Crop Yield Prediction Using Machine Learning

[1]Mrs.M.Sailaja, [2]S.Nandini,[3] N.Nandini , [4]P.Dhanusri, [5]M.Harini
[1]Guide [2,3,4,5]U.G Scholar
[1,2,3,4,5]Computer Science and Engineering
[1,2,3,4,5]Ravindra College of Engineering for Women

ABSTRACT

Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms have been applied to support crop yield prediction research. In this study, we performed a Systematic Literature Review (SLR) to extract and synthesize the algorithms and features that have been used in crop yield prediction studies. Based on our search criteria, we retrieved 567 relevant studies from six electronic databases, of which we have selected 50 studies for further analysis using inclusion and exclusion criteria. We investigated these selected studies carefully, analyzed the methods and features used, and provided suggestions for further research. According to our analysis, the most used features are temperature, rainfall, and soil type, and the most applied algorithm is Artificial Neural Networks in these models. After this observation based on the analysis of machine learning-based 50 papers, we performed an additional search in electronic databases to identify deep learning-based studies, reached 30 deep learning-based papers, and extracted the applied deep learning algorithms. According to this additional analysis, Convolutional Neural Networks (CNN) is the most widely used deep learning algorithm in these studies, and the other widely used deep learning algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN).

I. INTRODUCTION

Agriculture is the basic source of food supply in all the nations of the world whether it's under- developed, developing or developed countries. According to the Bangladesh Bureau of Statistics (BBS), 2017, about 17 % of the country's Gross Domestic Product(GDP) is a contribution of the agricultural sector. The decreasing crop production and shortage of food across the world is one of the crucial criteria of agriculture now-a-days is selecting the right crop for the right piece of land at the right time. In our research,we suggested a system for recommending

the most appropriate crops for a specific lands based on the analysis of the data on certain affective parameters using machine learning. We used Random Forest Classifier, Gaussian Nave Bayes, Logistic regression, Support Vector Machine, K-Nearest Neighbor and Artificial Neural Network for crop selection in our research.. We have trained these algorithms with the training data and later these were tested with test dataset. After That we compared the performances of all the tested methods to arrive at the best

outcome.

For a country, one of the most crucial aspects of its development circles around its capacity to produce food. At this time, the rate of urbanisation is by far our civilization's most superior goal. In doing this, we are ignorantly diminishing our capacity for agriculture; especially in terms of land and fertility. We will have to focus on making the most of what we have because the amount of land available will not increase in this era of urbanisation and globalisation. Due to this issue, we have to device newer ways to farm and extract the absolute most from these limited land resources. If effectively implemented in this ageof technology and data science, the agricultural sector might be dramatically impacted. However, machine learning techniques can be applied in this field for far greater precision and stability of selection. In this research, we have attempted to come up with a few techniques that will lead us to choose suitable crops based on specific state, specific district, season, and some other environmental aspects. By examining all of these challenges and problems, such as weather, temperature, and a variety of other elements, there is no proper solution and technologies to overcome the situation faced by us.

In India, there are numerous options for increasing agricultural economic growth. There are numerous methods for increasing and improving agricultural output and quality. Data Mining is also useful for the crop yield prediction.

Generally data mining is the process of analyzing data from different perspective and summarize it into useful information. Technically, data mining is the process of finding correlations or patterns among in the fields of large relational databases. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about the historical patterns and future trends.

Crop yield forecasting is a significant agricultural issue. Each and Every farmer is always trying to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depending on weather conditions, pests and planning of harvest operation. Accurate information about the history of crop yield is an important thing for making decisions related to agricultural risk management. This research focuses on evolution of a prediction model which may be used to predict crop yield production. The proposed method uses data mining technique to predict the crop yield production based on the association rules.

Data Mining is emerging research field in crop yield analysis. In agriculture, yield prediction is a critical issue. Any farmer wants to know how much he may expect in terms of yield. Yield prediction used to be done by taking into account a farmer's previous experience with a certain field and crop. The prediction of yield is a critical challenge that has yet to be resolved

using existing data. For this, data mining techniques are the best option. In agriculture, many Data Mining techniques are applied and analysed to forecast crop production for the coming year. This study suggests and tests a technique for predicting agricultural productivity based on historical data. This is accomplished through the use of association rule mining on agricultural data. This paper presents a brief analysis of crop yield prediction using data mining technique based on association rules for the selected region. The experimental results show that the proposed work efficiently predict the crop yield production.

