A Final report for a dissertation that will be submitted in partial fulfilment of a University of Greenwich Master's Degree

MSc Project Title

Crop Prediction Using Machine Learning Algorithms

Name: Maulikkumar Rakeshbhai Patel

Student ID: 001207677

Program of Study: MSc Computer Science

Module Leader: Dr. Tatiana Simmonds

Module Supervisor: Dr. Bernadette M Byrne

Submission Date: 30/04/2023

Word Count: 12275

ACKNOWLEDGEMENT

I would like to thank my supervisor for their consistent support and guidance during the running of this project. Furthermore I would like to thank the Kaggle for their data collection which has been used in this crop prediction project. Finally, I am grateful to “Statista” for their statistical analysis.

Abstract

This dissertation gives information about Crop classification and prediction based on the "Machine Learning" technique. Nowadays in the agriculture sector also, Machine Learning is used widely. Different kinds of algorithms are there that help Machines to understand the behaviour of the data. After that machine starts analysing the data and understands their characteristics to provide a conclusion. For this research three different datasets have been selected that are associated with several features like Crop, State, Cost of Cultivation (`/Hectare) A2+FL, Cost of Cultivation (`/Hectare) C2, Cost of Production (`/Quintal) C2, Yield (Quintal/ Hectare), temperature, humidity, rainfall, soil, production, etc. From these features Machine Learning algorithms can learn and provide various kinds of decisions that can help farmers to increase productivity. In a very simple word, ML helps to take perfect decisions based on certain conditions so that there will be a very high chance of positive outcomes.

Different techniques of "Machine Learning" have been demonstrated so that pertinent outcomes can be derived. There has been the implementation of the KNN classifier and Naive-based classifier algorithm so that an appropriate form of prediction is possible after implementing this procedure. Also, there has been an adaptation of the SVM algorithm so that appropriate accuracy is incurred after the application of this method.

However, KNN is selected as the best classifier as the best accuracy is implemented after applying this approach. Rainfall has been selected as a significant factor that can yield positive production of crops. Therefore, it has been identified as an effective approach and hence, these procedures are adopted after applying this strategy. There are ideas of appropriate standards for seeds after applying this method. Appropriate methods are involved after the application of procedures of irrigation. Technologies are also understood for the derivation of necessary opportunities.

Table of Contents

Chapter 1: Introduction

1.1 Introduction

“Machine Learning” is depicted as an essential technology that is necessary for supporting several decisions and prediction of data. In this research paper, this technique is significant for classification and prediction of crop. Various kinds of algorithms are applied for classification and prediction that can allow yield of crops. Moreover, there are probabilities to check revenues and costs that are linked with cultivation. Hence, these are described as one of the most significant approaches that can cause enhancement of crop productivity.

1.2 Background

There has been use of software for incorporation of innovative procedures that are undertaken after the involvement of this method. Moreover, these tools and techniques can help in derivation of new crops that are accumulated after incorporating this specified procedure. It has been reported that agricultural sector is accountable for giving support to growth of economy of country. Therefore, this field is significant in suitable progress of economy that is associated with this country.

Hence, there has been support of this ML technology that is associated in development of production of crops. There is utilisation of “Naive-Bayes classifier” for accomplishment of techniques that are associated with techniques of yielding crops. Moreover, “KNN algorithm” is used for carrying out this specified technique. As per the opinion of Liakos *et al*. (2018), a detailed analysis about all techniques is followed for selection of appropriate techniques for yielding crops. Additionally, yield of water is essential for giving depiction regarding technologies that are implemented after making appropriate predictions. There is also combination of technologies that are associated with robotics that can support production in this particular field.

1.3 Rationale

Harvesting in fields can become effective after accomplishment of procedures that are connected with “machine learning” (Pallathadk *et al.* 2021). Therefore, abilities are enhanced after

monitoring all techniques and these are presented as one of the most essential methods that are presented with this technology.

It is depicted as an issue in present situation because this field is contemplated for having contribution to economy of chosen country. Hence, this issue is compulsory to be abolished so that there is implementation of practices that can involve ML technologies to be practiced in this situation. These are depicted as effective approaches that are enabled so that implementation of technologies is becoming enabled for involvement of necessary measures that are effective in this situation. As per the statement of Pandith *et al*. (2020), there are essential elements that are associated with a technique that can have effective usage in this specified field. Hence, evaluation of methods is relevant in concerned procedure.

Moreover, information is acquired regarding usage of fertilizers can support yielding crops after adopting this specified procedure. Moreover, capacities are enabled that are affected after evaluation of this approach (Setiadi *et al.* 2020). This ***dissertation gives information*** about latest technology that is "Machine Learning" based on that crop can be classified with certain parameters as well as crop can be predicted. It will be very much helpful for the farmers or the agriculture department to take any decision associated with crop production with following several criteria's like temperature, humidity, rainfall, soil ingredients etc.

1.4 Aim

Aim of the research paper is classification and prediction of crops after application of several "Machine Learning" algorithms.

1.5 Objectives

- Understand about the ML techniques and its significance in agriculture sector
- To identify "Naive Bayes classifier" of "machine learning" to assume costs for production of crops
- To analyse elements such as transaction, minimum amount of support for making estimations regarding prices of products
- To evaluate conditions that are suitable for application of "KNN algorithm" so that prediction about revenues can become effective

- To recommend best methods so that prices of products, costs of cultivation are effective for making assumptions regarding costs

1.6 Questions of research

1. What is ML and why it is significant in agriculture?
2. What is the algorithm that can forecast price of crops?
3. What are various elements that are brought into contemplation for making estimations regarding prices?
4. What are certain techniques in ML that are associated with assumption of revenues?
5. What are recommendations regarding factors that are taken into contemplation for implementation of algorithms of "machine learning" for ascertaining profits?

1.7 Significance of this research

This research can enable to know about various applications that are associated with "machine learning" technology. Hence, acquaintance of these technologies is accomplished so that significance of these methods is implicated for defining processes of this research. There are effects of alterations of climatic situations that are having a significant effect in crop production. Hence, ML is depicted as a valuable technique that can be implemented for adaptation of necessary decision at the time of cultivation. Therefore, this research is contemplated as one of the effective research. It is associated with understanding of all elements that can support in constructive progress for determination of agriculture strategies. This chapter gives description about essential tasks that are compulsory for designing effective classifier algorithms that are suitable for designing strategies for production. Hence, these are selected as essential approaches that are associated for evaluation of large numbers of strategies that can help in escalation of profits and assuming techniques for growing crops.

1.8 Scope

This research can be necessary for making validation about ML technology that is evaluated after designing this research procedure. Hence, these are depicted as essential measures that are associated after application of these procedures.

1.9 Research limitations

Limitations are absence of sufficient time. In absence of ample amount of time, it becomes difficult as a researcher to gather all information regarding all tasks of research. Therefore, not all data regarding “Machine Learning” technologies are accumulated after designing this procedure of research. Moreover, absence of sufficient budget is also identified as a major limitation that has been associated with this research.

1.10 Structure of dissertation

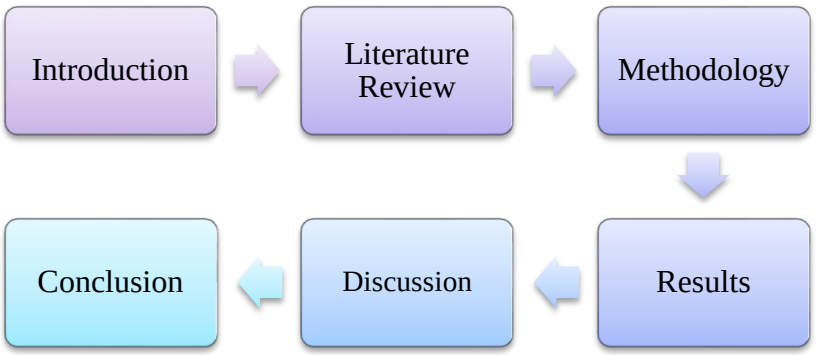

Figure 1.3: Dissertation structure

Chapter 2: Literature Review

2.1 Preface

This chapter gives description about ML techniques that are followed for undergoing predictions for yielding of crops. All these information is adopted from research papers, articles, journals that are published previously. Gaps in literatures are also depicted in this chapter.

2.2 Role played by "Machine learning" in predicting crop production

Technologies are incorporated so that ML can play active role in making specification of production of crops. There is use of supervised strategy in cases of determination of primary issues that are associated with this situation. Moreover, techniques of regression are incorporated so that these procedures are implicated with definitive approaches. Issues in this process can incorporate incomplete data that are used for computing and hence, irrelevant outcomes are noticed.

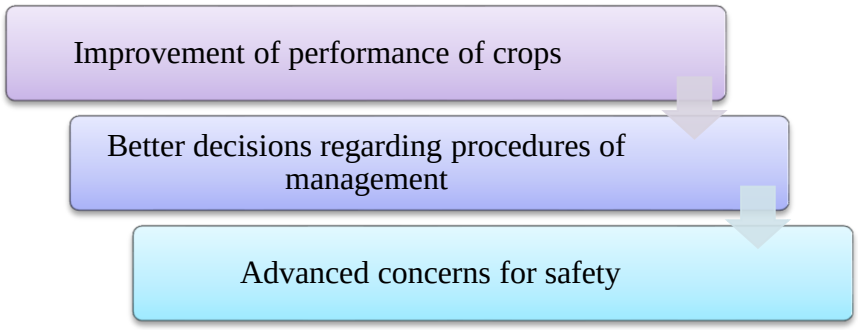

Figure 2.1: Advantages of incorporating "machine learning"

(Source: Chandana and Parthasarathy, 2022)

There are issues of presentation of data that are associated with procedures of training. Moreover, inaccurate consequences are visible after getting sufficient information that are involved in this technique. As per the statement of Chandana and Parthasarathy (2022), optimal uses of all kinds of fertilizers are relevant after adaptation of suitable techniques that are implemented in this specified technique. There are series of steps that are followed in pre-processing data involved in this procedure. All these steps are illustrated in this table.

Name of steps	Significance
Pre-processing of data	This step is noticed for pre-processing information that is suitable for research technique. Ingestion of raw information can become effective after incorporation of this procedure.
Cleaning of data	Automatic detection of cortex and scrubbing outliers are consistent after involving this step.
"Feature engineering"	There is identification of relevant features that are quantifiable after implementing this approach.
Selection of models	Validation of "training and testing models" is relevant after depiction about this procedure.
Generation of prediction	There us involvement of a "regression model" for understanding probabilities that can enable suitable strategies that supports this procedure.

Table 2.1: Description of steps in "machine learning"

(Source: Kaur *et al.* 2019)

Hence, these are identified as necessary approaches and advantages are implicated for determination in this specified field.

2.3 "Naive Bayes classifier" for understanding features of crop cultivation

There are probabilities for escalation of accuracies that are associated with techniques of producing crops. Improvement is possible in efficiency and it has been identified as an essential

technique for enhancing accuracy of yielding crops. Systems are recommended so that prediction regarding analysis can become effective. As per the statement of Kaur *et al.* (2019), classifiers are taken into contemplation so that there is development of a ***"neural network"*** for understanding features that can have positive significance in yielding crops. However, there is opinion of other authors who have said different things about cultivation. As per the statement of Medar *et al.* (2019), there are specified techniques of regression that are involved for getting advantages for effective use of this algorithm. Hence, prediction regarding techniques is compulsory for selection of relevant procedures and selection of pertinent procedures are related with incorporation of these approaches. Moreover, as per the statement of Archana and Saranya (2020), recommendations are available regarding fertilizers after involvement of this specified technique. Therefore, these are depicted as essential features that are taken into contemplation so that there are noteworthy consequences about making certain kinds of assumptions.

