**TITLE OF THE MAJOR PROJECT**

*A project report submitted in partial fulfillment of*

*the requirements for the award of the degree of*

**Bachelor of Technology**

in

**Computer Science & Engineering**

*Submitted by*

**Divya Mahakul (19btcse13)**

**Priti Smita Nayak (19btcse28)**

**Satyajit Sahoo (19btcse40)**

**Shiva Swain (20lbtcse08)**

*Under the Supervision of*

**Ms. Sushree Pradhan**

Assistant Professor



Department of Computer Science Engineering Sambalpur University Institute of Information Technology Burla, Odisha

Dec 2023



**Sambalpur University Institute of Information Technology**

**Jyoti Vihar, Burla, Odisha, Pin: 768019**

**CERTIFICATE**

This is to certify that the thesis entitled **‘HARVESTIFY’** submitted by **Student1**(**19btcse13), Student2**(**19btcse28)** and **Student3**(**19btcse40), Student4**(**20lbtcse08)** in partial fulfilment of the requirements for the award of Bachelor of Technology degree in Computer Science & Engineering at Sambalpur University Institute of Information Technology is an authentic work carried out by them under my supervision. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any Degree or Diploma.

Name and Signature of Student (s)

1.

2.

3.

4.

**This is to certify that the above statement made by the students is correct to the best of my knowledge.**

**Ms. Sushree Pradhan**

**Supervisor**

Assistant Professor

Department of Computer Science & Engineering

Sambalpur University Institute of Information Technology, Burla

**Counter-signed by**

**The project evaluation of the above candidates has been held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

| Dr. Sudarson Jena  HOD, Department of CSE&A | External Examiner |
| --- | --- |

**DECLARATION**

**I do hereby declare that the work embodied in this minor/major project report entitled “HARVESTIFY” is the outcome of genuine work carried out by us under the direct supervision of Ms. Sushree Pradhan,** Assistant Professor **, Department of Computer Science Engineering and Application is submitted by us to Sambalpur University Institute of Information Technology, Burla for the award of the degree of Bachelor of Technology. The work is original and has not been previously formed the basis for the award of any other degree or diploma.**

**Date:**

**Place:**

Divya Mahakul (19btcse13)

Priti Smita Nayak (19btcse28)

Satyajit Sahoo (19btcse40)

Shiva Swain (20lbtcse08)

**ACKNOWLEDGEMENT**

I feel honored to avail myself of this opportunity to express my deep sense of gratitude to my guide Guide Name, Department of Computer Science Engineering, Sambalpur University Institute of Information Technology, Burla, Odisha, India for his valuable inspiration, guidance, and warm encouragement throughout my research work. His profound knowledge and timely advice were very much helpful in my research work without which my thesis could never be able to see this day. Proud to be work under him and he has given me every freedom for working in this Project.

Divya Mahakul (19btcse13)

Priti Smita Nayak (19btcse28)

Satyajit Sahoo (19btcse40)

Shiva Swain (20lbtcse08)

**List of Figures:**

**CONTENTS page no**

**ABSTRACT** 1

1. Section Name 2

1.1 Sub section name 3

1.2. Sub section name 4

2. Section Name 5

2.1. sub Section name 6

2.2. sub Section name 7

2.3. sub Section name 8

LIST OF FIGURES 

Fig.2.1. Title of figure 5

Fig.2.2. Title of figure 11

Fig.2.3. Title of figure 12

Fig. 3.1. Title of figure 15

Fig. 3.2. Title of figure 16

**Abstract**

Agricultural industry plays a major role in the process of economic development as well as the Gross Domestic Product of Sri Lanka. One of the significant issues in the industry is lacking an accurate way to identify the best crop that can be grown with the available soil fertility in a particular land. Since most of the farmers have a lack of knowledge about soil nutrients, they start cultivations by believing myths in society and few of them use scientific approaches. This research mainly focuses on suggesting the best crop according to soil fertility of land and also it recommends a fertilizer plan to optimize the amount of fertilizers applied for suggested crops. The paper presents a tool with embedded sensors that measure soil fertility and developed a cross- platform mobile application to suggest the best crops according to available soil fertility. Further, a fertilizer plan will be suggested to optimize fertilizer usage in order to increase profitability and avoid soil degradation. Also In this paper, we propose a system that will use the different techniques of the image process to both analyze and detect the plant diseases. The results of the implementation show that the designed system could give a successful result by detecting and classifying the plant diseases. This project aims at designing a standalone application that will provide the framer with the necessary information about the crop, fertilizer and type of plant disease. The purpose of this project is to assist and provide efficient support to the monoculture farmers.