One of the goals of agricultural production is to achieve maximum crop yield at the lowest possible cost. Early detection and control of problems with crop yield indicators can aid in increasing output and profit.

By influencing regional weather patterns, large-scale meteorological phenomena can have a significant impact on agricultural production. Crop managers could use predictions to reduce losses in the event of unfavorable weather. Furthermore, when optimal growing conditions occur, these predictions could be employed to maximize crop prediction. Crop yield prediction, particularly for strategic plants like wheat, corn, and rice, has always been a fascinating research topic for agrometeorologists, as it is crucial for national and international economic planning.

Dry farming crop production, apart from relationship to the genetic of cultivator, adipic terms, effect of pests, pathology, and weeds, as well as management and quality control during the growing season and so on is severely depend to climatic events. As a result, using meteorological data, it is systems that can predict with greater precision. Nowadays, there are a lot of yield prediction models, that more of them have been generally classified in two group: a) Statistical Models, b) Crop Simulation Models. Artificial Intelligence (AI) techniques such as Artificial Neural Networks (ANNs), Fuzzy Systems, and Genetic Algorithms have recently demonstrated to be more effective in handling the problem. They can be used to construct models from complex natural systems with many inputs easier and more accurate. In this research it has been tried to develop a various crop yield prediction model using it can be ANNs can be used to estimate crop production in the long or short term, and they can also be obtained with enough and useful data.

PROBLEM STATEMENT

Agriculture is the main occupation of the majority of population. The district's farmers are heavily reliant on agriculture for their livelihood. Agriculture's development is influenced by a variety of factors, including soil type, relief, vegetation, climatic conditions, attitudes of different social groups of farmers toward agriculture, irrigation, HYV seeds, fertilizer, pesticides and insecticides, and the use of mechanical tools and implements, as well as proper scientific crop rotation. The impact of

these aspects of agriculture varies in different areas of the district. There are distinct variations in the magnitude of these concepts both over space and time. To have real understanding of the nature of agricultural development, scientific investigation and evaluation of different aspects of development become highly necessary.

Keeping these points in view, the department of agriculture research has been selected as the study area because there has been significant development in agriculture in the district in the post-independence. The level of agricultural development is not the same in all district which is inhabited by various social groups of people. This is because they live in different geographical areas and their attitudes to agriculture are different. The five social groups among various groups, viz. the indigenous Hindus, the indigenous.

MOTIVATION

On this research, we primarily focused on creating a platform that can be implemented on a national scale. We experienced our primary motivation from some worldwide statistics, Bangladeshi national statistics, and personal experience as well. According to World Food Program, one-in-nine people around the world, even on this day go to sleep on an empty stomach every night. That is about 821 million people. In our country, which is a developing country, the scenario is far worse. According to BRAC the agricultural land in Bangladesh is shrinking by 1% annually, while the population is growing by 1.2% due to rising sea- level, extreme weather patterns, frequent flooding, and loss of arable land. All these lead to a growing and sustained threat to food security.

Ways to avoid such calamity should be looked up with great emergency. In this era of technology, a sector as important as the agricultural sector should not go without utilizing the perks of it.

OBJECTIVES

Our project is to predict the maximum yield of the crops produced at minimum cost. Early detection and management of problems with crop yield indicators can aid in increasing yield and profit. By influencing regional weather patterns, large-scale meteorological phenomena can have a significant impact on agricultural production.

- To study the socio-spatial and temporal variations of agricultural land use pattern.
- To investigate the pattern of agricultural productivity, intensity of cropping, crop diversification and rotation of crops.
- To assess the contribution of various social groups to the agricultural changes in the region and examine the controlling factors behind such changes.