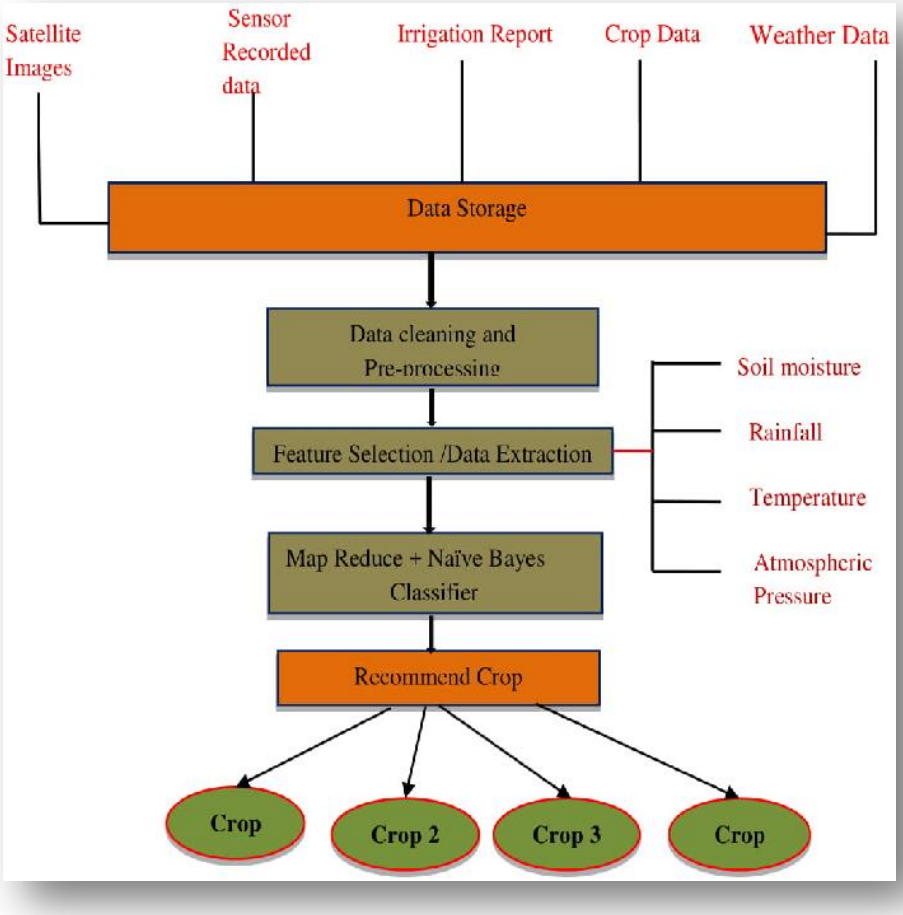

Figure 2.2: Application of "Naive Bayes classifier" for progress in crop cultivation

(Source: Kaur *et al.* 2019)

Therefore, there are probabilities of identification of procedures that are relevant for making appropriate strategies that can yield crops. Hence, these are reported as essential measures that are adopted for getting crops that are implicated in this procedure. It has been identified that structure of soil. There is advancement of approach of "data mining" that can have suitable growth after implementing this approach.

Implementation of the K-nearest neighbour Algorithm

As per the viewpoint of Abbad Ur Rehman *et al*. (2021), the algorithm of the "K-nearest neighbour Algorithm (KNN)" tends to be the non-parametric algorithm that is it rejects any kind of assumptions with the data sets that are already undecayed. Moreover, it is also known as the "lazy learner algorithm" because it does not have any kind of impact from the "training set" instead it stores the data during the tenure of classification and performs the action on the simulated set of data sets.

On the other hand, in case the data exists in a new point for the classification thereby the nearest neighbour of "K" is searched out from the "training data" sets. Accordingly, as per the report of Potratz *et al.* (2021), the procedure seems to be induced by calculating the overall distance between all sets of data within the available set of data sets as well as the input varies simultaneously. Most importantly, within the "K-nearest neighbour Algorithm (KNN)" the distance measurement is done by using the features such as "Euclidean distance", "Minkowski distance" and the "Manhattan distance". Therefore, according to the report of Tembusai *et al.* (2021), within the "K-nearest neighbour Algorithm (KNN)" the classifier of distance refers to the vivid simplicity and presumes the smallest distance between two points and the similarity segments within the two points. Moreover, the columns associated with the Therefore, within the "K-nearest neighbour Algorithm (KNN)" distance classifiers the columns act as the dimension. Lastly, the norms of the Manhattan distance use the analogy that systematically provides the distance metrics within its internal grid. The "Euclidean distance" within the space refers to the length of the line segments between the two determined points. Apart from that, it is the square

root of the associated sum of the differences of squares between the associated elements of the main two vectors. Furthermore, as per the viewpoint of Lin *et al.* (2021), the "Minkowski distance" appears to be the generalised distances with the metrics across the "Normed Vector Space". Thereby, the associated "Normed Vector Space" correlates to the systematic procedure of collection of the space in which the associated points have been operating as well as functioning with the help of a simulated function. Therefore, the associated vectors within the aspects must be a positive number that is the number should be greater than 0.

2.4 Prediction about yield of crops

There is incorporation of crops and elements are associated with parameters of meteorological characteristics. As per the opinion of Thompson *et al.* (2020), limits that are accustomed for adaptation of techniques of computation are involved for understanding techniques of "deep learning algorithms". Hence, these are depicted as certain approaches that are suitable for implementation of features that are enabled after adoption of this technique. Models are involved that can support models and significant features that can involve appropriate production of crops are enabled after depicting this technique.

Moreover, extraction of characteristics is featured after implementation of strategies that can incorporate essential feature in this context. Therefore, this procedure is essential for making effective decisions regarding strategies and tools that are essential for production of crops. Hence, these are derived as notable measures that are involved after making decisions in this field. There are applications of techniques of regression and correlation after making assumptions regarding operational procedures.

Consistent features are enabled after understanding procedures that are involved after a detailed identification of this technique. As per the statement of Sun *et al.* (2021), there are probabilities of getting consistent consequences that are enabled after making procedures that are associated in this situation. Hence, these are depicted as one of the most essential theories that are predicted for understanding all features that can have profound effects in controlling production of technique. Moreover, there are probabilities of getting progress that can support production after implementing this concept. Therefore, these are depicted as essential features that can support crop production.

2.5 Models and theories

2.5.1 Incorporation of "computational learning theory"

This theory is implemented for understanding processes of computation that are involved with this strategy. Principles that are relevant in this feature is involved so that a complete understanding is become enabled after involving this approach. Moreover, depicting outputs is an essential feature that is relevant to this technique. As per the opinion of Sun *et al.* (2021), there are consistent techniques that are evaluated after giving descriptions in this relevant field.

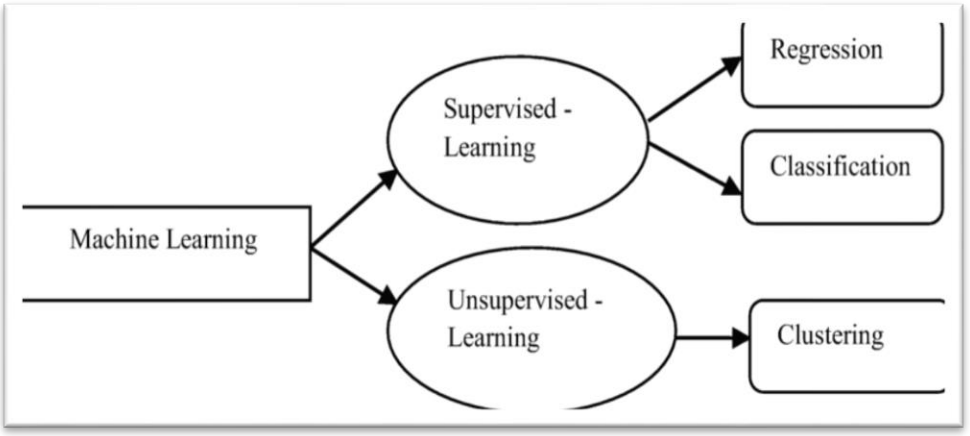

Figure 2.3: Characteristics of "computational learning theory"

(Source: Sun *et al.* 2021)

Hence, all elements that are compulsory to be acquainted are enabled after taking all these elements into contemplation. Moreover, prediction regarding all levels of class is enabled after application of this field. Ideas regarding "Euclidean distance" are enabled after enabling ideas in this approach.

2.5.2 Theory of "machine learning"

ML theory is used as a complete procedure that involves several stages those are used for fulfilling approaches. Hence, there are probabilities of standardisation of practices that are implication after making a complete incorporation of this strategy. Automation of flows of work is identifiable after a complete analysis of flows of work that are implicated in this procedure. Features are understood and outputs are predicted after inclusion of respective features that are

brought into contemplation in this circumstance. Moreover, automation of all kinds of flows of work is enabled for pre-processing several techniques of information.

Figure 2.4: Aspects of "machine learning"

(Source: Sun, 2019)

There are provisions for speeding up data and these are evaluated as notable opportunities that are implicated after involvement of this situation. Moreover, suitable algorithm is applied that are integrated so that notable consequences are acquired after evaluation of this structure. It has been observed that a large number of opportunities are incurred after application of this concept and progress in noticed in knowing production of several kinds of crops.

2.6 Conceptual framework

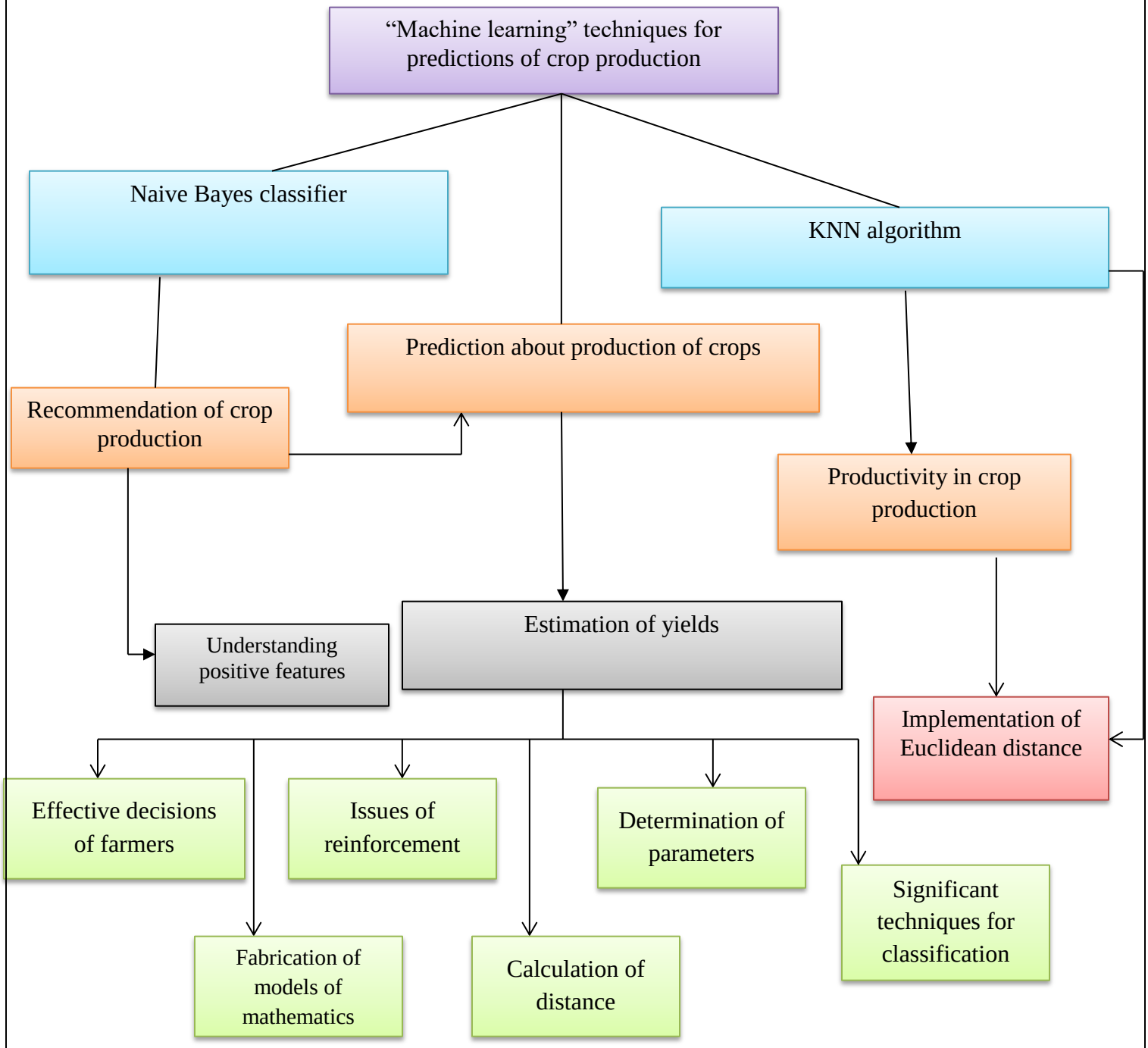

Figure 2.5: Conceptual Framework

(Source: Self developed)

2.7 Gaps in literature

In journals, that are published previously information is given about advantages of techniques of "machine learning" so that there are chances of assuming concepts of production of crops. However, practical instances in which these strategies are applied are not procured from these journals. Hence, these are depicted as one of the most noteworthy disadvantages that are present in these journals. Additionally, limitations are also involved in approaches of gathering data and these are contemplated as necessary approaches that are implicated in supporting these decisions.