**Keywords**: Image Processing; Plant Disease Detection; Standalone Application; crop; soil fertility.

**INTRODUCTION**

Agriculture plays an essential part in an economy’s life. They are the backbone of our country’s economy system. One of the key problems confronting farmers is selecting the right crop for cultivation. Selection of crops is determined by several factors such as temperature, soil composition, market prices etc. Machine Learning is a technique that uses complex algorithms and a collection of predefined rules to operate intelligently. It uses past data to read the patterns and then perform the intended task according to the defined rules and algorithms based on the analysis it produces. Machine Learning is an imminent field of informatics that can be applied quite efficiently to the agricultural sector. Machine Learning is everywhere throughout the entire growing and harvesting cycle. The major factors affecting crop yield are the soil type, land type and the macronutrients present in the soil. The purpose of this work is to categorize the soil samples according to the macro nutrients found there in and to predict the crops which can be grown in the soil.In this project a system is developed in which Voting Based Ensemble Classifier is applied to recommend the appropriate crops. This system also proposes the required fertilizer to boost the nutrients contained in the soil and thus enhance the yield of the crop. Thus there arises a need for suggesting suitable crops and fertilizers using machine learning algorithms.

Image processing and machine learning models can be employed for the detection of plant diseases. In this project, we have described the technique for the detection of plant diseases with the help of their leaves pictures. Image processing is a branch of signal processing which can extract the image properties or useful information from the image. Machine learning is a sub part of artificial intelligence which works automatically or give instructions to do a particular task. The main aim of machine learning is to understand the training data and fit that training data into models that should be useful to the people. So it can assist in good decisions making and predicting the correct output using the large amount of training data

literature survey

Crop Predic

1. A review by N. Balakrishnan et al. (2020) titled "Crop yield prediction using machine learning techniques: A comprehensive review" provides an overview of the recent advancements in crop yield prediction using machine learning techniques. The authors discuss various machine learning algorithms and techniques such as regression analysis, neural networks, decision trees, and support vector machines used for crop yield prediction.

Fertilizer Prediction:

1. An article by K. Ramakrishnan et al. (2020) titled "Fertilizer recommendation system for sustainable agriculture using artificial intelligence techniques: A review" discusses the use of artificial intelligence techniques such as decision trees, fuzzy logic, and neural networks for fertilizer recommendation. The authors review various studies that have implemented these techniques for fertilizer prediction and recommendation, highlighting their effectiveness in optimizing fertilizer use.

Plant Disease Detection:

1. A review by S. Singh et al. (2020) titled "Plant disease detection using machine learning techniques: A review" provides an overview of the recent advancements in plant disease detection using machine learning techniques. The authors discuss various machine learning algorithms such as support vector machines, decision trees, and convolutional neural networks used for plant disease detection. They also highlight the importance of deep learning techniques for accurate and efficient plant disease detection.

Overall, these literature surveys suggest that machine learning and deep learning techniques are effective for crop prediction, fertilizer recommendation, and plant disease detection. These technologies can greatly benefit farmers by providing them with accurate and timely information to improve crop yield and reduce crop loss due to diseases.

