**CROP DAMAGE PREDICTION**

**GROUP-28**

**ARCHANA - 2103A52140**

**NITHYA - 2103A52181**

**VAISHNAVI - 2103A52182**

**ABSTRACT:**

Everyone know that Agriculture is the back bone of our country. Cultivating crops i.e farming plays an important role all over the world. Without farming there will be no scope for surviving .

But now-a-days crop damage became a major problem for many of the farmers. This is happening due to many factors like Temperature, Season, Types of crops growing and many other factors. Many of the cultivators are not showing interest towards agriculture due to these factors which are becoming the major problems for getting less yield. So to get proper yield with in a desired time farmer should know the consequences before selecting the crop.This research will help the farmer to know about the crop production rating before the farmer gets crop yield. Depending on the parameters which have been provided by the farmer we can predict the output in the form of accuracy using some classification algorithms which will give the farmer, whether the farmer will get good yield or damaged crop.The analysis of environmental,usage of pesticides and fertilizers can predict crop disease, increase crop yield, along with that it increase the quantity and the quality of agricultural output less input.It helped to identify how the technology of artificial intelligence helps to improve the crop yield. The research study clearly gives the idea and need of machine learning in the field of agriculture. It also shows how it outperforms the other networks such as Logistic regression,KNN,SVM,Decision tree,Random forest. The results were calculated and with the obtained outcome future perspectives can be drawn.

**INTRODUCTION**

Agriculture is an important occupation all over the world.Agriculture will help the country to grow its economic status.It is the large commercial sector among all other sectors.One can definitely get more profits if they had a good command in farming[1].As the population is growing across the world,to meet everyone’s needs the crop production should be increased.There is huge demand for agriculture lands now-a-days.But there is lack of technology in agriculture sector so,we need to improve our technologies to get more yield in short span[2].When the farmers get awareness regarding the technology then there will be more benefits to farmers as well as our country economic condition.We all know that our country will do more and more imports than exports this will effect the country’s development.So,when we have a good crop production we can reduce the imports.Hence,to improve crop production and predict the crop damage before its yielding time we came with a solution using machine learning.Machine Learning techniques can help agriculture experts make early decisions during complex situations[3].Classifying the stages of crop production is a challenging task,but this can be outstandingly handled by ML Classification techniques.Some of the standard methods used for classification are Logistic Regression and KNN etc…Agriculture holds a crucial position in India's economy, being the most extensive sector and a key driver of the nation's development[4].With over 60% of the land dedicated to agriculture, it serves as the backbone for catering to the necessities of India's vast population of 1.3 billion people[5]. Therefore, the adoption of modern agricultural technologies becomes imperative, as it can pave the way for increased profitability among our country's farmers.Farmers rely on their past experiences and the preferences of neighboring farms to decide which crops to plant on their land[6]. They lack sufficient knowledge about soil nutrient levels, including nitrogen, phosphorus, and potassium. In this scenario, crop rotation is not practiced, and inadequate nutrient application can result in reduced yields, soil pollution through acidification, and damage to the topsoil layer.Considering these problems into an account we just implemented a model which would be farmer friendly and with this we can even develop our country’s economy[7].Machine learning (ML) has revolutionized the agriculture sector, presenting innovative possibilities for data-intensive scientific applications within the multidisciplinary field of agri - technology. ML, a subset of artificial intelligence, has seamlessly integrated with big data technologies and high-performance computing to unlock fresh avenues for agricultural advancement.In the realm of agriculture, machine learning is not an enigmatic illusion or a magical solution; rather, it comprises well-defined models designed to gather specific data and employ precise algorithms to attain predetermined outcomes.These input data attributes are utilized in machine learning predictive algorithms such as Support Vector Machine (SVM)and Decision Tree[8] .These algorithms are employed to detect underlying patterns within the dataset and subsequently make informed decisions based on the specified input conditions.The existing solution available in now –a-days is created by the experts on considering the present demanding crops and they worked on them by taking the factors like climatic conditions ,soil type and availability of water etc, to know whether the crop will sustain or not[9].