2.8 Summary

It is summarized that there are certain features that are essential to be accounted for making decisions about creation of crops. These are depicted as some of the notable tools that are involved for identification of necessary approaches. Decisions are supported for yielding crops and these are identified for making assistance in assumptions of productions. Positive outcomes are noticed after implementation of this strategy that is associated with cultivation after knowing "machine learning" techniques.

Chapter 3: Methodology

3.1 Introduction

The study is based on crop-yielding production with the application of the “Machine learning” technology. Moreover, this chapter consists of a systematic illustration of the associated methods of the research study along with the various tools as well as techniques for coherently completing the research study. Therefore, this also focuses on the mitigation of the barriers associated with the research study with the implementation of the appropriate techniques to maintain the objective of the study.

3.2 Research Design

The research design appears to be the most significant spectrum because it describes the methods as well as the technique for the research study. Most importantly, according to the report of Zhou *et al.* (2021), it helps in streamlining the research in a coherent direction in order to attain the main goal of the study mitigating the barriers. There are three key types of research designs for the completion of the research study and those are “explanatory, exploratory, and descriptive”.

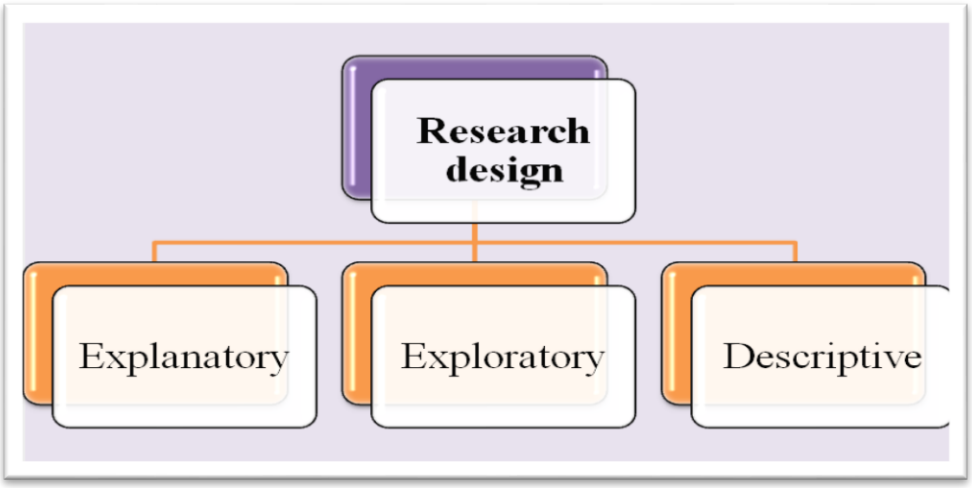

Figure 3.2: Research Design

(Source: Self-created)

The above figure displays the three key research designs and within the research of crop-yielding production with the application of the ML technology. The research is dependent on descriptive research design. The main reason associated with the same is because it helps in delivering a vivid description of the methods as well as data applied in the research study. Moreover, as per the viewpoint of Zhang, (2022), it can be stated that the descriptive research design also played the keenest role in demonstrating the theoretical background of the research in a coherent manner.

3.3 Research Approach

The research approach is the most appropriate aspect that helps in order to ensure that the purpose of the study is satisfied. Moreover, the key types of research approaches are inductive and deductive.

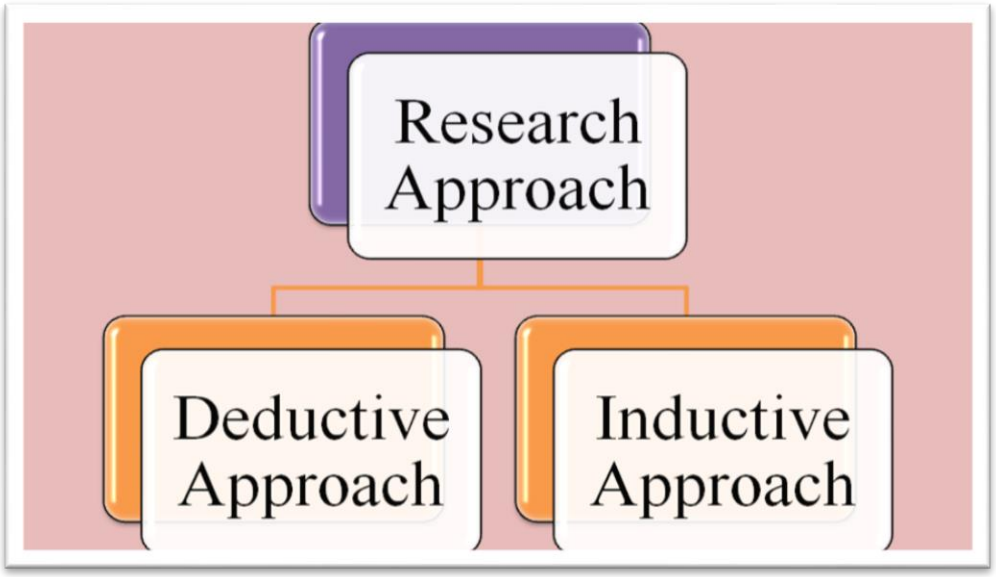

Figure 3.3: Research Approach

(Source: Self-created)

The research approach that has been utilized for the research study regarding crop-yielding production with the application of ML technology is deductive. As per the report of Rumsey *et al*. (2021), this helps to maintain the criticality of the research study and the statistical data

regarding the study are implemented by maintaining the authentic layout. Moreover, it also systematically enhanced the development of the hypothesis and thereafter validation as well as testing of the same. Apart from that, according to the report of Goldstein, (2022), it helps to maintain the statistical layout of the research study so that the implementation of the gathered data is done smoothly by mitigating the barriers and literature errors. Lastly, according to the report of Pereira *et al.* (2021), it helps to systematically enhance the development of a profound conclusion whereas the inductive research approach is time-consuming and less statistical.

3.4 Tools and Techniques

The main purpose of the research study is crop prediction by applying the ML technology. The advantages of the systematic incorporation of machine learning in crop prediction have been discussed within the study. There are issues regarding the presentation of data that are associated with the procedures of training and even inaccurate consequences can be visible after getting sufficient information is involved in this technique (Zhou *et al.* 2022).

Machine Learning:

In order to proceed on the ML work, python programming language has been used. That is one of the most efficient programing languages for ML, AI, data analysis, etc. Moreover, within the research, ***"Jupyter notebook"*** is used with the goal of developing program. It is an open-source software, open standards as well as services for interactive computing across various platforms. Similarly, it is the original set of web applications for the creation and thereafter sharing of the data regarding crop prediction by applying the ML technology and "Jupyter notebook" offered a simple, streamlined and document-centric experience regarding the research study. Apart from that, within the study KNN, SVM, "Naive-Bayes" classifier has been used for the accomplishment of techniques that are associated with techniques of yielding crops.

Python with Jupyter Notebook:

On the other hand, the application of the "Jupyter notebook" allows for compiling the associated aspects of the data projects in a single place and thereby with the application of the same the visualization has been created. The ***Python*** language has been used and it emphasises the readability of the code with the use of significant indentation (Zhang, 2022). It supported the paradigms of the multiple sources of programming and the structure of object-oriented,

structured and functional programming. The libraries that have been used within the research study are ***Jumpy, Pandas, Matplotlib, and Seaborn.*** Jumpy has been used for the purpose of programming the vivid range of mathematical operations within the arrays. Moreover, it also correlates to the various significant functions in order to work within the domains that are pertaining to the matrices such as "Fourier transform, and linear algebra".

Python Libraries:

The Pandas are the library that refers to the rendering of utilising the high performance “data structures as well as other associated tools of the systematic analysis of the extracted data of the research study. However, the module is subjected to running on the top-notch portion of the library that is functionally known as “***Numpy***”. On the other hand, according to the report of Rumsey *et al*. (2022), “matplotlib” is the systemic visualisation library that is focused on the layout of developing two-dimensional plots with respect to arrays.

Moreover, within the nodes of the “matplotlib” there appear to be the “multi-platform” libraries that are focused on the pertinent of the data visualisation build upon the arrays of ***NumPy***. Most importantly, it has been sequentially designed in order to work with the border stack known as “***SciPy***” (Goldstein, 2022). Lastly, “Seaborn” appears to be the specific library that consists of the top position of the library belonging to “matplotlib”. Therefore, it appears to have kind of similarities related to “matplotlib” as per the design layout of the visualisation. However, “Seaborn” appears to be better as well as significant as compared to matplotlib regarding features of operations as well as functions.

3.5 Data Collection

The data regarding the research study of crop prediction by applying the “Machine Learning” technology have been gathered from ***Kaggle.*** This appears to be the crucial aspect regarding the systematic induction of collaboration amongst the users and thereby extracting the data from coherent sources in order to publish the data sets systematically. Most importantly, it helped in order to deliver immense computational power throughout the induced servers of the system. Apart from that, within “Kaggle” there tends to be a significant list of data sets as well as the option in order to search by name option for specific data sets ingested (Pereira *et al.* 2021). Therefore, the table has been developed by extracting the data and within the data set, there are six columns which denote the type of crop, place, cost of cultivation per hector, cost of

production per quintal and the value of yield that is calculated by dividing the cost of production per quintal and cost of cultivation per hectare.

3.6 Data Analysis

The data regarding the research study has been gathered from various authentic sources that are focused on the foundational layout of the implementation of the algorithms of ML. Thereafter, the next step consists of ***data cleaning*** which consists of fixing or even removing the unnecessary incorrect, corrupted and inappropriately formatted data within the data sets. This plays the keenest role in mitigating the barriers of unreliable algorithms and outcomes and even though it may look corrects (Sosa-Díaz *et al.* 2022). Thus it systematically established a significant layout of fixing the errors of the structures because of the filtration procedure regarding the unwanted outliers. Henceforth, the validation within the research study has been done by maintaining a coherent manner as per the desired outcomes at the initial stage of the research study.

The next step that has been done is ***data pre-processing*** which has been done in order to aesthetically prepare for the whole data sets regarding the operations of crop prediction using the ML. The step is induced by gathering the data in order to sequentially import the same into the software platforms.