PROPOSED SYSTEM

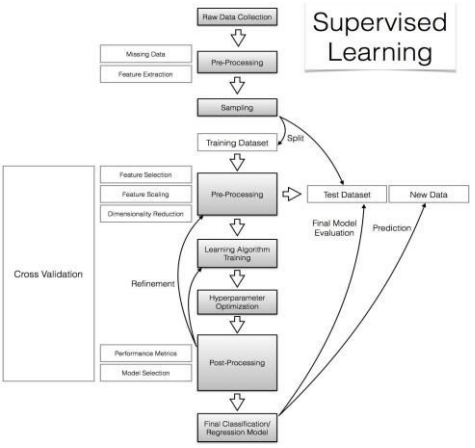
Supervised Learning

The search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances is known as supervised machine learning. Inother words, supervised learning aims to create a compact model of the distribution of class labels based on predictor features. When the values of the predictor features are known, the resulting classifier is used to assign class labels to the testing instances but the value of the class label is unknown. It simply is the task of learning a function that maps an input to an output based on input-output pairs of examples. A supervised learning algorithm analyzes the training data produces an inferred function, which can be used for mapping new examples. The algorithm will be able to correctly determine the class labels for unobserved instances in an ideal scenario. This necessitates the learning algorithm's ability to reasonably generalize from the training data to unknown situations.

Supervised learning models have some benefits over unsupervised learning models, but they also have drawbacks. Because humans have provided the foundation for decisions, the systems are more likely to make judgments that humans can relate to. However, in the case of a retrieval-based method, supervised learning systems have trouble dealing with new information

Data Classification

Classification is a data mining function that assigns items in a collection to target categories or



classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify a particular crop as to the possibility of its production given a set of features, as per our research.

- Dataset collection.
- Data visualization.
- Data pre-processing in the form of data cleaning and feature extraction.
- Data splitting into train and test sets.
- Fitting the algorithm.
- Parameter tuning (only for Artificial Neural Network).
- Testing the accuracy of the model.
- Data post-processing in the form of performance metrics.

Figure 3.1 Block diagram of our model.

Data Collection

For a research of this sort, it is crucial to have an available dataset to work upon. It is very difficult to find legible and reliable datasets of this sort. It took us a lot of time and effort to find one suitable for us. The dataset that we finally found contained all the accurate features that we really wanted. The features were perfect for a research of this sort. There was a total of 12 columns and around 250,000 rows. The columns were “State Name”, “District Name”, “Crop Year”, “Season”, “Area”, “Rainfall”, “Humidity”, “Temperature”, “Previous Year’s Rainfall”, “Previous Year’s Humidity”, “Previous Year’s Temperature” and “Crop”. 4 of the columns contained data which were in string notation. The rest of the columns contained data which were numerical. The “State Name” column had the names of a number of important states in the country. The “District Name” column had names of districts in those states.

Data Pre-Processing

One of the primary tasks we completed was to convert all of our string data into numerical data. In order to do this, we converted all the string features into dummy variables. This greatly increased our column number. We then cleaned our data in a very singular pattern. We had a small portion of null values in the production column. Due to the miniature amount, we dropped the fields off of the dataset. This did not affect much, as the amount was minimal. The dataset was also subjected to feature extraction.

Our data was categorical in some cases and continuous numeric in others. Algorithms are always hampered by this type of mixed data. As a result, we used standard feature scaling to bring all of the data into a common scale. Because most algorithms have a lot of internal calculations, feature scaling is extremely important. In addition, feature selection was carried out in order to avoid overfitting issues.

Data Splitting

Data splitting is the process of splitting the dataset into training and testing data. This process is very useful for any machine learning process as the main idea of machine learning depends on training and testing data and finding the accuracy of the machine given result. In our research, we divided our dataset because we trained our algorithms on the test dataset where the particular crop had its data in there. Here the algorithms were trained using that data. We deduced that data having 1 to be yes, and 0 to be no. The algorithms were trained and we will apply the trained algorithm to our test set and measured the accuracy of the machine

Algorithm Fitting

The most crucial part of the model was to fit the algorithm with the data. All the algorithms were easily fitted as the programming of this part was comparatively easy. Simple method callings were all that were required. The algorithms, upon being implemented, processed all the data using all the internal calculations.

Testing Accuracy

To test the accuracy, we implemented different methods on different algorithms based on requirement. Some were direct accuracy-check method calls from scikit-learn libraries. While in some other algorithms, we implemented manual accuracy checks, again based on the algorithm itself. In Random Forest of instance, mean was calculated. In Artificial Neural Network, despite calling an accuracy-check method, all the accuracy from all the epochs were taken in for mean value of accuracy. The accuracy check is crucial in understanding the viability of the algorithms and also the research itself. A very low accuracy in all the algorithms would mean the entire research was a dead end. It would mean this method altogether is not viable for this research. A low accuracy in a few algorithms and a high accuracy in the others would mean the ones with the low accuracy are not efficient in this model, but the others are. We would have discarded the low accuracy yielding algorithms. However, in our case, all the algorithms yielded a very high accuracy. Although this issue did cause us a few problems later on when comparing the accuracy amongst all the algorithms, the fact that the accuracy is high on all, was a strong point in determining the methods to be viable in this field of research. An Effective Model is a model which basically predicts the testing data most accurately as compared to other models and hence, can be deployed successfully.