**LITERATURE REVIEW**

As we have seen above results we came to know that there is no model which is giving best accuracy to all the crops.By the survey we concluded that the people who had done all the experiments were did only on specific crops using specific parameters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Authors** | **Parameters Used** | **Model Used** | **Merits** | **Limitations&Drawbacks** |
| Yelamarthi et al.(2020)[1] | 1.kernel  2.normalize | 1.Support Vector Machine  2.Linear Regression | 1.Improved Accuracy  2.Data-driven Insights  3.Real-time Monitoring  4.Precision Agriculture  5.Reduced Resource  6.Waste Scalability | 1.Data Dependency Complexity and Expertise  2.Initial Investment Uncertainty and Risk |
| A Priyadharshini et al.  (2021)[2] | 1.max\_depth  2.splitter  3.metric  4.normalize  5.alpha  6.norm  7.epochs  8.degree | 1.Decision Tree  2.K-Nearest Neighbour  3.Linear Regression  4.Naïve-Bayes  5.Neural Network  6.Support-Vector Machine | 1.Personalized 2.Recommendations,Increased 3.Yield and Profitability,Risk 4.Mitigation,Time 5.Savings,Continuous Learning | 1.Initial Investment  2.Complexity  3.Environmental Impact  4.Market Fluctuations |
| GFenu et al.(2020,January 21)[3] | 1.kernel  2.ReLU  3.f1 score | 1.Support-Vector Machine  2.Artificial-Neural Networks  3.Logistic Regression | 1.Early Detection  2.Accuracy  3.Reduced Costs  4.Sustainable Practices | 1.False Positives and Negative,  2.Complexity  3.Data Availability  4.Lack of Real-time Data |
| A.Sharma et al.  (2021)[4] | 1.gamma  2.pruning parameters  3.bootstrap | 1.Support Vector Machine  2.Decision Tree  3.Random Forest | **1.Optimized Resource Allocation**  **2.Real-time Monitoring**  **3.Reduced Environmental 4.Impact**  **5.Scalability**  **6.Real-time Monitoring** | 1.Initial Investment  2.Market and Economic 3.Factors,Ethical and Privacy  Concerns  4.Adaptation to Dynamic Conditions  5.Technology Dependence |
| MK.Dharani et al.  (2021)[5] | 1.learning rate  2.padding  3.batchsize | 1.Artificial Neural Network  2.Convolutional Neural Networks  3.Recurrent Neural Network  **N**eural **N**etwork | **1.HighPrediction Accuracy**  **2.Non-linear Relationships**  **3.Feature Extraction**  **4.Transfer Learning**  **5.Temporal and Spatial Patterns** | 1.Overfitting  2.Black Box Nature  3.Ethical and Bias Concerns  4.Dependency on Historical Patterns  5.Complexity |
| A.Nigam et al.  (2020)[6] | 1.max\_features  2.lambda  3.weighted KNN  4.precision. | 1.Random-Forest  2.XGBoost  3.KNN  4.Logistic Regression | 1.Accurate Predictions  2.Data-Driven Insights  3.Early Warning  4.PrecisionAgriculture  5.Optimized Resource Allocation | 1.Model Complexity  2.External Factors  3.Resource Intensive  4.Privacy Concerns  5.Dependency-on Historical Data |
| R.Medar et al.  (2019)[7] | 1.smoothing  2.metric parameters | 1.Naive Bayes  2.K-Nearest Neighbour | 1.Precision Agriculture  2.Informed Decision-Making  3.Adaptability  4.Scalability  5.Data-Driven Accuracy  6.Early-Warning System | 1.Data Quality and Availability  2.Resource Intensiveness  3.Privacy Concerns  4.Local Variability |
| YJN.Kumar et al.  (2020)  [8] | 1.random state  2.max\_depth | 1.random Forest  2.Decision Tree | 1.Accurate Predictions  2.Precision Agriculture  3.Applicability to Various Crops  4.Informed Decision-Making | 1.Data Quality and Quantity  2.Model Complexity  3.External Factors  4.Bias in Data  5.Feature Selection |
| D.Thomas  (2020)[9] | 1.meta regressor  2.solver  3.Feature scaling | 1.Stacked Regression  2.Kernel Ridge  3.Lasso | 1.Diverse-Data Sources  2.Accurate Insights  3.Localized Solutions  4.Mitigating Risks  5.Government Policy Impact | 1.Data-Collection Challenges  2.Smallholder Farming  3.Changing Patterns  4.Resource Constraints  5.Interpretability |
| Akshay Dhande et al.(2019)[10] | 1.accuracy.  2.precision  3.recall | 1.Deep convolutional model(DCN) | 1.Predict crop images with high efficiency | 1.Time consuming  2.Deep models are prone to overfitting. |
| Nischitha K(2020)[11] | 1.Accuracy  2.F1 score | 1.SVM  2.Decision tree | 1.Recommends crops for farmers  2.Recommends the amount of nutrients to be added for the predicted crop | 1.Data quality,Bias  2.Overfitting at times |
| Tiago Domingues et al.(2022)[12] | 1.Pooling layer  2.Activation functions  3.Thresholding  4.kernel  5.min\_samples\_leaf | 1.CNN  2.LSTM  3.Object segmentation  4.SVM  5.RF | 1.Early disease detection  2.Precision agriculture | 1.Ml-models performance is influenced by the quality and type of input images |
| Donggeun Han et al.(2023)[13] | 1.Accuracy  2.F1 score | 1.Equilibrium displacement model  2.KNN | 1.Awareness&preparedness | 1.Complexity and data challenge  2.Lack of resources |
| Thuanha et al.(2022)[14] | 1.NDVI  2.AUC metrics | 1.Equilibrium displacement model  2.KNN | 1.Awareness and preparedness | 1.Complexity and data challenge  2.Lack of resources |
| Asif Ali et al.  (2013)[15] | 1.F-static  2.Residuals  3.R-squared | 1.Linear regression  2.Hyperbolic regression  3.Sigmoidal regression | 1.Timely Intervention  2.Higher Crop Yields  3.Resource Efficiency  4.Improved Decision-Making | 1.Unforeseen Factors  2.Inaccurate Early Estimations  3.Data Requirements  4.Model Complexity |
| G.Belanger et al.(2006)[16] | 1.Ginni coefficient loss | 1.Random Forest | 1.Yield Optimization  2.Climate Resilience  3.Understanding Vulnerability  4.Adaptive Management | 1.Changing Climate  2.Resource and Funding Constraints  3.Unpredictable Weather Events  4.Practical Application |
| Sathwik Gara et al.(2021)[17] | 1.Pooling layer  2.Learning rate  3.max\_depth | 1.CNN  2.RNN  3.Decision Tree | 1.Improved Decision-Making  2.Enhanced Resource  Management  3.Higher Crop Yield and Quality  4.Real-Time Monitoring: | 1.High-Initial Investment  2.Complex Integration  3.Data Privacy and Security  4.Expertise Requirement |
| Johnathn shook et al.(2021)[18] | 1.kernel  2.optimizer  3.Alpha | 1.SVM  2.LSTM  3.LASSO Regression | 1.Accurate Yield Prediction  2.Data-Driven Insights  3.Adaptive Farming Practices  4.Reduced Resource Waste | 1.Data Quality and Quantity  2.Complex Model Interpretation  3.Model Overfitting  4.Need for Expertise |
| Guiping hu et al.  (2019)[19] | 1.Ginni coefficient loss  2.Normalization  3.coefficient shrinkage | 1.random Forest  2.LASSO regression  3.ridge regression | 1.Precision Agriculture  2.Resource Management  3.Support for Policy Decisions | 1.Data Privacy and Security  2.Model Maintenance  3.Overfitting  4.Complexity and Interpretability |
| Hemantkumarawan et al.(2017)[20] | 1.kernel  2.precision  3.splitter  4.max\_depth  5.bootstrap  6.alpha | 1.SVM  2.Logistic Regression  3.KNN  4.Decision tree  5.Random forest  6.Neural Networks | 1.Early Detection  2.Increased Yield and Quality  3.Resource Efficiency  4.Customization for Crops | 1.Model Complexity  2.False Positives and Negatives  3.Maintenance and Updates  4.Interpretability |