This technique is vividly focused on converting the associated raw data into a set of clean data sets. Thus, at the initial layout for the project, the data are gathered from various sources and gathered in a raw format which may not be feasible for the analysis this has been done for enhancing the feasibility of the analysis of the research study (Ullah and Rafiq, 2022). Apart from that, it is the most crucial aspect in order to enhance the reliability as well as accuracy level of the research study by converting as well as eliminating the raw data so that the consistency level of the data can be improvised. Moreover, ***exploratory data analysis*** that has been used helped in the layout of enhancing the data analysis by inducing the visualisation process. It also helped in order to discover systematic trends as well as patterns so that the assumptions regarding the data sets can be checked with the help of statistical sequence as well as the graphical representation of the research study in a coherent manner.

The next step of data analysis consists of ***feature selection*** which has been done for reducing the inputs of the associated variables by using the relevant data so that the research can be completed

without any kind of noisy data. It consists of the procedure of automatically including the relevant feature of machine learning to solve the associated problems. The ***train test split*** is correlated with the estimation for the performance of the algorithms of machine learning that are related to the prediction of crop yielding. This helps in the design layout of enhancing the modulation as well as fastens the functionality of the project by preventing over fitting (Lee *et al.* 2021). Lastly, ***model development*** within the research study correlates to the data acquisition from the various available coherent sources and the processing of the data to build the model as per the requirements for the research study. Most importantly, it helps in the layout of recognising various types of patterns arrived during the tenure of the research study. This plays the keenest role regarding the enhancement of the algorithm so that it can be used as the sequential aspect regarding the completion of the research study systematically and these can also be constituted as the decision-making algorithms.

3.7 Selected Model:

KNN:

The "K-nearest neighbour Algorithm (KNN)" appears to be the simplest form of "Machine learning (ML)" algorithm that is based on the "supervised learning technique".

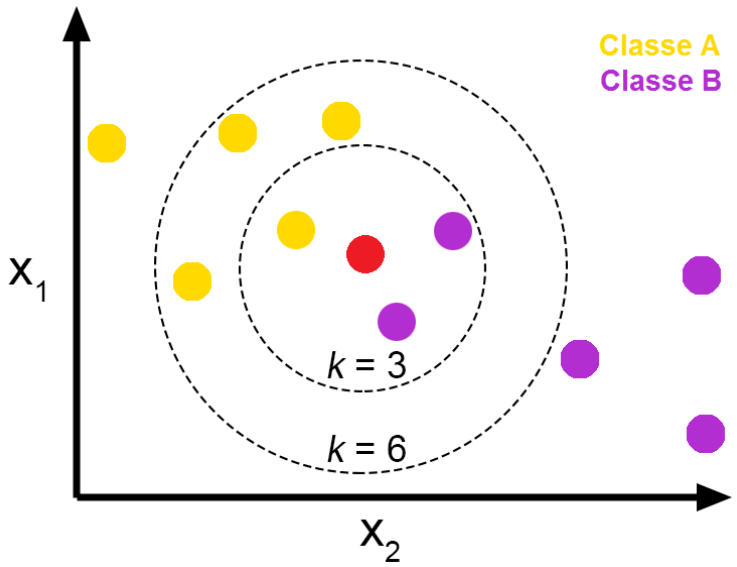

Figure 3.4: KNN working

Source: (towardsdatascience, 2018)

This is based on the assumption of the similarity within the new case data as well as the available set of data and thereby the procedure is induced of putting the new case into the category that appears to be the most symmetric with the available set of categories. Apart from that, according to the report of Wazery *et al.* (2021), the algorithm of the "K-nearest neighbour Algorithm (KNN)" tends to be systematically implemented for obstacles regarding the classification as well as the regression. The working modulation of the same is focused on the determination of the sole parameters known as "K" and correlates with the sequential quantity of the nearest neighbour.

Naïve Bayes:

An example of a classification algorithm based on Bayes' theorem is the naive Bayes classifier. naïve Bayes classifiers place a heavy (or naïve) emphasis on the independence of data point properties. Spam filters, text analysis, and medical diagnosis are just a few examples of where naive Bayes classifiers have found success. Since these classifiers are straightforward to implement, they see extensive use in machine learning.

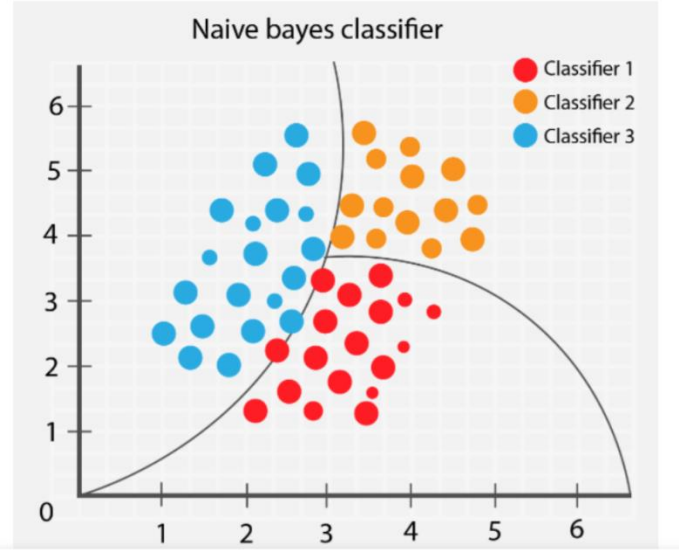

Figure 3.5: Naïve Bayes working

Source: (analyticsvidhya, 2022)

Using probability theory, a naive Bayes classifier assigns labels to data. Bayes' theorem is used by the Naive Bayes classification system. The essential idea behind Bayes' theorem is that fresh information may be used to revise an event's probability.

SVM:

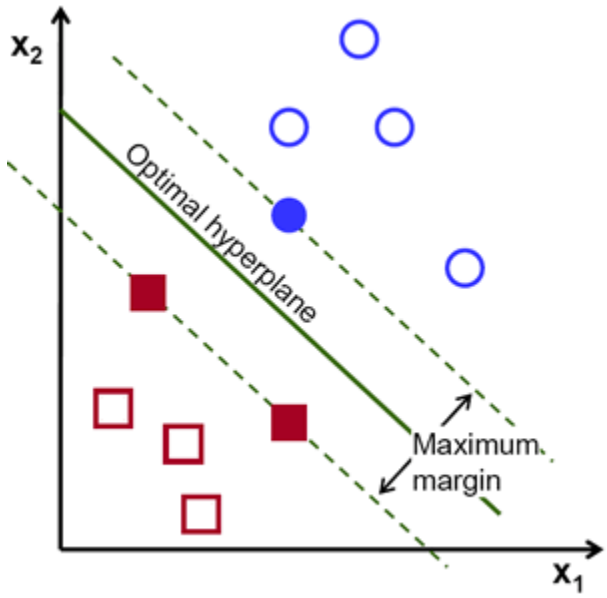

Figure 3.6: SVM working

Source: (analyticsvidhya, 2022)

A basic yet effective Supervised ML technique, the SVM may be used to construct both regression and classification models. Whether your data is linearly or non-linearly separable, SVM will perform exceptionally well. The support vector machine technique continues to impress even when given a little dataset.

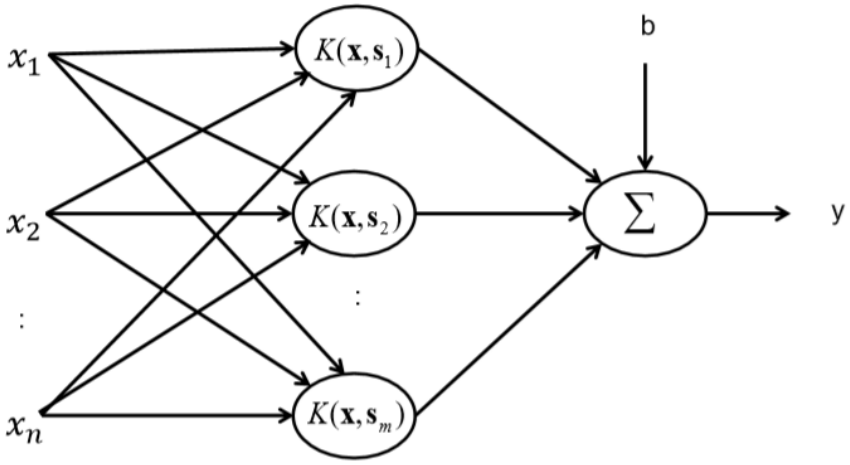

Figure 3.7: SVM Architecture

With the ability to perform linear and nonlinear classification, regression, and outlier identification, a support vector machine is a crucial and flexible machine learning technique. Although support vector machines (SVM) are used to solve both classification and regression problems, they are more commonly employed for the former. In comparison to competing classification methods, it requires significantly less processing power while providing comparable accuracy.

The implementation has been done based on these three ML algorithms.

3.8 Ethical Consideration

The highest, as well as significant ethical standards have been maintained within the research study the sources of the gathered data are cited within the study. Moreover, the gathered data are not used for any kind of commercial purpose or any kind of purpose rather than the research study. Moreover, the research study is subjected to any kind of arbitration or duplication of the data within the course of the research study. The issue regarding copyright as well as any kind of infringement has been systematically ingrained within the course of the research study.

Chapter 4: Result

4.1 Introduction

This chapter focuses explicitly on secondary quantitative data analysis through the help of various machine-learning techniques. The data set collected from the secondary source had comparable data on crop yield, various meteorological factors (rainfall, temperature), and pesticides. Through the help of the Naive-Bayes Classifier and K-nearest neighbor Algorithm (KNN), SVM, correlation analysis and heat map projection have been presented in this section. The correlation analysis through the help of an ML has directly predicted interdependence between crop yield (dependent factor) and temperature, rainfall, cost of production, and cultivation (predictors). In this regard, this section has delivered a critical evaluation of the crop yield under the influence of various economic and meteorological predictors.

4.2 Implementation

4.2.1 Data set description

Three specific data sets (Cost of production, Total production, and Rainfall) have been implemented through the Python Jupyter Notebook in terms of analyzing KNN, Naive Bayes classifier, and data visualization. Now, each of these data sets has its own unique characteristics. For instance, the first data set contained data on the cost of cultivation and total yield. Figure 1 below is explaining that there are seven major features such as crop, state, cost of cultivation(x2), cost of production, and yield within this first data set. However, the CROP and STATE feature represents the state wise crop production. This explains that the first data set is relevant for further analysis.