Data Post-Processing

After all the accuracy have been taken into account, a few other data processes can still be implemented. This part of the model is not necessary for the primary target of the research, but we still used it for certain confirmation purposes. We implemented a method which would create a confusion matrix.

METHODOLOGY

. In this paper crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. For that purpose, we are used artificial neural network (ANN). This project shows the ability of artificial neural network technology to be used for the approximation and prediction of crop yields at rural district

NEED OF CROP PREDICTION

Prediction of crop yield mainly strategic plants such as wheat, corn, rice has always been an interesting research area to agro meteorologists, as it is important in national and international economic programming. Dry farming crop production, apart from relationship to the genetic of cultivator, adaphic terms, effect of pests and pathology and weeds, the management and control quality during the growing season and etc. is severely depend to climatic events. Therefore, it is not beyond the possibility to acquire relations or systems which can predict the more accuracy using meteorological data.

Nowadays, there are a lot of yield prediction models, that more of them have been generally classified in two group a) Statistical Models, b)Crop Simulation Models (e.g. CERES). Recently, application of Artificial Intelligence (AI), such as Artificial Neural Networks (ANNs), Fuzzy Systems and Genetic Algorithm has shown more efficiency in dissolving the problem. Application of them can make models easier and more accuracy from complex natural systems with many inputs. In this research it has been tried to develop a wheat yield prediction model using ANNs. If we design a network which correctly learn relations of effective climatic factors on crop yield, it can be used to estimate crop production in long or short term and also with enough and useful data can get a ANNs model for each area. Furthermore, using ANNs can find the most effective factors on crop yield. Therefore, some factors that their measurements are difficult and cost effective can be ignored. In this the effect of climatic factors on wheat yield has only been applied.

IMPLEMENTATION

Artificial Neural Network(ANN)

We followed standard procedures for all the algorithms apart from the Artificial Neural Network (ANN). Starting from the data pre-processing, to algorithm fitting, to accuracy checking, and finally predicting the desired outcome, the methods and codes used nearly remained the same. The only major variation was in the part where we called the algorithm fitting functions. However, the deep learning model that we created for the ANN, was different to the other algorithms to a certain extent. We created the hidden layers and perceptions manually. Even after the predictions were being generated, we performed further checks in the form of k-Fold Cross Validation and Grid Search CV to ensure the highest viability of the accuracy obtained. The first step was to calculate the activation of one neuron given an input. The input was a row from our training dataset, a sin the case of the hidden layer. It might also be the outputs from each neuron in the hidden layer, in the case of the output layer. Neuron activation was calculated as the weighted sum of the inputs, much like linear regression.

$$activation = sum(weight_i * input_i) + bias$$

Where weight was a network weight, "i" was the index of a weight or an input and bias was a special weight that had no input to multiply with. Once a neuron was activated, we needed to transfer the activation to see what the neuron output actually was. Different transfer functions could be used. It was traditional to use the sigmoid activation function, but we could also use the tan h (hyperbolic tangent)function to transfer outputs. However, due to the recent popularity and increase in efficiency, the rectifier transfer function had been used by us in our deep learning model. The sigmoid activation function looks like an S shape; it is also called the logistic function. It can take any input value and produce a number between 0 and 1 on an S-curve.

It is also a function of which we can easily calculate the derivative (slope) that we use later when back propagating error.