**Table-1 Literature Review table**

**SUMMARY:**

**T**hese papers introduce a range of machine learning approaches,including regression models,decision trees,random forests,support vector machines,and neural networks.Each algorithm's suitability is assessed based on historical yield data,climate variables,soil information,and related inputs.Comparative evaluations highlight the algorithms' respective capabilities in handling diverse datasets and generating precise yield predictions Suruliandi.A et al.(12 Mar 2021)[10].

**T**hese papers commence by emphasizing the pivotal role of crop selection in agricultural success,highlighting the need for a data-driven approach to navigate the intricacies of this decision Usha Devi, Sheela Selvakumari(3 May 2022)[15].The ICRS is introduced as a comprehensive solution,aimed at delivering personalized crop suggestions based on a comprehensive analysis of multifaceted factors A Priyadharshini et al.(2021)[12].

**D**ata acquisition involves compiling comprehensive datasets that include historical disease records,environmental factors,crop characteristics,and geographical information.Pre processing techniques are applied to ensure data quality by addressing missing values,normalizing variables,and handling outliers.Data collection involves compiling extensive datasets encompassing historical yield records, climate data, soil characteristics, and crop attributes G Fenu,FM Malloci(2020,January 21)[3]. Pre processing techniques are applied to ensure data quality through handling missing values, normalizing variables, and mitigating outliers, thus laying a robust foundation for subsequent analysis.the research paper demonstrates the potential of machine learning techniques to transform crop disease forecasting A.Sharma,A.jain,P.Guptha,V.Chowdhary(2021)[4] .By combining historical data, environmental factors, and advanced algorithms, these techniques offer a proactive approach to disease management Mahmudul Hasan et al.(10 August 2023)[16] .The study highlights the importance of ongoing research to refine algorithms, incorporate new data sources, and adapt to evolving disease patterns, ultimately contributing to more resilient and productive agricultural systems A.Nigam et al.(2020)[6].To validate algorithm effectiveness, the study employs performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R2). Additionally, model interpretability and explainability are emphasized to build trust among stakeholders in the agricultural sector MK.Dharani et al.(2021)[5].