	Crop	State	Cost of Cultivation (`/Hectare) A2+FL	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)
0	ARHAR	Uttar Pradesh	9794.05	23076.74	1941.55	9.83
1	ARHAR	Karnataka	10593.15	16528.68	2172.46	7.47
2	ARHAR	Gujarat	13468.82	19551.90	1898.30	9.59
3	ARHAR	Andhra Pradesh	17051.66	24171.65	3670.54	6.42
4	ARHAR	Maharashtra	17130.55	25270.26	2775.80	8.72
5	COTTON	Maharashtra	23711.44	33116.82	2539.47	12.69
6	COTTON	Punjab	29047.10	50828.83	2003.76	24.39
7	COTTON	Andhra Pradesh	29140.77	44756.72	2509.99	17.83
8	COTTON	Gujarat	29616.09	42070.44	2179.26	19.05
9	COTTON	Haryana	29918.97	44018.18	2127.35	19.90
10	GRAM	Rajasthan	8552.69	12610.85	1691.66	6.83

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

	Crop	Production 2006-07	Production 2007-08	Production 2008-09	Production 2009-10	Production 2010-11	Area 2006-07	Area 2007-08	Area 2008-09	Area 2009-10	Area 2010-11	Yield 2006-07	Yield 2007-08	Yield 2008-09	Yield 2009-10	Yield 2010-11
0	Total Foodgrains	158.8	168.6	171.3	159.4	178.9	128.5	128.8	127.6	126.0	131.7	123.6	130.9	134.3	126.5	135.9
1	Rice	200.8	207.9	213.3	191.6	206.4	168.5	168.9	175.1	161.2	164.8	119.2	123.1	121.8	118.9	125.2
2	Wheat	131.6	136.4	140.1	140.3	150.8	115.0	115.2	114.0	116.9	119.5	114.4	118.4	122.8	120.0	126.3
3	Jowar	124.3	137.8	126.0	116.5	121.8	120.7	110.6	107.3	111.0	105.2	103.0	124.6	117.4	105.0	115.8
4	Bajra	136.4	161.5	143.9	105.4	167.9	94.5	95.1	87.0	88.5	95.6	144.3	169.7	165.4	119.0	175.8

Figure 4.1: Description of the data sets

(Source: Acquired from Jupyter Notebook)

4.2.2 Data pre-processing

	Cost of Cultivation (`/Hectare) A2+FL	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)
count	49.000000	49.000000	49.000000	49.000000
mean	20363.537347	31364.666735	1620.537755	98.086735
std	13561.435306	20095.783569	1104.990472	245.293123
min	5483.540000	7868.640000	85.790000	1.320000
25%	12774.410000	19259.840000	732.620000	9.590000
50%	17022.000000	25909.050000	1595.560000	13.700000
75%	24731.060000	35423.480000	2228.970000	36.610000
max	66335.060000	91442.630000	5777.480000	1015.450000

	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117

Figure 4.2: Mean, mode, and standard deviation of the datasets

(Source: Acquired from Jupyter Notebook)

Figure 2 is explaining the distribution of the data where maximum and minimum distribution can be seen as major measurement parameters. It can be seen that 25% of the data falls under the maximum yield value of 9.59quintal/hectare against the maximum cost of cultivation of 12774.41. Moreover, the highest yield value can be witnessed as 1015.45 quintal/ hectare against 66335.05 as the cost of cultivation. The standard deviation value below the mean value explains that there is a significant variance in the data set (Patil *et al*. 2020). Similarly, it can be seen that SD<Mean within these two data sets thus it can be said that data are distributed around the mean. Therefore, it can be said that data set used for the analysis purpose have accuracy, reliability and dependence characteristics.

4.2.3 Model used

Three particular ML models have been used in order to analyze these specific crop yield data sets. The first model as in the Naive Bayes model is known for its flexibility in handling linear numbers of predictors. This model is highly scalable with different types of data (Ghadge *et al*. 2018). Similarly, KNN Model offers opportunities for handling multiclass data sets. It should be mentioned that both of these models do not require much training time thus this research got faster. Finally, Support vector machines or SVM have been used in terms of delivering understandable margins in these data sets.

4.3 Result analysis

4.3.1 Visualization

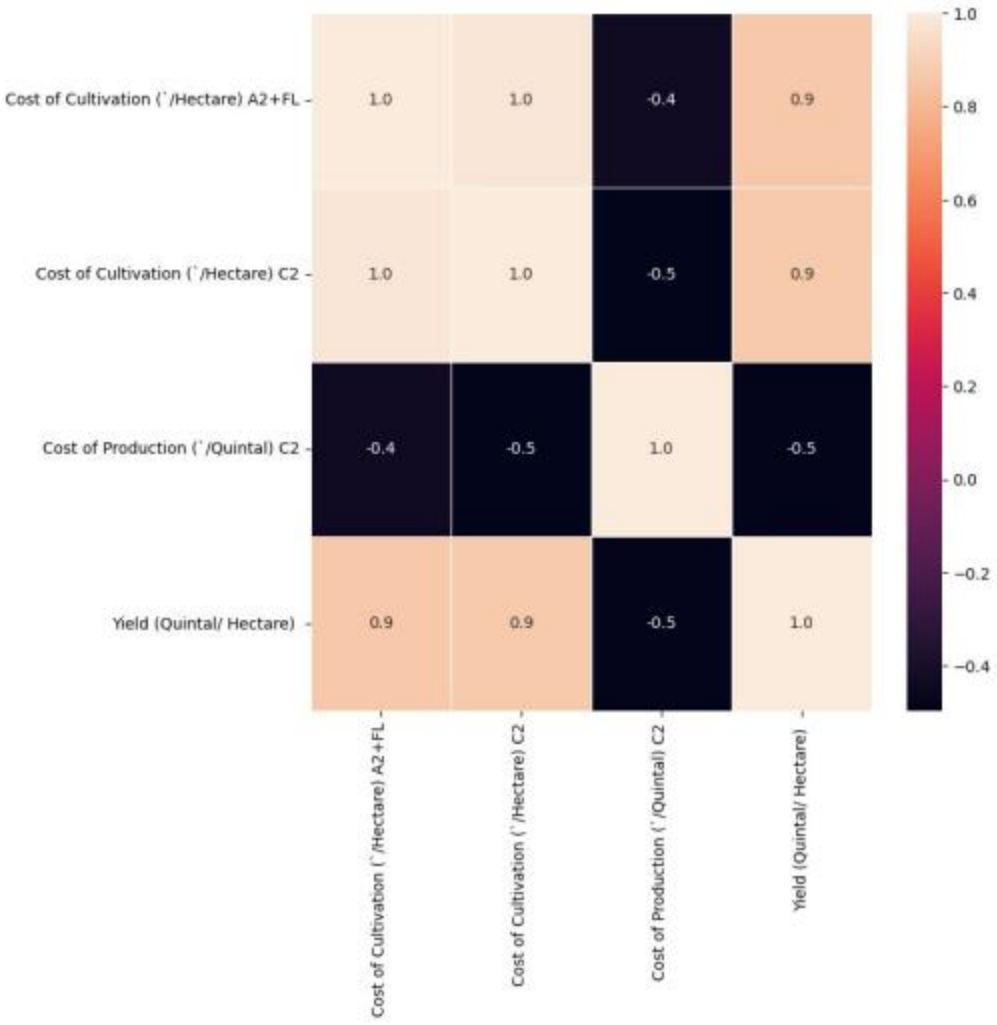

Figure 4.3: Correlation heat map (Cost of production vs yield)

(Source: Acquired from Jupyter Notebook)

The above illustration is described on the correlation heat map between the predictor cost of production and response factor maximum crop yield. It can be seen that correlation values between these two factors observe a varying dynamically. For instance, the cost of cultivation (/heater) has witnessed a negative correlation against the cost of cultivation (correlation value -0.4). On the other hand, while presenting the correlation value between the cost of cultivation (/heater) and crop yield (quintal/ hectare), it was observed that there is a positive correlation between them. It should be mentioned that having a correlation value of 0.9 suggests that these two factors observe a strong positive interdependence between them. Now, other subsequent factors like cost of production have witnessed a strong negative correlation between them

(Correlation-0.5). Changing trends where a positive correlation should be explaining an incremental and negative correlation must be indicating a decreasing trend within the data (Palanivel and Surianarayanan, 2019). Similarly, in this regard, having a positive correlation between yield and cultivation (/heater) only defines that there should be an incremental trend with the observed change in any of these factors. In other words, an incrementing trend in the cost of production or simply increasing cost of cultivation in that region has seen a major increase in crop yield.

Precisely, a crop yield in a particular region is positively dependent on the cost of cultivation. In support of this claim, maximum transaction on a specific crop has a direct positive correlation with crop production (Abbas *et al*. 2020). Similarly, observing correlating dynamics between profitability and investment particularly agricultural production has witnessed a positive trend (Nishant *et al*. 2020). Therefore, the correlation heat map produced via this study also presents a similar trend where the increasing cost of cultivation has directly affected the maximum yield rate in a specific region. Moreover, it can also be said that policymakers within a particular region should be considering this cost of cultivation factor significantly as it can have a major effect on the annual yield rate. Thus, through the correlation heat map, it gets evident that the cost of cultivation and production is an influencing factor that can either increase or decrease the overall annual crop yield rate

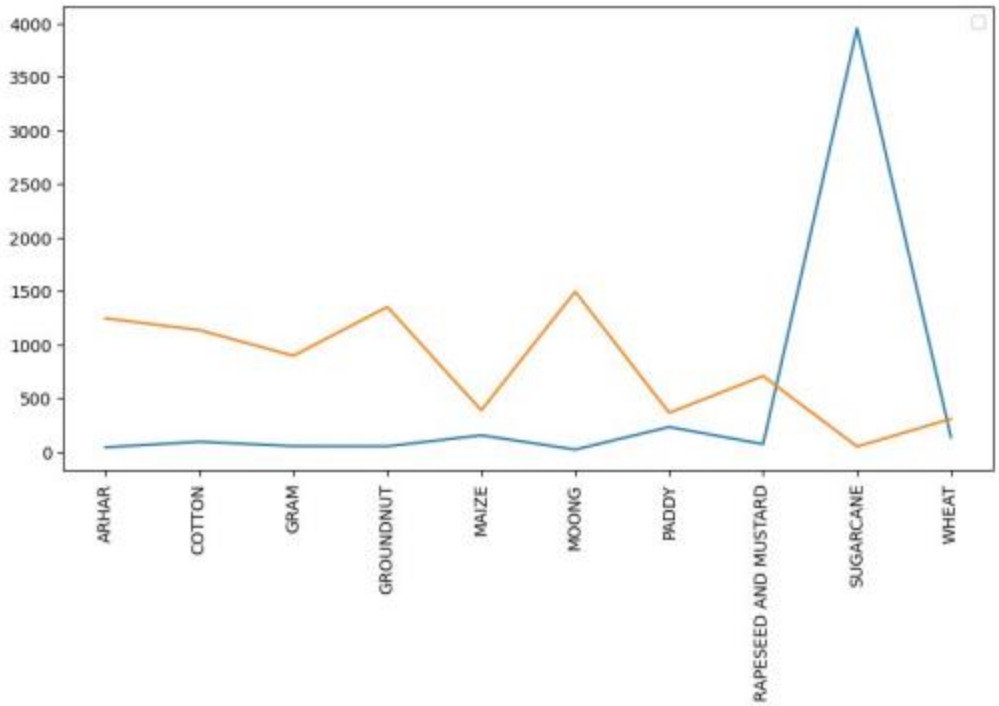

Figure 4.4: Maximum Yield Vs. cost of production

(Source: Acquired from Jupyter Notebook)

Here, in this section, a graphical presentation between Maximum Yield and cost of production has been produced. The X- axis defines the maximum yield around different types of crops. On the contrary, Y-Axis is presenting the total cost of production for each of these crops. It can be seen that the maximum production cost for sugar cane is the highest among all the mentioned crops. However, despite having a major production cost, the maximum yield is high for moong crops. Sugar cane as a crop requires major costs in terms of production. Similarly, this data set also observed a significant production cost value for crops like sugarcane and paddy. This illustration also defines that grams, ground nuts, and rapeseeds are some of the specific crops that observe an incremental effect on maximum yield concerning the cost of production.

