

Methodology

Han bhai bore ho gye the to design laga diye



-

Joe mama

## Methodology

This section comprises of various research methodology and relevant methods of application implementation broken down in how we collect data which are used for the training and validation, how these can predict for the desired outcome. A web-based application approach is discussed, and suitable deep learning techniques are made use of – for plant disease detection while achieving high levels of accuracy.

## Data collection

The methodology includes a brief explanation of the process in data collection. For the fulfilment of this research comprehensive data gathering was done and most data were collected from Kaggle. Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore, and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.[1]  .

1. *PREDICTION SYSTEMS*

SOIL FERTILITY PREDICTION :

YIELD PREDICTION : The dataset consists of factors like temperature, rainfall, humidity, ph. The datasets have been obtained. The data set has large number of instance or data that have taken from the past historic data. It includes many parameters or features like the temperature, humidity, rainfall temperature, land type etc.

1. *RECOMMENDATION SYSTEMS*

FERTILIZER RECOMMENDATION SYSTEM:

IRRIGATION RECOMMENDATION SYSTEM:

CROP RECOMMENDATION SYSTEM:

1. *PLANT DISEASE DETECTION*

*LEAF DISEASE DATASET:*

## Prediction System

We have implemented our prediction using various machine learning algorithms.

1. k-Nearest neighbor (KNN) Algorithm

The methodology is wide used adopted thanks to its potency [9]. The key plan of the algorithmic rule is to categorize a brand new sample within the most frequent class of its nearest neighbor within the coaching set. This is often the foremost selection formula on the category labels of its neighbors. The k-nearest neighbor classification algorithmic rule may be divided into 2 phases: coaching section and testing section. KNN is similar to kernel methods with a random and variable bandwidth. The idea is to base estimation on a x th number of observations k which are closest to the desired point.

(Formula)

1. Support Vector Machine Algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm which is a very 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.

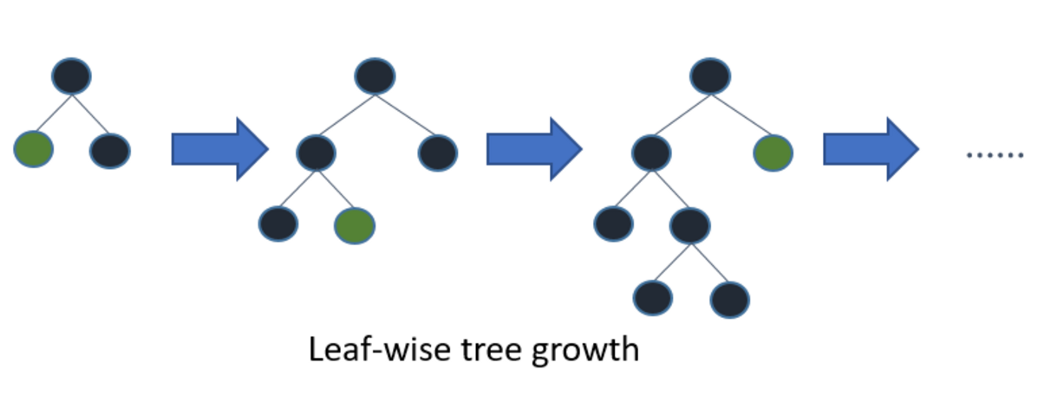
Support Vector Support vector regression is the natural extension of large margin kernel methods used for classification to regression analysis. It retains all the properties that characterize maximal margin algorithms of support vector machines such as duality, sparseness, kernel and convexity. It has become a powerful technique for predictive data analysis with many applications in varied areas of study like biological contexts, drug discovery, civil engineering, sunspot frequency prediction, image tracking, image compression etc.,. The problem of regression is that of finding a function which approximates mapping from an input domain to the real numbers on the basis of a training sample. This refers to the difference between the hypothesis output and its training value as the residual of the output, an indication of the accuracy of the fit at this point. To decide how to measure the importance of this accuracy, as small residuals may be inevitable even need to avoid large ones.