In a systematic exploration, the paper introduces various machine learning algorithms, including regression models, decision trees, support vector machines, and neural networks, tailored to the unique characteristics of Indian agriculture. The assessment of these algorithms underscores their capability to process intricate datasets and provide dependable yield forecasts. Comparative analyses shed light on their individual merits and limitations concerning the diverse variables intrinsic to the Indian agricultural context.Hemantkumarawan,Nilima ashtanka(7 January 2017 )[20].

The research paper offers an illuminating examination of the transformative potential of employing deep learning techniques for crop prediction. The study underscores the pivotal role of accurate crop forecasting in optimizing agricultural practices and resource allocation, positioning deep learning as a promising solution in this context Asif Ali et al.(31 July 2013)[15].

This research paper offers a comprehensive perspective on the profound implications of integrating machine learning techniques within the realm of precision agriculture. By seamlessly merging advanced algorithms with the wealth of agricultural data available, machine learning has the potential to shape more sustainable, productive, and resource-efficient farming practices. This study contributes significantly to the evolving landscape of precision agriculture, advocating for the continued exploration and refinement of machine learning applications to address the complex challenges faced by the agricultural sector Akshay Dhande,Rahul malik(10 August 2022)[10].

The core of the paper revolves around an exploration of various machine learning algorithms that are exceptionally suited for precision agriculture. The research delves into regression models, neural networks, clustering algorithms, and decision trees, among others. Each algorithm is meticulously evaluated based on its aptitude to comprehend intricate datasets and provide actionable insights that can potentially reshape agricultural practices. This section elucidates the algorithms' individual strengths, ultimately showcasing how they can collectively optimize crop production and sustainability G.Belanger et al.(January 2006)[16].

**PROBLEM STATEMENT:**

**Initial Investment,Uncertainty,Risk and Imbalanced data**

**G**enerally most of the farmers do not have any idea about crop choosing.They generally go with the flow that means they cultivate the crops according the trend going on.This may effect crop yield because they literally don’t know which crop is suitable to that soil,temperatures,weather conditions and season etc…

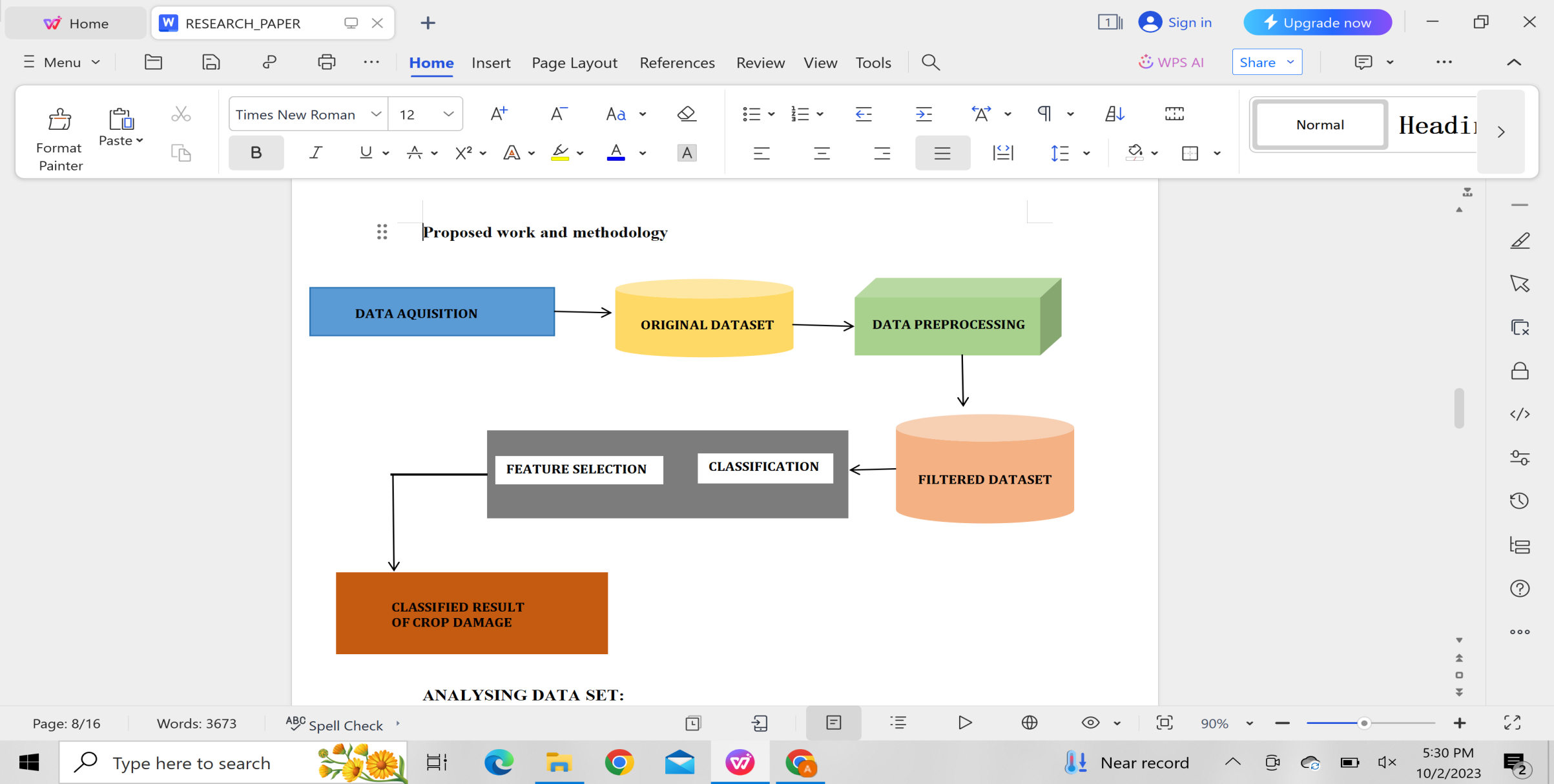
**T**he farmers now-a-days are habituated to use hybrid varieties which will give fast production.This will help the farmer for only one type of crop but as the time passes that will definitely effect the soil fertility.If the farmer want to rotate the crop the soil may not give better yield.So,to avoid such consequences the farmer should use some technologies to know about crop prediction.