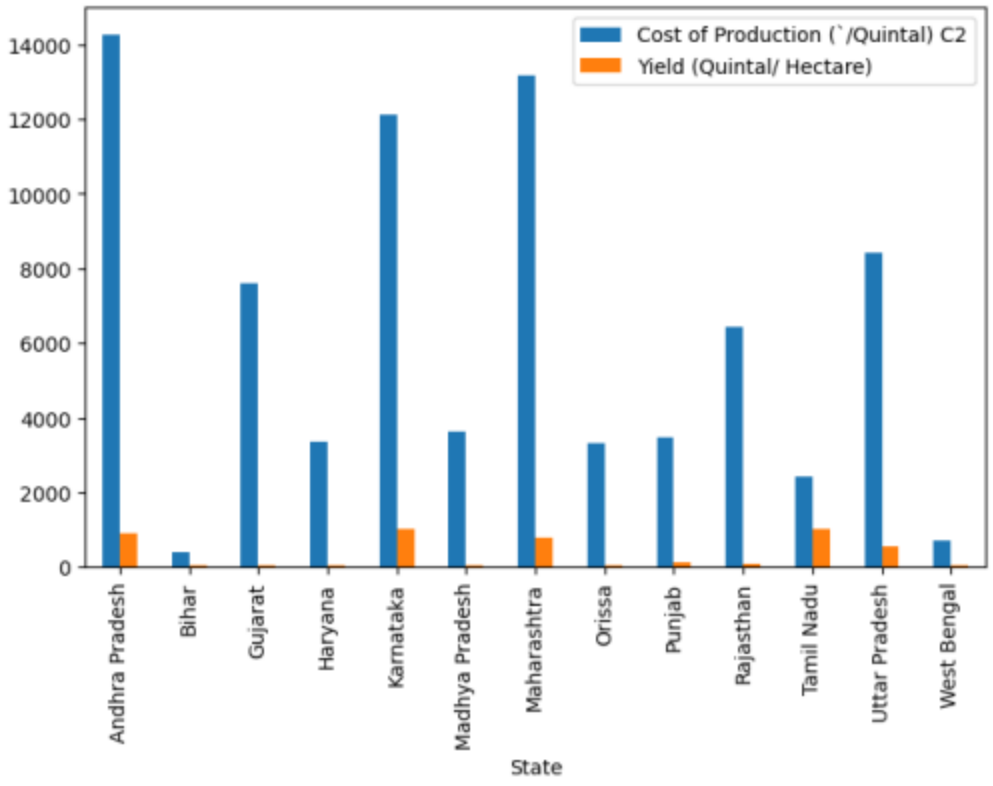

Figure 4.5: State Wise maximum cost of product and maximum yield

(Source: Acquired from Jupyter Notebook)

The above figure demonstrates the comparative relation between the cost of production and total yield. It can be seen that some states have observed a significant difference between the cost of production and yield. For instance, Crop yield in Andhra Pradesh has observed a major significant difference with the total cost of production. Similarly, states like Karnataka and Maharashtra also observed a significant difference between these two aspects. Factors other than only the cost of production (Kim et al. 2016) affect crop yield within a specific region. Similarly, in this illustration, the significant difference between the cost of production and maximum yield should be explained by the relative impact of factors like precipitation (low), temperature (high), and many more. In other words, the cost of production can affect the maximum yield in a specific region.

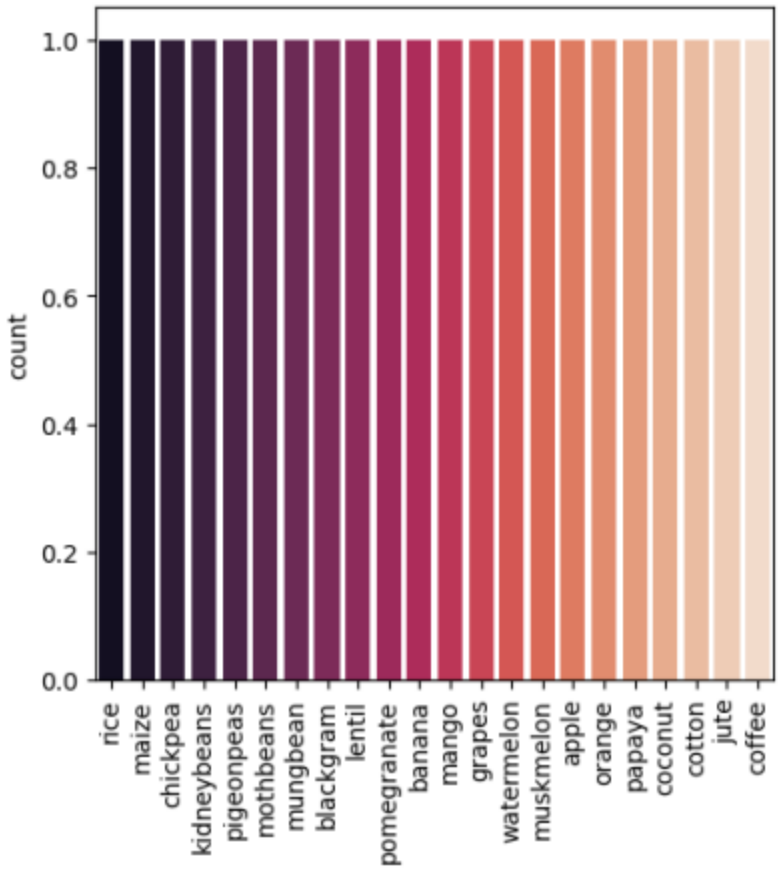

Figure 4.6: Crop production

(Source: Acquired from Jupyter Notebook)

The above figure is suggesting that crop production has seen a diverse result according to the type of crops. It can be seen that rice, maize, and chickpeas are the three major crops that have the highest production rate. On the other hand, jute, cotton, and coffee are the three major crops that can be suggested as the least produced crops in those regions. Now, jute, cotton, and coffee can be associated with low cost of cultivation as well as low cost of production (Rajak *et al.* 2017). These crops need a significantly low cost of production and still manage to deliver the highest production rate or yield rate in a region. In contrast, jute, coffee, and cotton are some of the crops that need high costs of production and cultivation. Despite that, these crops recorded significantly low yield rates in this region. Therefore, from this particular illustration, it can be said that cost of production has a significant association with the maximum yield around the crop.

KNN Accuracy

```
classifier = KNeighborsClassifier(n_neighbors=K_opt+1)
classifier.fit(xtrain, ytrain)
y_pred = classifier.predict(xtest)

accuracy = acc(ytest, y_pred)*100
print('Accuracy of the training Model : ', round(accuracy, 3), '%')
```

```
Accuracy of the training Model :  99.545 %
```

Figure 4.7: KNN accuracy

(Source: Acquired from Jupyter Notebook)

Figure 4.7 is describing on the KNN accuracy numbers. It can be seen that KNN accuracy has been observed as 99.545% which is also suggesting a good data set. It should be mentioned that KNN classifiers can determine the class of data by pointing out the majority numbers. Similarly, prediction is also done via recognition of the majority class. Therefore, in the case of observing 99.545% accuracy, it can be said that predictors like rainfall or cost of production do have a significant prediction rate against the maximum crop yield. Similarly, having 99.545% of accuracy in the KNN model also explains that factors like rainfall or cultivation cost tend to be more robust against the response factor maximum crop yield (Nikhil *et al.* 2020). The nearest number recognition through the KNN model has observed a 99.545% accuracy thereby explaining that considering factors like average precipitation or cost of production does have a significant accuracy in the prediction of the crop yield within a region.

Naive bayes accuracy

```
In [35]: X_train, X_test, y_train, y_test = train_test_split(xtrain,ytrain, test_size = 0.2)

In [37]: from sklearn.naive_bayes import GaussianNB

In [38]: nb = GaussianNB()

In [39]: nb.fit(xtrain, ytrain)
Out[39]: GaussianNB()

In [40]: nb.score(X_test, y_test)
Out[40]: 0.9943181818181818
```

Figure 4.8: Naive Bayes accuracy

(Source: Acquired from Jupyter Notebook)

The above figure is explaining on naive Bayes accuracy which is one of the faster and more reliable classification models. It needs to be mentioned that the Naive Bayes classifier assumes that the impact of a particular class within an independent feature can have a direct or indirect impact on other response classes. Similarly, hair rainfall and cost of production are the two predicting classes that can have a direct or indirect impact on the response class crop yield. The results show that there has been 99.14% accuracy as the naive Bayes model is describing. It should be mentioned that the Naive Bayes algorithm should be highly accurate in terms of predicting the impact of any predicting factor on the response factor (Mishra *et al.* 2016). Similarly, having 99.14%, accuracy through this Naive Bayes model explains that predicting crop yield through rainfall and crop cultivation cost can have significant prediction accuracy as well.

SVM accuracy

```
print('SVC accuracy: {:.2f}%'.format(SVCAcc*100))
```

```
[[15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 12  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 20  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0 11  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  1  0  0  0  0  0  0  0 13  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 11  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0 11  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  0 26  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0 15  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 16  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 25  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  8  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 15  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 13  0  0]
 [ 0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0 19  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 18]]
SVC accuracy: 98.30%
```

Figure 4.9: SVM accuracy

(Source: Acquired from Jupyter Notebook)

The above figure is illustrating the SVM accuracy numbers where maximum accuracy can be seen as 98.01%. It can be seen that SVM accuracy is 98.01% which is directly affecting on the

prediction rate. Having 98.01% accuracy through the SVM model is only suggesting that the prediction of crop yield can be done through predictors like precipitation and cost (PS, 2019). In this regard, it was observed that crop prediction in a particular region is directly dependent on the total precipitation rate and cost of cultivation.

4.4 Summary

Therefore, on the basis of high KNN accuracy, SVM accuracy, and Naive Bayes accuracy, it can be said that there is a significant relation between Crop Yield and rainfall. Similarly, from those same accuracy parameters, it can also be observed that there is a significant interdependence between crop yield and cost of production/cultivation. Precisely, both of these predictors have observed a direct positive strong relationship with the response factor Crop yield (Corelation heat map observation 0.9). In such a case, it can be said that increasing crop yield is positively dependent on the increase in rainfall. On the other hand, an increasing trend in the cost of cultivation should be directly affecting the maximum yield/hectare.

Chapter 5: Discussion

5.1 Introduction

The application of ML in crop prediction appears to be consists of various areas such as the detection of weeds and diseases and thereby, the prediction of the quality of the crop and thus the systematic prediction of the yield. This has been done by gathering data from the coherent sets in order to demonstrate insight as well as offer predictions regarding the production of the livestock of crop production. Moreover, crop-yielding prediction is the most important and essential aspect regarding making significant decisions at the regional as well as national levels. However, the prediction of the crop yielding appears to be one of the challenging segments associated with the precision of agriculture and therefore requires various data sets because it is dependent on various factors such as weather, type of soil, climatic condition and many others. Accordingly, the discussion portion consists of the discussion of the results and the discussion of the analysis derived from the results of the Jupyter notebook.

5.2 Discussion on the result

The initial table of the data sets consists of the type of crop, geographical area, the required cost of cultivation per hectare, cost of production per quintal and lastly the yield. Moreover, within the cost tail, it can be seen that “Tamil Nadu” appears to have the heist yield in the protection of sugarcane crops. The main reason can be the availability of the significant climatic condition of that particular place (Lin *et al*. 2022). Moreover, within the production data head, the production of the wheat crop appears to have sequentially grown as compared to the production data of 2006-07 to the production of 2010-11. Moreover, within the graph of the line plot, it can be seen that sugarcane seems to have gained hold of the highest position as compared to the production of other crops such as “arhar, cotton, gram, groundnut, maize, moong, paddy, wheat, rapeseed and mustard. On the other hand, as per the graphical representation of the bar plot, the cost of production per quintal of the crop is maximum in Uttar Pradesh whereas it is lowest in Bihar. Eventually, the yield is highest in Tamil Nadu on the other hand; it is lowest in “Bihar, Gujarat, Haryana, Madhya Pradesh, Orissa, and West Bengal.

On the other hand, as per the crop recommendation as rainfall within the Jupyter notebook, it can be seen that rice requires the maximum level of rainfall and the maximum value is 298.560117.

Simultaneously, EDA showcases the density graph of the rainfall as well as the temperature and according to the graph, it can be stated that both are proportional to each other. However, the yielding of crops needs to be systematically dependent on the rainfall factors and as per the KNN cotton seems to have the highest value of (N vs. Crop). Apart from that, the KNN algorithm is focused on enhancing the highest level of accuracy of the prediction level of crop yielding (Zheng *et al.* 2021). Accordingly, according to the split rainfall data into the train and test the optimal value of K is 2 and as per KNN accuracy, the accuracy of the training model is 99.545%. Lastly, in the "Naive Bayes and SVM" part the overall accuracy level of the SVC is 98.01%.