(Image)



1. LightGBM

Light GBM is a gradient boosting framework that uses tree based learning algorithm.Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.Light GBM can handle the large size of data and takes lower memory to run. Another reason of why Light GBM is popular is because it focuses on accuracy of results. LGBM also supports GPU learning and thus data scientists are widely using LGBM for data science application development.



1. *Soil Fertility Predictor*

Soil fertility is the ability of soil to sustain plant growth and optimize crop yield. This can be enhanced through organic and inorganic fertilizers to the soil. Nuclear techniques provide data that enhances soil fertility and crop production while minimizing the environmental impact.In soil fertility prediction, we take various attributes of the soil such as Ph, EC, OC, OM, N, P, K, etc. to determine if the soil is fertile or not.

Using the mentioned machine learning algorithms , soil fertility data set is analyzed and determined the optimal parameters for the predicting if the soil is fertile or not. Multiple linear regressions are used to find the significant attributes and form the equation for the yield prediction.

Here we had train the model using the training data set it is generated from the actual data set dividing it into the 75% of training data set and 25% of these data set 75% of data set is given to the different machine learning algorithms. After completion of the trainings data set is given to the trained model in that model is is tested with test data set it will produces the accuracy of the different model which is shown below.

(Model Table)

We analyze that k nearest neighbor algorithm algorithm hass produced 72% of accuracy shown in Table which is least among the three algorithms. Support vector machine produced 80% of accuracy which is comparatively acceptable as compared to k nearest neighbor algorithm. While Lightgbm outperformed both of them and got the best accuracy , so we used the same for soil fertility predcition.

1. *Yield Predictor*

Crop yield is a very useful information for farmers . It is very beneficial to know the yield which results in reduction in loss. In the past the yield prediction is done by experienced farmers. The proposed system also works in a similar way. It takes the previous information and uses it to predict the future yield.

The crop yield mainly depends on weather and pesticides. This prediction is proportional to the accuracy on information provided. Therefore, the proposed system predicts the yield and decreases the loss.

Using the mentioned machine learning algorithms , crop data set is analyzed and determined the optimal parameters for the crop production. Multiple linear regressions are used to find the significant attributes and form the equation for the yield prediction. This model is simple, does not required any sophisticated statistical tools, required data for crop growing periods, yield data for past years and provides marginally good prediction. Therefore it can be used for district, agro climatic zone and state level prediction.

Here we had train the model using the training data set it is generated from the actual data set dividing it into the 80% of training data set and 20% of these data set 80% of data set is given to the different machine learning algorithms. After completion of the trainings data set is given to the trained model in that model is is tested with test data set it will produces the accuracy of the different model which is shown below.

(Model Accuracy Table)

We analyze that support vector machine algorithm is produced 72% of accuracy shown in Table which is least among the three algorithms. In k nearest neighbor algorithm produced 80% of accuracy which is comparatively acceptable as compared to support vector machine algorithm. While Lightgbm outperformed both of them and got the best accuracy , so we used the same for crop yield prediction.

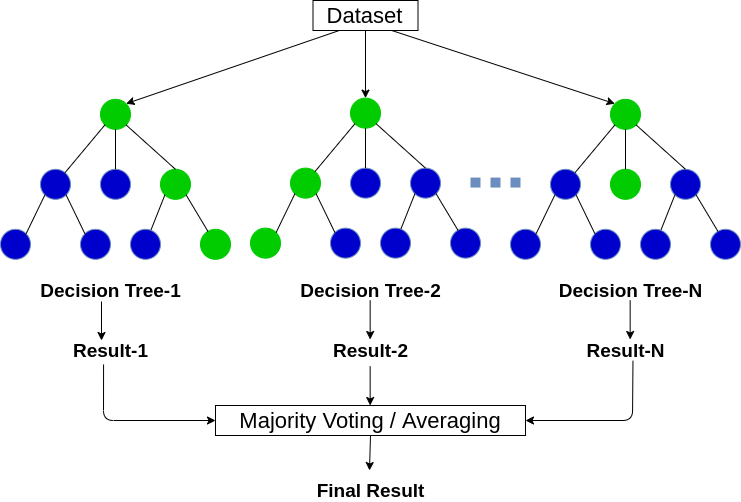
Recommendation System

We have implemented our prediction using various machine learning algorithms.