**B**y implementing few algorithms of Machine Learning we can find the accuracy and it predicts whether the crop gets damaged or we get a good yield.This will become a boon to many farmers which help them to get benefit and earn benefits.

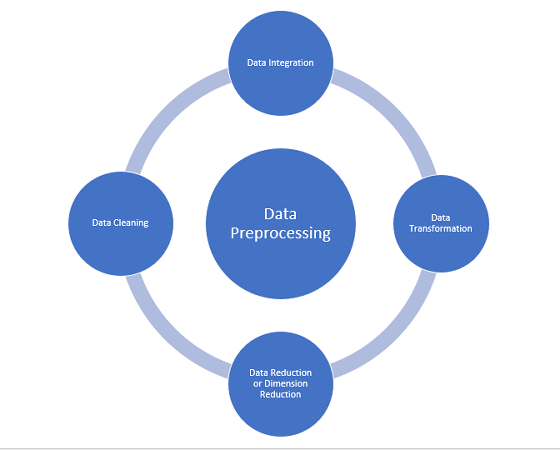
**T**he main motive is to prove the prediction accuracy using the different classification models and compare which model performs better regarding the problem.

**I**mbalanced data set gives the low accuracy hence,we can overcome this limitation by balancing the data set using some techniques.We are going to compare each and every accuracy.We are going establish the machine learning algorithm which gives the best accuracy.

**PROPOSED WORK AND METHODOLY:**



#### **Figure1 System Architecture**



#### **Figure2 Data Pre-Processing representation**

##### ANALYSING DATA SET:

#### This phase includes both data acquisition and original data set from the architecture.Before starting the project,it's crucial to perform an analysis of the data set.Our data set comprises 10 columns and 88858 rows,and it's worth noting that there are 9000 null values in Number\_Weeks\_Used.We replaced that missing values using mean fill method.

#### **DATA COLLECTION AND PRE PROCESSING:**

This phase includes data processing stage from the architecture.We acquised our data set from Kaggle. Initially, we must check for any duplicate values and, if found, remove them using .drop\_duplicates().

In this data set, null values are represented by '..', but the system interprets them as strings rather than null values. Our primary task is to replace all instances of '..' with pandas.NA.

For NaN values, we can opt to replace them with the mean, mode, or median. Alternatively, if a row contains a significant number of NaN values, it may be prudent to remove that row.

There are a total of 534 rows with missing values, amounting to a total of 920 missing values in the data set.

Since we have separate training and testing datasets, all pre processing techniques should be applied to both.

**DATA SPLITTING:**

This phase includes filtered data set stage from the architecture.We split the data into training and testing datasets.Evaluation can be directly applied to the testing dataset.

**MODELS USED:**

This phase includes classification stage from the architecture.As the dataset consist of output and the output can be 0,1 and 2 so different classification techniques were applied such as Logistic Regression,Decision Tree ,Support Vector,Random Forest,K-Nearest Neighbor.Evaluation were done using these models on the data set.