5.3 Analysis of results

Analysis of EDA

There is presence of a temperature that is greater than 0.09 and a density is greater than 0.5 for a depiction of suitable production of crops. Moreover, plots have been generated for derivation of models and rainfall is contemplated as a suitable factor that can yield crops. There is an incorporation of temperature that is lesser than 30 degrees centigrade and rainfall is greater than 120 mm for development of certain types of crops. Rainfall is depicted as a notable factor that can play an active part in cultivating crops. Therefore, these are selected as necessary measures that are involved for escalation of entire process that has been incorporated into production of several kinds of crops (Feng *et al.* 2018). Crops that are produced after implementation of these conditions of weather are "rice", "kidneybeans", "pigeonpeas", "apples", "papaya", "coconut", "jute" and "coffee" Therefore, these are depicted as essential measures that are involved after application of this procedure.

Analysis of model development

There has been incorporation of a "KNN classifier" for making depictions for predictions about crops. There has been implementation of a library that has been kept for downloading all information that is present in source codes. Moreover, a "Confusion matrix" is incorporated so that there is an appropriate definition of performance. Therefore, a complete summarization of performance is depicted after depicting this particular technique. This matrix is utilized for making a complete analysis of performance after involving this approach (Ruuska *et al*. 2018). Therefore, a complete determination regarding accuracy is involved after depiction of this

procedure. There are probabilities of getting appropriate separations after implementing this framework (Patil *et al.* 2020). Therefore, this has been selected as an essential approach that has been selected after appropriate form of application in this field.

After this procedure, there has been a classification of all sets of data into training and testing sets. This classification has been done in 70:30 ratios. There has been incorporation of 70 percent of data for training and 30 percent of data is left for purpose of testing. At first, training process has been conducted on 70 percent of data sets. After conduction of this procedure, execution has been done based on results of testing. Therefore, there is a minimization of probability of misinterpretation of information after incorporating this technique. There are possibilities for sustainable applications after execution of this procedure (Elavarasan and Vincent, 2020). There these are depicted as necessary procedures that are executed after getting a detailed application in this specific context. Moreover, there are probabilities of getting better forms of dimensions after making appropriate inclusion in this field (Domingues *et al.* 2022). Moreover, humanized methods are involved in making comparisons after depictions regarding these approaches.

Model Comparison

After implementation of "KNN classifier", an accuracy of approximately 99.545 percent has been acquired. Therefore, there has been a generation of the most appropriate accuracy after implementing this procedure. Prediction about classifiers has been implemented so that the most relevant result is available. There are provisions for making dissociation of all types of classes after involvement of this specified model (Palanivel and Surianarayanan, 2019). There is an availability of accuracy of approximately 98 percent accuracy after application of "Naive Bayes classifier" algorithm. Hence, it has been selected as one of the most notable practices that are brought into practice after depiction of this procedure. Moreover, there has been an application of SVM model and there is a generation of approximately 98.01 percent accuracy after depiction of this procedure. This comparison has been carried out in this table.

"KNN classifier"	**SVM model**	**"Naive Bayes classifier"**
There has been a generation of approximately 99.545	There is a determination of about 98 percent accuracy	It is seen that 98.01 percent accuracy has been acquired

percent accuracy.	after implementation of this approach.	after getting ideas in this field.
Accuracy of this model has been identified as the highest.	Accuracy of this specified model has been reported as a lower amount as compared to "KNN classifier".	Accuracy is lesser than "KNN classifier".

Table 5.1: Comparison between accuracies of three models

(Source: Self-developed)

Explanation

It is seen that seen the "KNN classifier" is the best algorithm that has been adopted for making appropriate predictions regarding cultivation of crops. Moreover, there is a requirement of a lesser amount of time that has been necessary for giving appropriate training to all sets of data (Sethy *et al.* 2020). Therefore, it has been contemplated as one of the most identical approaches that have been selected after adopting this approach.

Through the use of ML models, it has been observed that crop yield prediction has a major dependence on the factors like rainfall and cost of production. It can be seen that that these factors can have a direct positive impact on yield rate. The prediction value observed within these models has been over 98% which explains that in terms of accurately predicting crop prediction, these factors can be taken as influencers. In such a case, it can be also said that increased crop yield prediction rate has major dependence over the crop production cost and the average precipitation rate within a region.

Thus, this study has been significantly empirical towards data where a correlation between factors like crop production cost and total yield has been presented. The empirical nature of the result should be taken as major advantage that has been provided through the implementation of various ML algorithms. Similarly, data visualization through the help of this ML technology can have a direct effect on the reliability and accuracy of the result. Henceforth, it can be said crop yield prediction through secondary data sets has successfully fulfilled the predetermined research

scope. In such a case, this research has fulfilled all its objectives as well as answered all the questions. Crop yield prediction must be done considering factors like rainfall and cost of cultivation. Hence, it has been adopted as a best method that is incurred in this procedure.

Chapter 6: Conclusion and recommendations

6.1 Conclusion

It is concluded that appropriate analysis, classification and prediction regarding crop classification are has been done successfully after incorporation of ML strategies. This research involves strategies for making decisions regarding crops after implementing concepts from "***Naive Bayes classifier***" and "***KNN classifier algorithm***". There are probabilities of getting appropriate accuracies of data and making relevant strategies of acquisition of decisions for making calculations regarding probabilities. Improvement is seen in prediction after incorporation of this method. Frameworks are adopted so that there are adaptations of several procedures for selection of necessary approaches for yielding better qualities of crops. Hence, these are selected as essential approaches that are implemented in this strategy. There is a derivation of 99.545 percent accuracy and it has been derived from "KNN classifier". Therefore, it has been selected as one of the most significant methods for execution of this procedure.

6.2 Linking with objectives

This research has fulfilled the first objective, as there is a complete analysis of predicting prices of crops after implementation of this research. Ideas about probabilities of classes are generated for making predictions regarding target types of variables. There are ideas about types of variables acquired after incorporating this classifier in this type of analysis. There are probabilities of analyzing performances after implementing a definite framework (Saritas and Yasar (2019). Hence, applications are available to make this kind of prediction after incorporation of this procedure.

There has been fulfillment of second objective as contemplation of factors is involved for making an estimation of prices. Moreover, there are chances of making pertinent forms of assumptions regarding costs. Support is also presented for making predictions about prices. There is also incorporation of minimum support that is needed for getting this kind of

information. Hence, there is an enhancement in production of several types of crops after involving this classifier. There is a requirement for frameworks after derivation of these ideas. There is existence of multiple kinds of frameworks and these frameworks are contemplated as effective frameworks in making assumptions about crops.

In this research, third objective is met as revenues are generated regarding revenues. Therefore, there is an incorporation of strategies after accomplishment of this method. Moreover, prediction about values is possible so there has been a clear idea after involvement of this strategy. Alternative methods are generated after implementation of necessary opportunities that are implicated in this procedure. There are probabilities of making predictions so that desired amount of accuracy is generated after getting ideas in this kind of prediction (Ning *et al.* 2019). Hence, decisions of precision are acquired after involvement of necessary ideas in this strategy of assuming price of crops.

Forth objective is met after prediction of profits and determination of costs. Implementation of a simplified method is generated after involvement of ideas that are implicated in this perspective. Ideas about accuracy of crops are involved and strategies are assumed for requirement of several kinds of resources that are essential in yielding crops. Therefore, predictions are made about an escalation of profits after executing this classifier algorithm. Moreover, costs for cultivation are incurred after procuring complex ideas regarding prediction of yield of crops. Suitable measures are adopted after knowing the best factors that can help in cultivation of certain types of crops.

6.3 Recommendations

Recommendation 1: Increase of profits after making accurate assumptions for costs

There are probabilities of making accurate assumptions so that costs are predicted after implementing this method. Therefore, these are depicted as approaches that are incorporated so that there is a prevalence of mechanisms and these are selected as opportunities that are involved in this method. Determination of profits is derived as a significant impact after implementing these measures (Safri *et al.* 2018). Therefore, these are depicted as notable measures that are taken into contemplation for involvement of necessary procedures in this specified context. There are chances of making improvements in identification of diseases that are caused to plants

(Ahmed *et al.* 2022). Hence, better decisions are possible in this regard and overall improvement is noticed in making assumptions about crops.

Recommendation 2: Standard approaches are needed for scaling so that there are chances of getting the best accuracy after application of this method

There are necessities of introducing standard strategies of scaling for better accuracy and these are contemplated as essential measures that are implemented after a description of this approach. Therefore, these are selected as one of the best approaches that are taken into contemplation so that there are probabilities of getting the most suitable outcomes. Therefore, decisions are adopted so that utilization of this approach is adopted to get the best solutions. There is a need for involvement of procedures that are adopted for generation of best approaches for adopting pertinent procedures of scaling (Patil *et al*. 2020). Hence, these are depicted as essential strategies that are implemented for escalation of advantages in cultivation of crops.

Recommendation 3: Implementation of an effective method for computation of probabilities in using "Naive Bayes classifier"

There is a need for adaptation of notable procedures so that there is a depletion of redundant characteristics that are present in this model. Moreover, approaches are essential for segmentation of all kinds of data that are implemented for creation of strategies after incorporation of this strategy (Safri *et al.* 2018). Therefore, this strategy has been incorporated for making better forms of decisions and for adoption of best practices that are involved in this specified procedure. Moreover, incorporation of probabilities that are involved with distribution is implemented so that ideas about appropriate outcomes are adopted.

Recommendation 3: Utilization of a generative model for making suitable ideas that are incorporated into prediction of crops

There are possibilities of getting accurate outcomes after using this specified model. Therefore, ideas are generated for making detailed assumptions of all types of crops and hence, these are determined as necessary measures that are involved in this approach. Moreover, continuous types of variables are being handled after incorporation of this specified framework. Hence, these are selected as appropriate procedures that are adopted so that pertinent decisions regarding

cultivation are possible (Ahmed *et al.* 2022). Additionally, there are probabilities of handling all kinds of data after involvement of this procedure.

Recommendation 4: There is a need of making accurate predictions of weather about yield of crops

There are probabilities of making suitable predictions so that proper procedures are followed for making better strategies for crops. Moreover, there is a need to select pertinent measures that can help in adaptation of necessary measures for irrigation. Hence, these are contemplated as the best decisions that are adopted in this approach. There are chances of making an appropriate kind of analysis after depiction of necessary frameworks that are implemented in this approach (Taher *et al.* 2021). Additionally, decisions are possible in selection of approaches for an implication of appropriate factors that can take an active part in yielding crops.

6.4 Future Work

In the future, a "Naive classifier" algorithm can be used to solve classification issues. Moreover, there are probabilities of using datasets that are high dimensional for incorporation of ideas in this strategy. Additionally, there are opportunities of using both continuous sets of data and discrete sets of data after using this specified classifier. Predictions about sensitive characteristics are possible and depicted as necessary measures in this scenario. "KNN algorithm" can be used for getting highest amount of accuracy. Appropriate management of methods of cultivation is brought into practice after incorporation of procedures that are relevant to this algorithm (Abbas *et al.* (2019). Measurement of distance is procured after making a detailed application, and hence, in future prediction about revenues can be done more accurately. Therefore, standards for making predictions are enhanced after application of this method and it can help in pertinent procurement of decisions.

6.5 Implications of this research

Academic implications

This research can help in depiction of necessary measures so that researchers are acquainted with factors that impact crop yield. Hence factors such as temperature, humidity, ph. and rainfall have direct impact on the crop yield.