The first step was to calculate the error for each output neuron, that gave us our error signal (input) to propagate backwards through the network. The error for a given neuron was calculated as

follows.

$$error = (expected - output) * transfer\ derivative(output)$$

Where "expected" was the expected output value for the neuron, "output" was the output value for the neuron and transfer_derivative() calculated the slope of the neuron's output value, as shown above. This error calculation was used for neurons in the output layer. The expected value was the class value itself. In the hidden layer, things are a little more complicated. The error signal for a neuron in the hidden layer was calculated as the weighted error of each neuron in the output layer. The error traveled back along the weights of the output layer to the neurons in the hidden layer. The back-propagated error signal was accumulated and then used to determine the error for the neuron in the hidden layer, as follows.

$$error = (weight\_k * error\_j) * transfer\_derivative(output)$$

Where"error_j"wastheerrorsignalfr omthe"j"thneuronintheoutputlayer,"weight _k" was the weight that connects the "k" th neuron to the current neuron and output was the output for the current neuron. The network was trained using stochastic gradient descent. This involves multiple iterations of exposing a training dataset to the network and for each row of data forward propagating the inputs, back propagating the error and updating the network weights. Once errors were calculated for each neuron in the network via the back-propagation method above, they could be used to update weights. Network weights were updated as follows.

$$weight = weight + learning\_rate * error * Input$$

Where "weight" was a given weight, "learning_rate" was a parameter that we specify, "error" was the error calculated by the back-propagation procedure for the neuron and "input" was the input value that caused the error. The same procedure might have been used for updating the bias weight, except there was no input term, or input was the fixed value of 1.0. Learning rate controls how much to change the weight to correct for the error. For example, a value of 0.1 will update the weight 10% of the amount that it possibly could be updated. Small learning rates are preferred that cause slower learning over a large number of training iterations. This increases the likelihood of the network finding a good set of weights across all layers rather than the fastest set of weights that minimize error**.** If errors were accumulated across an epoch before updating the weights, this would be called batch learning or batch gradient descent. Once all the aforementioned steps were completed, the network became eligible for training. The rest of the procedure followed a similar structure to the rest of the algorithms, as discussed above.

Algorithms for Proposed Model

Logistic Regression

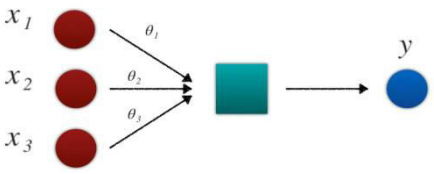


Figure: Flow Chart of Logistic

Regression

Methods involving regression are essential to any data analysis models which attempt to describe the association between a response variable and any number of predictor variables. Situations involving discrete variables constantly arise. For instance, the dataset we have implemented has an outcome involving the presence or absence of a particular crop, given a set of features. Logistic regression analysis extends the techniques of multiple regression analysis to investigate and inquire situations in which the outcome is categorical, which is, taking on multiple values. This is a very basic branch of data science. Although the name suggests a regression technique, logistic regression is a statistical classification model which deals with categorical dependent variables. Classification is decision. To make an optima decision we need to assess the utility function, which implies that we need to account for the uncertainty in the outcome ,i.e. a probability. Logistic regression is emphatically nota classification algorithm on its own. It is only a classification algorithm in combination with a decision rule that makes dichotomous the predicted probabilities of the outcome. This is one of the very first algorithm any machine learning practitioner attempts when faced with a classification problem. The basic mechanism and output of this algorithm is similar to many other machine learning algorithms. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous or binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

This algorithm works with binary data, where either the event happens, represented by “1”, or the event does not happen, represented by 0. So given some feature “X”, it tries to find out whether some event “y” happens or not. So “y” can either be “0” or “1”. In the case where the event happens, “y” is given the value “1”. If the event does not happen, then “y” is given the value of “0”. For example, if “y” represents whether a particular crop among a huge variety of crops, then “y” will be “1” if the crop does grow or “y” will be“0” if it does not. This is known as Binomial Logistic Regression. There is also another form of Logistic Regression which uses multiple values for the variable “y”. This form of Logistic Regression is known as Multinomial Logistic Regression. Figure3.3 shows a simple flow chart representation of logistic regression.

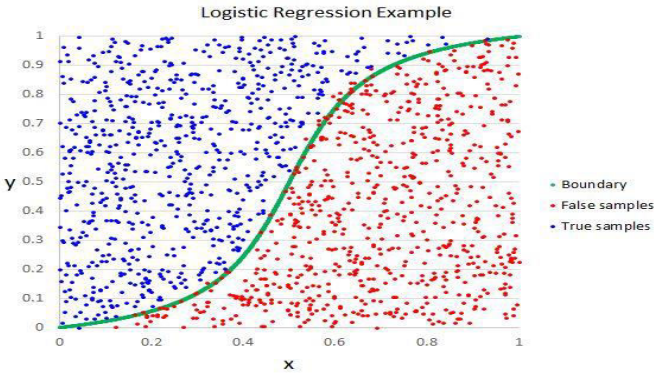


Figure: Classification of data by Logistic Regression.