**MODEL TRAINING:**

This phase includes feature selection stage from the architecture.Model training involves optimizing both variance and bias to attain optimal values.During this phase,the model comprehensively grasps all its features and captures the inherent structure within the provided data.The primary objective of model training is to derive a mathematical function based on the available data.This function is subsequently employed to process inputs and generate outputs,essentially making predictions based on this established function.

**MODEL EVALUATION:**

This phase includes results stage from the architecture.Model evaluation serves the purpose of assessing the predictive capability of the model and gauging its performance.This assessment is carried out using a testing data set.Various metrics such as Accuracy,Precision,Recall,F1 score,Confusion Matrix,and are used for this purpose.

**MODEL SELECTION:**

**T**he model is chosen by comparing various metrics,and the most optimal one is utilized for making predictions.

**PROCEDURE TO SOLVE THE GIVEN PROBLEM**

The crop damage prediction data set is imported into our project from Kaggle in comma separated values (csv) format. With the use of pandas, numpy, and scikit-learn, the data set is analyzed. Scene is used as a tool for information visualization. After using Scene to pull insights from the data set, we identify the key factors, or factors that have the most impact on how much expenses have changed. The elements with inconsequential values are omitted from the overall result. The data set is divided into two separate sets: the testing set and the preparation set. With the use of the preparation set, multiple machine learning models are created. The performance of each machine learning model is then evaluated using the testing set. An accuracy score is computed.

**LOGISTIC REGRESSION**

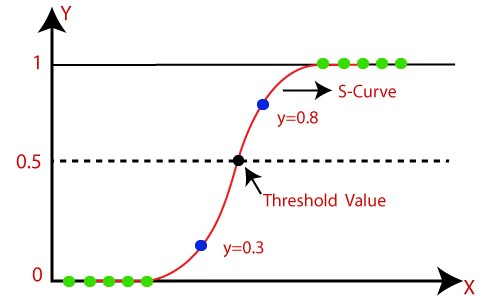
One of the most well-known machine learning algorithms, pertaining to supervised learning methods, is logistic regression. Using a predetermined set of free factors, it is used to predict a categorical secondary variable.

• A categorical dependent variable's result can be predicted by logistic regression. Therefore, the outcome must be graded or treated with discretion.

• Rather than providing the exact values of and 1, it provides the probabilistic values that fall between and 1. It can be either Yes or No, or 1, true or incorrect, etc.

• In logistic regression, we fit a computed task that looks like the letter "S," predicting the two most extreme values (0 or 1). This is done in place of fitting a regression line..

• We use the idea of a threshold value in logistic regression, which describes the probability of either or 1. For instance, estimates are higher at values beyond the limit and lower at values below the limit, when estimates tend to be 0. The sigmoid's behaviour is determined by the logistic regression represented by this condition.



**Figure2 Logistic Regression**

**K-NEAREST NEIGHBOURS:**

The K-NN working can be explained on the basis of the below algorithm:

**Step-1:** Select the number K of the neighbors

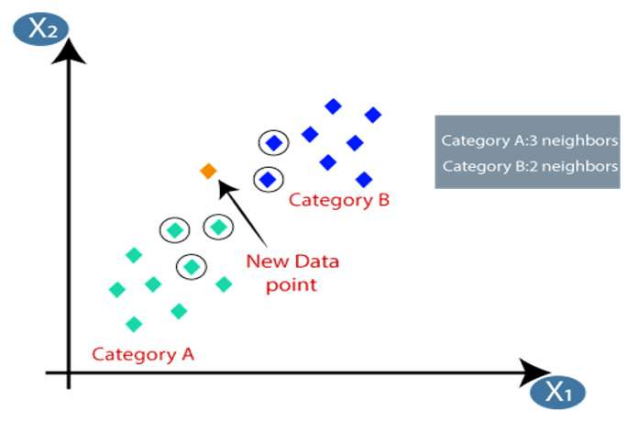
**Step-2:** Calculate the Euclidean distance of K number of neighbors

**Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.

**Step-4:** Among these k neighbors, count the number of the data points in each category.

**Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

**Step-6:** Our model is ready.



**Figure3 KNN**

Firstly, we will choose the number of neighbors ,so we will choose the k=5.Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between points, which we have already studied in geometry. By calculating the Euclidean distance, we get the nearest.Neighbors,as three nearest neighbors in category A and two Nearest neighbors in category B.