**PROBLEM STATEMENT**

The prediction of crop yield and fertilizer recommendation problem aims to provide accurate recommendations to farmers about the appropriate type and amount of fertilizer to use to optimize crop growth and yield. This involves analyzing a range of factors, such as soil quality, weather patterns, and pest infestations, to predict crop yields and determine the best fertilizer strategy. The ultimate goal is to help farmers improve their crop yields and profitability while minimizing environmental impact. Various technologies and strategies, such as soil testing, weather monitoring, and machine learning algorithms, can be used to address this problem, also plant diseases are impacting the agriculture in general and the monoculture more precisely. The detection of plant diseases most of the time is done only throughout the naked eye observation. This is a traditional way to deal with the detection of plant diseases, the farmers used to consult the experts who spend a lot of time trying to specify the type of disease that is affecting the plants’ leaves. The experts based their analytics that includes categorizing andclassifying the diseases, only on the visual symptoms that are usually shown on the leaves. The thing that is time consuming and costs a lot for the framers of this kind of agriculture. This operation of detecting plant diseases is expensive since the framers must consult experts, and time consuming as those experts need time to detect the disease and classify its type and need a continuous control. Knowing that some farmers could not have the capabilities to consult the experts regularly, so the risk of contagion between plants is extremely high. Also, this will have bad impacts on the environment due to the extensive use of chemicals and pesticides that are used in a random quantity, also, the production process is affected especially that in this case we are dealing only with one type of crops.

**SUGGESTED SYSTEM**

A crop and fertilizer recommendation system is a type of artificial intelligence application that helps farmers optimize crop yield and improve soil health by providing recommendations for crop selection and fertilization.

These systems use data analysis and machine learning algorithms to process large amounts of data about soil composition, weather patterns, crop types, and fertilizer types to generate recommendations for farmers.

As technology and data science continue to advance, the accuracy and usefulness of these systems are likely to improve. In the coming years, we can expect to see more sophisticated crop and fertilizer recommendation systems that take into account even more factors, such as local market conditions and pest and disease risks.

Ultimately, these systems have the potential to revolutionize agriculture by helping farmers improve crop yields, reduce waste, and promote sustainable farming practices.

As a suggested solution for this issue that every farmer is facing, we are suggesting the use of the latest technologies to detect the plant disease detection which is image processing. He will be uploading the image of the leaf to the application, once the image is processed and the detection is executed, the application will display to the farmers the type of disease along with the affected region, and the accuracy. The core idea behind this system is to remedy the disease with minimum impact on the environment, and to guarantee for the user or the farmer a fast and economical way in detecting leaf disease and categorizing it

**Proposed Method**

**Data collection:-**

Data Sources:

The data for our project was collected from various sources. The crop recommendation and fertilizer suggestion datasets were custom-built, whereas the disease detection dataset was obtained from a publicly available dataset. The details of each dataset are as follows:

Crop Recommendation Dataset:

1. The crop recommendation dataset was created by augmenting datasets of rainfall, climate, and fertilizer data available for India. The dataset consists of six features: Nitrogen (N), Phosphorous (P), Potassium (K) content in soil, temperature, humidity, and pH value of the soil. The data was collected from various sources and verified for accuracy before being used for the project.

Fertilizer Suggestion Dataset:

1. The fertilizer suggestion dataset was also custom-built and consists of five features: Crop, Nitrogen (N), Phosphorous (P), Potassium (K), pH value of the soil, and soil moisture. The data was collected from various sources and verified for accuracy before being used for the project.

Plant Disease Detection Dataset:

1. The plant disease detection dataset was obtained from a publicly available dataset that consists of 38 different classes of healthy and diseased crop leaves. The dataset contains approximately 87,000 RGB images of healthy and diseased crop leaves. The dataset was augmented offline to increase the number of images and improve the quality of the dataset.

Data Collection Process:

The data collection process for the project involved the following steps:

1. Identify the data requirements for the project.
2. Search for publicly available datasets or build custom datasets.
3. Verify the accuracy and authenticity of the data collected.
4. Preprocess the data to remove any outliers or inconsistencies.
5. Augment the dataset to increase the number of samples and improve the quality of the dataset.
6. Divide the dataset into training and validation sets.