Industrial modes of implications

Temperature, humidity, pH, and rainfall are just a few of the variables that have an impact on crop output that are significantly influenced by the industrial sector. Increased atmospheric greenhouse gases may result from industrial activities including the production of energy, transportation, and chemicals, which may influence weather patterns and the climate (Pandith *et al*. 2020). Crop yields may be directly impacted by these changes due to different growth circumstances, such as temperature swings and water availability.

6.6 Limitations present in this research

Limitations can include an absence of sufficient time. With involvement of this disadvantage, there are chances of making limitations in making analysis and hence, there are probabilities of getting relevant outcomes. Moreover, there is an absence of sufficient budget and it has been described as a disadvantage. This is because information is not accumulated from paid types of sources because of existence of this limitation. Hence, these are contemplated as notable limitations after involvement of this procedure.

References

Abbad Ur Rehman, H., Lin, C.Y. and Mushtaq, Z., 2021. Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. *Journal of the Chinese Institute of Engineers*, *44*(1), pp.77-87.

Abbas, F., Afzaal, H., Farooque, A.A. and Tang, S., 2020. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, *10*(7), p.1046.

Abbas, M., Memon, K.A., Jamali, A.A., Memon, S. and Ahmed, A., 2019. Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), p.62.

Ahmed, A.S., Obeas, Z.K., Alhade, B.A. and Jaleel, R.A., 2022. Improving prediction of plant disease using k-efficient clustering and classification algorithms. IAES International Journal of Artificial Intelligence (IJ-AI), 11(3), pp.939-948.

Archana, K. and Saranya, K.G., 2020. Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. SSRG Int. J. Comput. Sci. Eng, 7, pp.1-4.

analyticsvidhya (2022) : Building Naive Bayes Classifier from Scratch to Perform Sentiment Analysis [Online] Available at: https://www.analyticsvidhya.com [Accessed on 28th April 2023]

Chandana, C. and Parthasarathy, G., 2022. Efficient Machine Learning Regression Algorithm using Naïve Bayes Classifier for Crop Yield Prediction and Optimal Utilization of Fertilizer. International Journal of Performability Engineering, 18(1).

Domingues, T., Brandão, T. and Ferreira, J.C., 2022. Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. Agriculture, 12(9), p.1350.

Elavarasan, D. and Vincent, P.D., 2020. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE access, 8, pp.86886-86901.

Feng, P., Wang, B., Liu, D.L., Xing, H., Ji, F., Macadam, I., Ruan, H. and Yu, Q., 2018. Impacts of rainfall extremes on wheat yield in semi-arid cropping systems in eastern Australia. Climatic change, 147(3), pp.555-569.

Ghadge, R., Kulkarni, J., More, P., Nene, S. and Priya, R.L., 2018. Prediction of crop yield using machine learning. *Int. Res. J. Eng. Technol.(IRJET)*, *5*.

Goldstein, S.B., 2022. A systematic review of short-term study abroad research methodology and intercultural competence outcomes. *International Journal of Intercultural Relations*, *87*, pp.26-36.

Goldstein, S.B., 2022. A systematic review of short-term study abroad research methodology and intercultural competence outcomes. *International Journal of Intercultural Relations*, *87*, pp.26-36.

Kaur, J., Bhambri, P. and Sharma, K., 2019. Wheat Production Analysis based on Naive Bayes Classifier. International Journal of Analytical and Experimental Model Analysis, 11(9), pp.705-709.

Kaur, P., Chahal, J.K. and Sharma, T., 2021. A Data Mining Approach for Crop Yield Prediction in Agriculture Sector. Advances in Mathematics: Scientific Journal, 10(3), pp.1425-1430.

Kim, Y.H., Yoo, S.J., Gu, Y.H., Lim, J.H., Han, D. and Baik, S.W., 2016. Crop pests prediction method using regression and machine learning technology: Survey. *IERI Procedia*, *6*, pp.52-56.

Lee, B.A., Ogunfemi, N., Neville, H.A. and Tettegah, S., 2021. Resistance and restoration: Healing research methodologies for the global majority. *Cultural Diversity and Ethnic Minority Psychology*.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D., 2018. Machine learning in agriculture: A review. Sensors, 18(8), p.2674.

Lin, G., Lin, A. and Gu, D., 2022. Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient. *Information Sciences*, *608*, pp.517-531.

Lin, J., Cao, Y., Zhu, K., Yan, F., Shi, C., Bai, H., Ge, G., Yang, J., Yang, W., Li, G. and Zeng, H., 2022. Ultrahigh energy harvesting properties in temperature-insensitive eco-friendly high-performance KNN-based textured ceramics. Journal of Materials Chemistry A, 10(14), pp.7978-7988.

Medar, R.A., Rajpurohit, V.S. and Ambekar, A.M., 2019. Sugarcane crop yield forecasting model using supervised machine learning. International Journal of Intelligent Systems and Applications, 11(8), p.11.

Mishra, S., Mishra, D. and Santra, G.H., 2016. Applications of machine learning techniques in agricultural crop production: a review paper. *Indian J. Sci. Technol*, *9*(38), pp.1-14.

Nikhil, R., Anisha, B.S. and Kumar, R., 2020, July. Real-time monitoring of agricultural land with crop prediction and animal intrusion prevention using internet of things and machine learning at edge. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1-6). IEEE.

Ning, B., Junwei, W. and Feng, H., 2019. Spam message classification based on the Naïve Bayes classification algorithm. IAENG International Journal of Computer Science, 46(1), pp.46-53.

Nishant, P.S., Venkat, P.S., Avinash, B.L. and Jabber, B., 2020, June. Crop yield prediction based on indian agriculture using machine learning. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1-4). IEEE.

Palanivel, K. and Surianarayanan, C., 2019. An approach for prediction of crop yield using machine learning and big data techniques. International Journal of Computer Engineering and Technology, 10(3), pp.110-118.

Palanivel, K. and Surianarayanan, C., 2019. An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, *10*(3), pp.110-118.

Pallathadka, H., Mustafa, M., Sanchez, D.T., Sajja, G.S., Gour, S. and Naved, M., 2021. Impact of machine learning on management, healthcare and agriculture. Materials Today: Proceedings.

Pandith, V., Kour, H., Singh, S., Manhas, J. and Sharma, V., 2020. Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. Journal of Scientific Research, 64(2), pp.394-398.

Pandith, V., Kour, H., Singh, S., Manhas, J. and Sharma, V., 2020. Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. Journal of Scientific Research, 64(2), pp.394-398.

Patil, P., Panpatil, V. and Kokate, S., 2020. Crop prediction system using machine learning algorithms. Int. Res. J. Eng. Technol.(IRJET), 7(02).

Pereira, L., Santos, R., Sempiterno, M., Costa, R.L.D., Dias, Á. and António, N., 2021. Pereira Problem solving: Business research methodology to explore Open Innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(1), p.84.

Potratz, G.L., Canchumuni, S.W.A., Castro, J.D.B., Potratz, J. and Pacheco, M.A.C., 2021. Automatic lithofacies classification with t-SNE and K-nearest neighbors algorithm. *Anuário Do Instituto De Geociências*, *44*.

PS, M.G., 2019. Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. *Applied Artificial Intelligence*, *33*(7), pp.621-642.

Rajak, R.K., Pawar, A., Pendke, M., Shinde, P., Rathod, S. and Devare, A., 2017. Crop recommendation system to maximize crop yield using machine learning technique. *International Research Journal of Engineering and Technology*, *4*(12), pp.950-953.

Rashid, M., Bari, B.S., Yusup, Y., Kamaruddin, M.A. and Khan, N., 2021. A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE Access*, *9*, pp.63406-63439.

Roman, D., Saxena, S., Robu, V., Pecht, M. and Flynn, D., 2021. Machine learning pipeline for battery state-of-health estimation. Nature Machine Intelligence, 3(5), pp.447-456.

Rumsey, M., Stowers, P., Sam, H., Neill, A., Rodrigues, N., Brooks, F. and Daly, J., 2022. Development of PARcific approach: participatory action research methodology for collectivist health research. *Qualitative Health Research*, p.10497323221092350.

Rumsey, M., Stowers, P., Sam, H., Neill, A., Rodrigues, N., Brooks, F. and Daly, J., 2022. Development of PARcific approach: participatory action research methodology for collectivist health research. *Qualitative Health Research*, p.10497323221092350.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P. and Mononen, J., 2018. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behavioural processes, 148, pp.56-62.

Safri, Y.F., Arifudin, R. and Muslim, M.A., 2018. K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. Sci. J. Informatics, 5(1), p.18.

Saritas, M.M. and Yasar, A., 2019. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), pp.88-91.

Sethy, P.K., Barpanda, N.K., Rath, A.K. and Behera, S.K., 2020. Nitrogen deficiency prediction of rice crop based on convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 11(11), pp.5703-5711.

Setiadi, T., Noviyanto, F., Hardianto, H., Tarmuji, A., Fadlil, A. and Wibowo, M., 2020. Implementation of naïve bayes method in food crops planting recommendation. Int. J. Sci. Technol. Res, 9(02), pp.4750-4755.

Sosa-Díaz, M.J. and Valverde-Berrocoso, J., 2022. Grounded theory as a research methodology in educational technology. *International Journal of Qualitative Methods*, *21*, p.16094069221133228.

Statista. (2022). Crop Production - United Kingdom. Available at: https://www.statista.com/outlook/io/agriculture/crop-production/united-kingdom [Accessed on: 19/12/2022]

Statista. (2022). Total software spending in the agriculture industry in the United Kingdom from 2018 to 2024. Available at: https://www.statista.com/statistics/1275067/uk-software-spending-industry-agriculture/ [Accessed on: 19/12/2022]

Sun, R., 2019. Optimization for deep learning: theory and algorithms. arXiv preprint arXiv:1912.08957.

Sun, Y., Song, Q. and Liang, F., 2021. Consistent sparse deep learning: Theory and computation. Journal of the American Statistical Association, pp.1-15.

Taher, K.I., Abdulazeez, A.M. and Zebari, D.A., 2021. Data Mining Classification Algorithms for Analyzing Soil Data. Asian Journal of Research in Computer Science, pp.17-28.

Tembusai, Z.R., Mawengkang, H. and Zarlis, M., 2021. K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification. *International Journal of Advances in Data and Information Systems*, *2*(1), pp.1-8.

Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F., 2020. The computational limits of deep learning. arXiv preprint arXiv:2007.05558.

towardsdatascience(2018) : KNN (K-Nearest Neighbors) #1 [Online] Available at: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d [Accessed on 28th April 2023]

towardsdatascience(2018) : SVM: Feature Selection and Kernels [Online] Available at: https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c [Accessed on 28th April 2023]

Ullah, A. and Rafiq, M., 2022. Education and learning about research methodology: Views of LIS authors in Pakistan. *Information Development*, *38*(4), pp.688-703.

Wazery, Y.M., Saber, E., Houssein, E.H., Ali, A.A. and Amer, E., 2021. An efficient slime mould algorithm combined with a k-nearest neighbor for medical classification tasks. *IEEE Access*, *9*, pp.113666-113682.

Zhang, Y., 2022. Research methodology. In *Assessing Literacy in a Digital World* (pp. 51-71). Springer, Singapore.

Zhang, Y., 2022. Research methodology. In *Assessing Literacy in a Digital World* (pp. 51-71). Springer, Singapore.

Zheng, T., Yu, Y., Lei, H., Li, F., Zhang, S., Zhu, J. and Wu, J., 2022. Compositionally Graded KNN-Based Multilayer Composite with Excellent Piezoelectric Temperature Stability. *Advanced Materials*, *34*(8), p.2109175.

Zhou, Y., Li, H. and Sun, H., 2022. Metalloproteomics for Biomedical Research: Methodology and Applications. *Annual Review of Biochemistry*, *91*.

Zhou, Y., Li, H. and Sun, H., 2022. Metalloproteomics for Biomedical Research: Methodology and Applications. *Annual Review of Biochemistry*, *91*.