iterates over the decision tree algorithm to provide a 12 month price prediction of a graph using their base year price and the previous year trends.



Forecast Trends

Month	Price (per Qtl.)	Change
Sep 21	₹1495.8	0.0% ▲
Oct 21	₹1476.9	-1.26% ▼
Nov 21	₹1539.0	2.89% ▲
Dec 21	₹1672.65	11.82% ▲
Jan 22	₹1551.15	3.7% ▲
Feb 22	₹1572.75	5.14% ▲
Mar 22	₹1556.55	4.06% ▲
Apr 22	₹1574.1	5.23% ▲
May 22	₹1564.65	4.6% ▲

Fig 7.7: Graphical View of WPI Analysis

8. CONCLUSION & FUTURE SCOPE

In the current work, we tend to collected multiple datasets and performed applicable feature engineering to make one supply of information that accounts for all of the essential options to assist model correctness. to produce a comparison study of our neural network model' performance, we utilize a similar dataset to coach 3 further regression models, namely, the Multinomial Regression model, the Random Forest regression model, and also the support vector regression model. All three models' performance was evaluated victimization the same 2 measures delineate above: mean absolute error and R-Squared score.

It had been clearly seen within the graphical analysis that Neural Network not simply outperformed alternative classical Machine Learning algorithms in terms of Accuracy that was found to be 96.24%, however additionally was ready predict the result with minimum Error. A non-linear manner of deciphering the affiliation is important to demonstrate the interactions between the factors impacting crop production. because of the complexness of the factors impacting crop production, a linear technique reminiscent of regression toward the mean was deemed inadequate as an example the interactions between the parts and crop yield. For statement agricultural yield, ANN was thought to be a viable various to standard regression methods. A neural network not solely predicts non-linear correlation successfully; however, it may acknowledge sophisticated patterns in data and train applicably, one thing most ancient approaches fail to do.

The subsequent work can also be extended by victimization appropriate hardware to dynamically fetch the info from the attributes that have an effect on the crop yield reminiscent of soil, temperature, precipitation and rainfall. statement using Remote Sensing data can also be improved so as to eradicate the effort of handling the static data. However, these satellite pictures are often utilized in associations with the in-land information to produce a brand new dimension to the subsequent work.

9. BIBILOGRAPHY

[1] D. J. Reddy and M. R. Kumar, "Crop Yield Prediction using Machine Learning Algorithm," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1466-1470, Doi: 10.1109/ICICCS51141.2021.9432236.

[2] R. J. Brooks, et al., "Simplifying Sirius: sensitivity analysis and development of a meta-model for wheat yield prediction," European Journal of Agronomy, vol. 14, pp. 43-60, 2001.

[3] Gabhane Srushti, Shaikh Naushinnaaz, Sadavarte Shivani, Khan Huda, A.I. Waghmare, “Crop Yield Prediction to maximize profit using Machine Learning”, International Research Journal of Modernization in Engineering Technology and Science Volume:02/Issue:06/June-2020

[4] “Correlation Of Climatic Factors With Cereal Crops Yield: A Study From Historical Data Of Morang District, Nepal”, Badri Khanal, The Journal of Agriculture and Environment Vol: 16, June 2015

[5] Agro based crop and fertilizer recommendation system using machine learning, Preethi G, Rathi Priya V, Sanjula S M, Lalitha S D, Vijaya Bindhu B, European Journal of Molecular & Clinical Medicine ISSN 2515-8260

[6] Rice Crop Yield Prediction Using Artificial Neural Networks, Nikita Gandhi, Owais Petkar, Leisa J. Armstrong, 2016 IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR 2016)

[7] Rice Crop Yield Prediction Using Artificial Neural Networks, Nikita Gandhi, Owais Petkar, Leisa J. Armstrong, Amiya Kumar Tripathi, 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)

[8] Wheat Yield Prediction: Artificial Neural Network based Approach, Muhd Khairulzaman Abdul Kadir, Mohd Zaki Ayob, Nadaraj Miniappan, 2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T).