*1.Random Forest Algorithm*

Random Forest Machine Learning Algorithm Random forest is a supervised learning algorithm. As the name suggests, this algorithm creates a forest and using precision techniques, makes it random. The “forest” it builds, is an ensemble of Decision Trees, which are mostly trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction.

Random forest algorithm Random Forest is a ML algorithm. At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction.



1. *Gradient Boosting Algorithm*

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.The base estimator for the Gradient Boost algorithm is fixed and i.e. Decision Stump. However, if we do not mention the value of n\_estimator, the default value of n\_estimator for this algorithm is 100.

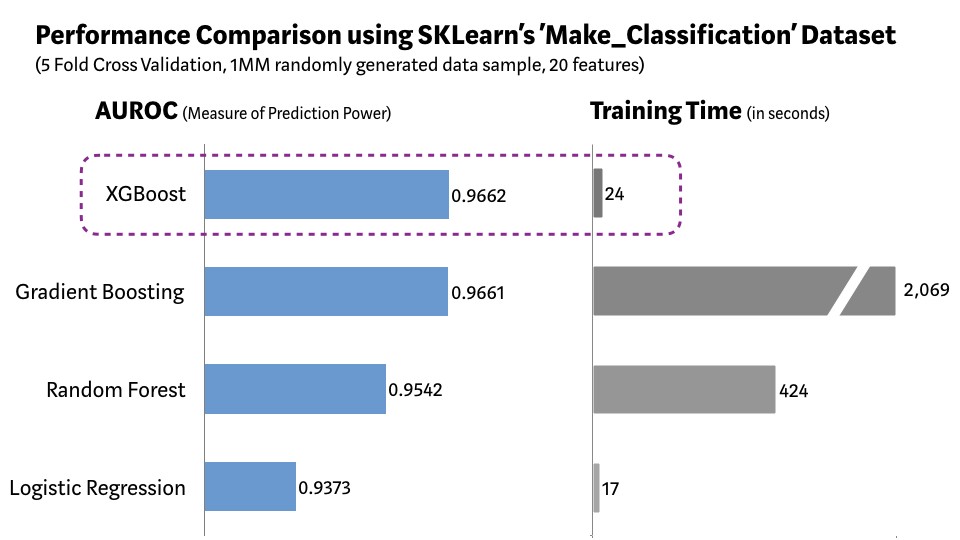
Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

(Formula)

*3.XGBOOST*

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.



As demonstrated in the chart above, XGBoost model has the best combination of prediction performance and processing time compared to other algorithms. Other rigorous benchmarking studies have produced similar results.

1. Fertilizer Recommendation System

The proposed system aims to estimate the nutrient content and recommend the suitable fertilizer to be used for higher productivity. Under application of fertilizer results in low yield due to insufficient nutrients present in the soil for the crop.Over usage of fertilizer results in soil pollution. The food products from the polluted soil will be food poisoning and health issues for the consumer .

In Fertilizer Recommendation System, we take various attributes like Temparature , Humidity , Moisture , Soil Type ,Crop Type,Nitrogen ,Potassium and Phosphorous.

Using the mentioned machine learning algorithms , fertilizer recommendation data set is analyzed and determined the optimal parameters for the predicting if the soil is fertile or not. Multiple linear regressions are used to find the significant attributes and form the equation for the yield prediction.

Here we had train the model using the training data set it is generated from the actual data set dividing it into the 75% of training data set and 25% of these data set 75% of data set is given to the different machine learning algorithms. After completion of the trainings data set is given to the trained model in that model is is tested with test data set it will produces the accuracy of the different model which is shown below.

(Model Table)

We analyze that gradient boosting algorithm hass produced 72% of accuracy shown in Table which is least among the three algorithms. Random Forest Algorithm produced 80% of accuracy which is comparatively acceptable as compared to gradient boosting algorithm. While XGBoost outperformed both of them and got the best accuracy , so we used the same for fertilizer recommendation system.

1. Crop Recommendation System

To compare and predict the initial data set, the environmental factors needed to be gathered. We will be will using a dataset that is already available at kaggle website.The gathered data are cleaned and processed by using clustering and other algorithms to pass the values to the next component for crop recommendation. Since the environmental conditions differ from region to region, a machine learning model is used to predict the best crop type for the selected land.