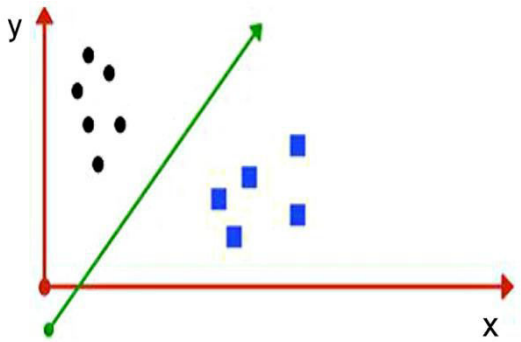
Logistic Regression uses the

logistic function to find a model that fits with the data points. The function gives an "S" shaped curve to model the data. The curve is restricted between "0" and "1", so it is easy to apply when "y" is binary. Logistic Regression can then model events better than linear regression, as it shows the probability for "y" being "1" for a given "x" value. Logistic Regression is used in statistics and machine learning to predict values of an input from previous test data. Scatter plot classification of data by Logistic Regression is shown in Figure

Support Vector Machine(SVM)

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which is avery useful technique for data classification. However, this learning algorithm can also be used for regression challenges. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables).

The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.



Linear SVM

Support Vector Machine classifier plots each data item with the value of each feature as a pointinann-dimensional space (where n is number of features) being the value of a particular coordinate. SVM maps data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Then, it performs classification by finding the hyper-plane that differentiates the two classes very well. points can be categorized, even when the data are not otherwise linearly separable. Then, it performs classification by finding the hyper-plane that differentiates the two classes very well.

When the data can be linearly separated in two dimensions, as shown in Figure3.5, any machine learning algorithm tries to find a boundary that divides the data in such a way that the mis-classification error can be minimized. Nevertheless, there can be several boundaries that correctly divide the data points as shown below in

VMisdifferentfromtheotherclassifie rsinthewaythatitchoosesthedecisionboundar ythatmaximizes the distance from the nearest data points of all the classes. This boundary has the maximum margin from the nearest points of the training class as well as the test class. As a result, SVM classifier does not only find a boundary; it finds the most optimal decision boundary. This boundary resulting from SVM is called the maximum margin classifier, or the maximum margin hyper plane. The nearest points from the hyper plane that maximizethe distance between the decision boundaries are called support vectors.

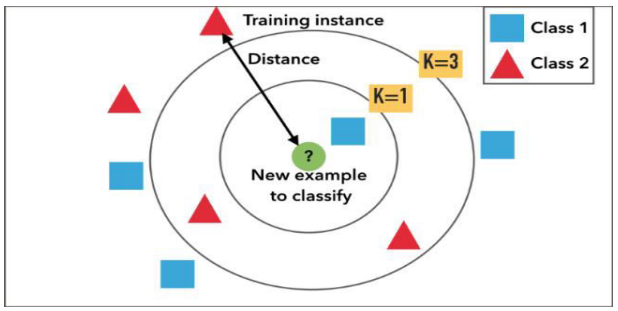


Figure: K-Nearest Neighbors Classification

K-Nearest Neighbors

K-Nearest Neighbors (KNN) algorithm is one of the simplest, easy to understand, versatile and one of the topmost machine learning algorithms. KNN is a non-parametric supervised learning algorithm. Additionally, it is an instance-based learning or a lazy algorithm. When a query to the database is made, the algorithm uses the training instances to spit out an answer. That is why, for KNN the training phase is very fast compared to other classifier algorithms. However, the testing phase becomes slower and costlier, that is in terms of time and memory.

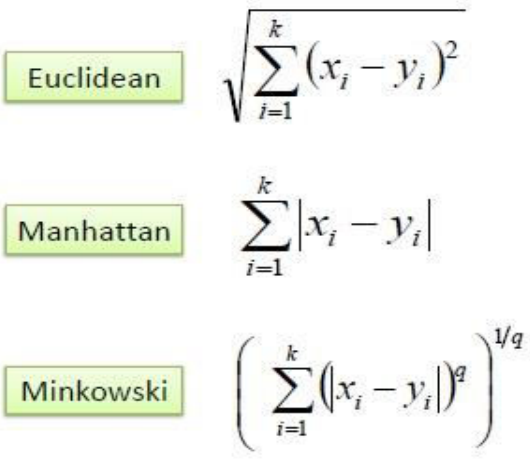


Figure: Formula for Distance Metric Calculation.