Data collection links

* [Crop recommendation dataset](https://www.kaggle.com/atharvaingle/crop-recommendation-dataset) (https://www.kaggle.com/atharvaingle/crop-recommendation-dataset)
* [Fertilizer suggestion dataset](https://github.com/Gladiator07/Harvestify/blob/master/Data-processed/fertilizer.csv) (https://github.com/Gladiator07/Harvestify/blob/master/Data-processed/fertilizer.csv)
* [Disease detection dataset](https://www.kaggle.com/vipoooool/new-plant-diseases-dataset) (https://www.kaggle.com/vipoooool/new-plant-diseases-dataset)

Data preprocessing

Data preprocessing is the initial step in any data analysis project. It involves cleaning, transforming, and organizing raw data to make it suitable for analysis. The main goal of data preprocessing is to prepare the data in such a way that it is easily understandable and interpretable by machine learning algorithms.

Here are some common data preprocessing techniques:

1. Data cleaning: Data cleaning involves handling missing values, removing duplicates, handling outliers, and correcting data inconsistencies. This step ensures that the data is accurate and reliable.
2. Data normalization: Data normalization involves scaling the data to a specific range, typically between 0 and 1. This step ensures that all features contribute equally to the analysis and prevents features with large values from dominating the analysis.
3. Feature selection: Feature selection involves identifying the most relevant features for the analysis and removing irrelevant or redundant features. This step helps to reduce the dimensionality of the data and improve the accuracy of the analysis.
4. Feature engineering: Feature engineering involves creating new features from existing ones or transforming the existing features to improve their predictive power. This step helps to increase the accuracy of the analysis by providing more relevant information to the machine learning algorithms.
5. Data integration: Data integration involves combining data from multiple sources into a single dataset. This step helps to provide a more comprehensive view of the data and can improve the accuracy of the analysis.
6. Data reduction: Data reduction involves reducing the size of the dataset by sampling or summarizing the data. This step is useful for large datasets that are difficult to analyze and can improve the performance of machine learning algorithms.

Overall, data preprocessing is an important step in any data analysis project, as it helps to ensure that the data is accurate, reliable, and suitable for analysis.

3.Missioning value identification

Missing value identification is the process of identifying if there are any missing values in the dataset. A missing value is a cell in the dataset that has no value assigned to it. This can occur due to various reasons, such as incomplete data collection or errors during data entry.

To identify missing values in a dataset, we can use various techniques such as:

1. Visual inspection: We can visually inspect the dataset by looking for cells that have no value assigned to them. In some cases, missing values may be represented by a specific symbol, such as "NA" or "NaN".
2. Statistical methods: We can use statistical methods to detect missing values in a dataset. For example, we can calculate the mean, median, and standard deviation of each column and look for any values that are significantly different from these measures.
3. Software tools: There are various software tools available that can help identify missing values in a dataset. For example, the pandas library in Python has functions such as isnull() and notnull() that can be used to identify missing values.

Once missing values are identified, we can either remove them or impute them. Removing missing values can result in loss of data, and therefore imputation is often preferred. Imputation involves filling in the missing values with some appropriate value, such as the mean or median of the column or using more advanced imputation techniques such as regression or k-nearest neighbors (KNN) imputation.

Feature Selection:-

Feature selection is a process of selecting relevant features from a larger set of features or variables that are available in a dataset. It is an important step in machine learning as it helps to improve model performance by reducing overfitting, improving accuracy, reducing training time, and reducing the complexity of the model.

There are several methods for feature selection, including:

1. Correlation analysis: This involves examining the correlation between each feature and the target variable. Features with high correlation are retained while those with low correlation are dropped.
2. Filter methods: These methods use statistical measures like chi-square, information gain, and correlation coefficient to rank the importance of features. The top-ranked features are retained.
3. Wrapper methods: These methods involve the use of a specific machine learning algorithm to evaluate subsets of features. The algorithm is trained and tested on different subsets of features, and the best subset is retained.
4. Embedded methods: These methods involve feature selection as part of the model building process. The algorithm automatically selects the best features as it builds the model.

The choice of feature selection method depends on the type and size of the dataset, as well as the nature of the problem being solved. A good feature selection method should strike a balance between reducing the dimensionality of the dataset while retaining the most relevant features for the problem at hand.

Splitting of Data:-

After data preprocessing, the next step is to split the data into training and testing sets. This is important to ensure that the machine learning model is trained on a portion of the data and tested on another portion that it has not seen before. This helps to evaluate the model's ability to generalize to new data.