To train the crop recommending model with the data gathered, above metioned machine learning algorithms are used to identify the best crop to cultivate with the highest probability of growing. Random Forest , Gradient Boosting and XGBoost algorithms are used to select the best crop type. From this model, it decided what type of crops that the farmer should grow. This is done by analyzing factors of humidity, temperature, soil moisture, pH level, and sunlight. Mainly the system suggests 4 crop types by analyzing the above-mentioned machine learning algorithms.

Here we had train the model using the training data set it is generated from the actual data set dividing it into the 80% of training data set and 20% of these data set 80% of data set is given to the different machine learning algorithms. After completion of the trainings data set is given to the trained model in that model is is tested with test data set it will produces the accuracy of the different model which is shown below.

(Model Table)

We analyze that gradient boosting algorithm hass produced 72% of accuracy shown in Table which is least among the three algorithms. Random Forest Algorithm produced 80% of accuracy which is comparatively acceptable as compared to gradient boosting algorithm. While XGBoost outperformed both of them and got the best accuracy , so we used the same for crop recommendation system.

1. Irrigation Recommendation System

The first one is the crop field level where different sensors are deployed in the field. Various sensors like soil moisture (EC- 1258), soil temperature (DS18B20), air temperature (DHT11), and humidity (DHT11) are used to gather all these soil and environmental attributes.All these values are collected and the dataset is publicly available in kaggle.

The collected data are then feed to the ML-based model for analysis. This ML unit is the heart of this intelligent system, which has two sections. One is the regression model that is used to predict the soil and environmental parameters in advance. By doing so, it can be used effectively to improve the performance of the system. The parameters that are considered from forecasted weather data are the atmospheric pressure, precipitation, solar radiation, and wind speed. These predicted values are passed through a clustering model to reduce the predicted errors. The other ML-based model takes the results of the clustering model along with the forecasted weather data as input. This binary classification model categorizes the predicted samples into two predefined classes: irrigation required (Y) or not required (N). The results of these ML models are stored in the database for future actions.

Using the mentioned machine learning algorithms , irrigation recommendation data set is analyzed and determined the optimal parameters for the predicting if the land requires irrigation or not. Multiple linear regressions are used to find the significant attributes and form the equation for the yield prediction.

Here we had train the model using the training data set it is generated from the actual data set dividing it into the 75% of training data set and 25% of these data set 75% of data set is given to the different machine learning algorithms. After completion of the trainings data set is given to the trained model in that model is is tested with test data set it will produces the accuracy of the different model which is shown below.

(Model Table)

We analyze that gradient boosting algorithm hass produced 72% of accuracy shown in Table which is least among the three algorithms. Random Forest Algorithm produced 80% of accuracy which is comparatively acceptable as compared to gradient boosting algorithm. While XGBoost outperformed both of them and got the best accuracy , so we used the same for irrigation recommendation system.

## Plant disease detection

For the objective of this implementational research to be fulfilled, a detection system was made by using the concept of object detection. Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.[2]

The tool that was used for the real time detection of media was OpenCV. OpenCV (Open-Source Computer Vision Library) is an open-source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. Being a BSD-licensed product, OpenCV makes it easy for businesses to utilize and modify the code.

The library has more than 2500 optimized algorithms, which includes a comprehensive set of both classic and state-of-the-art computer vision and machine learning algorithms. These algorithms can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, extract 3D models of objects, produce 3D point clouds from stereo cameras, stitch images together to produce a high resolution image of an entire scene, find similar images from an image database, remove red eyes from images taken using flash, follow eye movements, recognize scenery and establish markers to overlay it with augmented reality, etc. OpenCV has more than 47 thousand people of user community and estimated number of downloads exceeding 18 million. The library is used extensively in companies, research groups and by governmental bodies.[3]

As for the detection system

## References

[1] [Kaggle - Wikipedia](https://en.wikipedia.org/wiki/Kaggle)

[2] [Object detection - Wikipedia](https://en.wikipedia.org/wiki/Object_detection)

[3] [About - OpenCV](https://opencv.org/about/)