More formally, given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following three steps:

• It runs through the whole dataset computing d between x and each training observation. We'll call the K points in the training data that are closest to x the set A. Note that K is usually odd to prevent tie situations.

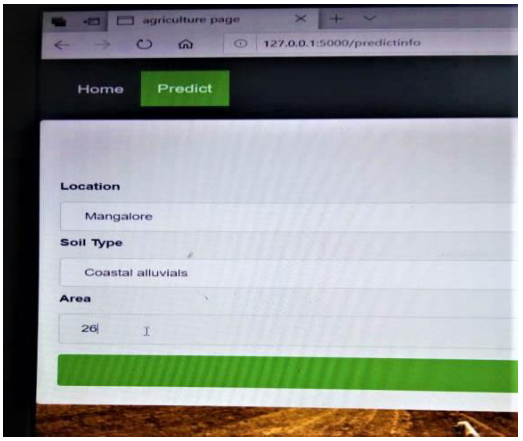
• It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label.

• Finally, the input x is assigned to the class with the largest probability.

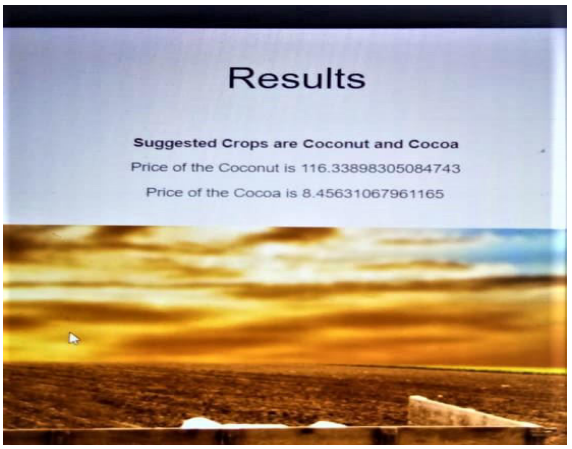
The number of neighbors (K) is a hyper-parameter that needs to be chosen at the time of model building. Research has shown that there is no optimal number of neighbors which suits all kind of data sets. Each dataset is different and needs to fulfill its own requirements. In most cases, it is better to choose it as an odd number if the number of classes is even. When the value of K is small, the region of a given prediction is being restrained. And the classifier is being forced to be "more blind" to the overall distribution. A small value for K provides the most flexible fit, which will have low bias but high variance. Graphically, the decision boundary will be more jagged, which is shown in the Figure below. On the other hand, a higher valued K averages more voters in each prediction. Hence it is more resilient to outliers. Larger

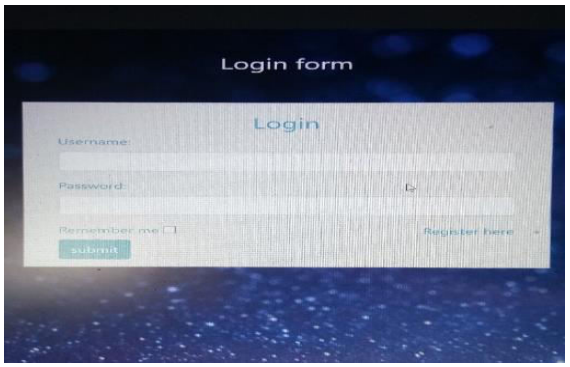
values of K will have smoother decision boundaries which mean lower variance but increased bias, shown below in the Figure below.

This graph clarifies that at K=1, the boundaries were being over-fitted. So, the error rate initially decreases and reaches minima. After the minima point, it then increases with increasing K. To get the optimal value of K, the training and validation can be segregated from the initial dataset. Then by plotting the validation error curve, it is easier to get the optimal value of K. This value of K should be used for all predictions.