The most common way to split the data is to use a 70/30 or 80/20 split for training and testing, respectively. This means that 70% or 80% of the data is used for training, while the remaining 30% or 20% is used for testing.

It's important to note that the data should be split randomly to avoid any bias in the training or testing set. This can be done using the train\_test\_split function in Python's scikit-learn library.

Machine learning classification:-

Machine learning classification is a type of supervised learning in which an algorithm is trained on a labeled dataset to predict the class label of new, unseen data points. The goal of classification is to find a decision boundary that separates different classes in the feature space.

There are several types of classification algorithms, including:

1. Logistic Regression: A statistical method used to model the relationship between a categorical dependent variable and one or more independent variables. It assumes a linear relationship between the dependent variable and the independent variables, and uses a logistic function to predict the probability of a particular outcome.
2. Naive Bayes: A probabilistic algorithm that assumes that the features are independent of each other given the class label. It calculates the probability of each class label given the feature values and selects the class with the highest probability.
3. Decision Trees: A method that recursively partitions the feature space into smaller regions based on the values of the features. It constructs a tree-like model of decisions and their possible consequences, including chance events and resource costs, to determine a course of action.
4. Random Forest: An ensemble method that constructs multiple decision trees using random subsets of the features and data points. It aggregates the predictions of the individual trees to produce a final prediction.
5. Support Vector Machines (SVM): A method that finds the hyperplane that maximally separates the different classes in the feature space. It maps the data to a high-dimensional space and finds the optimal hyperplane that maximizes the margin between the two classes.
6. Neural Networks: A method inspired by the structure and function of the human brain that consists of a series of interconnected layers of nodes that process the input data and produce an output. It can be used for classification tasks by using a softmax function to produce class probabilities.

Each classification algorithm has its strengths and weaknesses, and the choice of algorithm depends on the specific problem and data at hand.

**MODELING AND ANALYSIS**

**Linear Regression:**

**L**inear regression is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called dependent variable. The variable you are using to predict the other variables value is called the independent variable. Linear regression makes predicts for continuous / real or numeric variables such as sales,salary,age,product price etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a sloped straight line representing the relationship between the variables.

**Random Forest:**

Random forest is an ensemble learning method for classification. This algorithm consists of trees and the number of tree structures present in the data is used to predict the accuracy. Where leaves are corresponding to the class labels and attributes correspond to internal node of the tree. Here number of trees in forest used is 100 in number and Gini index is used for splitting the nodes.

**Naive Bayes:**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

**Decision Tree**:

Decision tree is non parametric classifier in supervised learning. In this method all the details are represented in the form of tree, where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. We have used Gini Index for splitting the nodes.

**SVM:**

SVM is supervised learning algorithm used for classification. In SVM we have to identify the right hyper plane to classify the data correctly. In this we have to set correct parameters values. To find the right hyper plane we have to find right margin for this we have choose the gamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads to miss classification.

**CONCLUSION**

The model predicts best crop that should be grown on land with less expenses among a number of crops available after analyzing the prediction of parameters. To the best of studies, there is so much work in existence that uses the same techniques in predicting the crops. Hence it is concluded that there is enhancement in the accuracy of this research when compared to the existing work that used another technique for prediction of crops, and it is provides the plant disease detection using image processing is the most convenient method to keep up with an efficient yield. The main goal of this paper was to prove how image processing tool is able to guarantee accurate results regarding the detection of the plant diseases and also how it can assist the farmers in increasing the yields. By the end of this project, we were able to achieve the intended objective which was the implementation of image processing in the plant disease detection, Prediction of Crop and Fertilizer Recommendation.

**SCOPE OF THE PROJECT**

**The scope of a crop and fertilizer recommendation and plant disease detection system project involves data collection on soil composition, weather patterns, crop types, fertilizer types, and disease symptoms, data processing using machine learning and data analysis algorithms, and providing recommendations on crop selection, fertilizer use, and disease management. The system would also feature a user-friendly interface and be deployed in the field for testing and evaluation. The project has the potential to significantly improve agricultural productivity and sustainability, reduce waste, and promote more efficient use of resources.**