VIII. RESULTS





CONCLUSION

RESULT ANALYSIS

The core aim of our research was to establish a model that will efficiently predict a particular crop based on a set of features. During the course of this research, we have also successfully created a model thatcan predict the possibility of a particular crop to be able to be grown, given a set of features. In order to do this, as mentioned above, we have implemented 6 different machine learning algorithms. Amongst them, The SVM was done using 3 different aspects, and the KNN was done using 2 aspects. This ensured us the ability to bring a comparison between varieties of different algorithms. Our target was to establish the best performing algorithm for this field of work, based on our data.

ACCURACY ANALYSIS

As mentioned before, we have extracted accuracy from 9 different processes. Only the ANN was run on one instance only. Apart from that, all the other algorithms were tested on a variety of instances of the same dataset. We took 5 samples of our primary dataset. The samples had an increasing number of rows starting from 2000 and ending at 11000. We ran each algorithm on these samples in 2 separate train test splits. Initially, we used the 60%-40%train test ratio. Eventually, we changed that to 80%:20%. Both gave us fair results, but the latter gave us a better accuracy, which led us to finalize the model on that. Two separate outcomes were extracted from our algorithms with slight tweaking. We could both predict a particular crop and the percentage chance of a specific crop to be able to be grown, given a particular set of features.

However, the former method fetched bad accuracy levels, which we found highly unconvincing. The highest accuracy we got was from the KNN (K=optimal) algorithm, which was 81.3%, on a set of 11000 features. In similar features, but with a specific crop as the dependent variable, all the algorithms gave strong accuracy levels. The lowest accuracy was given by Random Forest Classifier which was still 92.30%. The highest accuracy was again given by the KNN (K=optimal) algorithm. The ANN was tested only once with 11000 features, and only for the specific crop model.

We trained and tested the network up to 1000 epochs. The accuracy that we obtained was 96.95%. All the accuracy levels in this part of the thesis indicated to a strong viability of our research in this field. In the(Table. 3.1) and (Table. 3.2) below we have shown the comparison among all the classifiers we have implemented for specific crop possibility prediction and crop prediction respectively

Data	GNB	SVM	Linear SVM	RBF SVM	KNN (k=1)	KNN (k=optimal)	LR	RF	ANN
2000	80.61	95.07	92.97	94.38	90.44	95.50	95.22	95.07	
4000	84.26	94.95	96.35	94.53	93.26	95.65	96.07	95.16	
6000	89.04	97.28	97.09	97.84	96.16	96.44	95.78	95.78	
8000	85.36	95.27	95.59	95.74	95.59	96.81	95.31	93.71	
11000	95.5	97.0	97.2	96.23	96.38	97.70	97.38	92.30	96.95

Table Accuracy Comparison for Crop Possibility

Data	GNB	SVM	Linear SVM	RBF SVM	KNN (k=1)	KNN (k=optimal)	LR	RF
2000	70.18	70.50	65.44	67.13	65.44	66.01	62.58	30.04
4000	70/88	49.92	49.64	58.34	38.84	48.80	57.50	44.37
6000	55.97	63.12	67.50	62.92	54.02	56.46	52.52	45.41
8000	58.79	63.52	58.90	64.79	57.33	60.0	43.75	45.84
11000	56.2	63.7	56.8	68.8	61.38	81.3	66.46	55.0

REFERENCES

[1] C. Pu, “Evacuation Assisting Strategies in Vehicular Ad Hoc Networks,” in IEEE Proc. UEMCON, November 2018.

[2] Ecuador Earthquake: Death Toll Jumps to 272; More Than 2,500 Injured, https://www.cnn.com/2016/04/17/americas/ecuador-deadlyearthquake, Apr 18, 2016.

[3] H. Nishiyama, M. Ito, and N. Kato, “Relay-by-Smartphone: Realizing Multihop Device-to-Device Communications,” IEEE Commun. Mag., vol. 52, no. 4, pp. 56–65, 2014.

[4] Number of smartphone users in the United States from 2010 to 2022, 2018, https://www.statista.com/statistics/201182/forecast-ofsmartphone-users-in-the-us.

[5] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. V. Schreeb, “Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti,” PLoS medicine, vol. 8, no. 8, 2011.

[6] Z. Lu, G. Cao, and T. L. Porta, “Networking Smartphones for Disaster Recovery,” in Proc. IEEE PerCom, 2016, pp. 1–9.

[7] A. Pal and K. Kant, “E-Darwin2: A Smartphone Based Disaster Recovery Network using WiFi Tethering,” in Proc. IEEE CCNC, 2018, pp. 1–5.