**SVM(SUPPORT VECTOR MACHINE):**

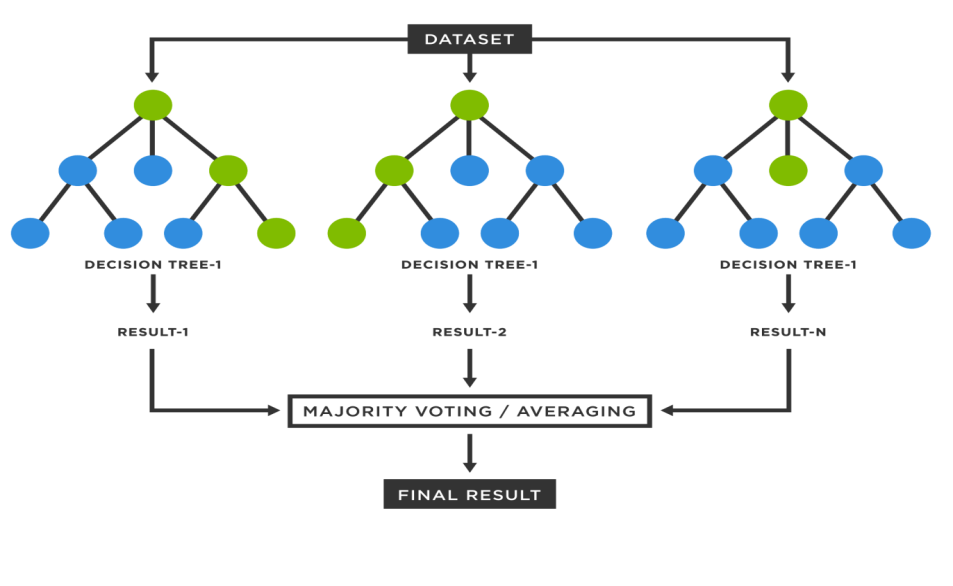
* Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. In the SVM method,we plot each data item as a point in n-dimensional space(where n is no.of features you have).
* We perform classification by finding the hyperplane that differentiates the two classes very well.
* We have three hyper planes(A,B and C). Now,identify the right hyper plane to classify star and circle.
* We want our data points to be as far away from the hyperplane as possible,while still being on the correct side of it.
* The distance between the hyperplane and the nearest data point from either set is known as margin.
* The goal is to choose a hyperplane with the greatest possible margin.
* There will never be any data point inside the margin.



**Figure4 SVM**

**RANDOM FOREST:**

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.Random forest algorithms have three main hyper parameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we’ll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged.



**Figure5 Random Forest**

**EXPERIMENTAL WORK:**

The data set crop damage prediction is taken from the Kaggle[21]. This dataset is used to

predict the crop damage and also used to know which factors are affecting most for the

damage caused to the crop.

This data set contains 88858 records,which were then classified into 3 classes:

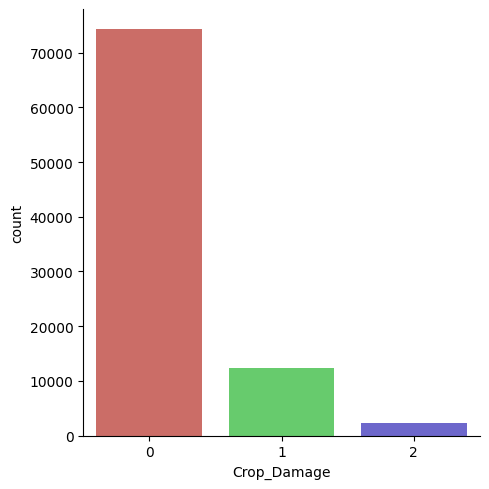
No Damage - 0

Pathological Damage - 1

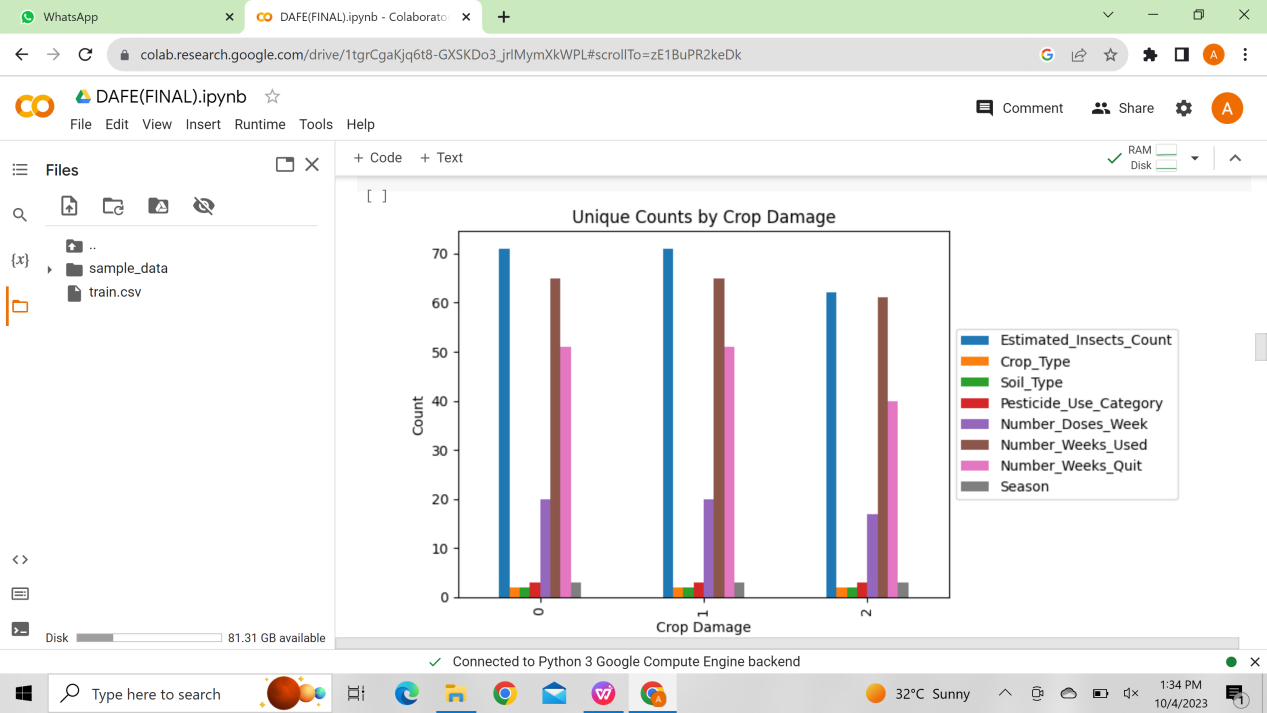
Damage- 2

The data set has a column with string values hence,we removed that column i.e,ID.The data set contain some missing values in the column so,we used mean of that particular column and filled the missing values.

The data set contains 88858 rows and 10 columns including target column i.e,Crop\_Damage.There are 9000 missing values in Number\_Weeks\_Used.We replaced that missing values using mean fill method.



**Figure6 Distribution of label column values(imbalanced data)**



**Figure7 Distribution of label column with all other columns**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Attribute name** | **Count** | **Mean** | **Std** | **Min** | **Max** | **Unique count** |
| Estimated\_Insects\_Count | 88858 | 1399.0122104 | 849.044003 | 150 | 4097 | 71 |
| Crop\_Type | 88858 | 0.284375 | 0.451116 | 0 | 1 | 2 |
| Soil\_Type | 88858 | 0.4584167 | 0.498268 | 0 | 1 | 2 |
| Pesticide\_Use\_Category | 88858 | 2.264185 | 0.461769 | 1 | 3 | 3 |
| Number\_Doses\_Week | 88858 | 25.849951 | 15.554340 | 0 | 95 | 20 |
| Number\_Weeks\_Used | 79858 | 28.623970 | 12.391804 | 0 | 67.0 | 64 |
| Number\_Weeks\_Quit | 88858 | 9.589986 | 9.900575 | 0 | 50 | 51 |
| Season | 88858 | 1.896959 | 0.701318 | 1 | 3 | 3 |
| Crop\_Damage | 88858 | 0.190562 | 0.454213 | 0 | 2 | 3 |

**Table-2 Data set description**

The data set on which we had been working is a classification related data set and following are the performance parameters.

In this framework we take information as content input from the client and after that we pre-process information of the client

Following we extricate the specified data from the information and after that it is sent for classification.

In classification information is classified utilizing prepare information set available in the framework and utilizing different calculation cost is anticipated.

**Classification Performance Metrics:**

Accuracy: Measures the proportion of correctly classified instances.

Formula: (TP + TN) / (TP + TN + FP + FN)

Precision: Measures the accuracy of positive predictions.

Formula: TP / (TP + FP)

Recall (Sensitivity or True Positive Rate): Measures the proportion of actual positives correctly predicted.

Formula: TP / (TP + FN)

F1-Score: Combines precision and recall into a single metric.

Formula: 2 \* (Precision \* Recall) / (Precision + Recall)

**DATA SET DESCRIPTION:**

**Data set Glossary (Column-Wise):**

* **ID:** Represents the ID of the crop
* **Estimated\_Insects\_Count:** Represents the insect count in a crop
* **Crop\_Type:** Represents the type of the crop
* **Soil\_Type:** Represents the soil type
* **Pesticide\_Use\_Category:** Represents the pesticide type
* **Number\_Doses\_Week:** Represents the no.of doses used per week
* **Number\_Weeks\_Used:** Represents the no.of weeks used
* **Number\_Weeks\_Quit:**Represents the no.of weeks quit
* **Season:** Represents the season
* **Crop\_Damage:**Represents the crop damage

**RESULTS**

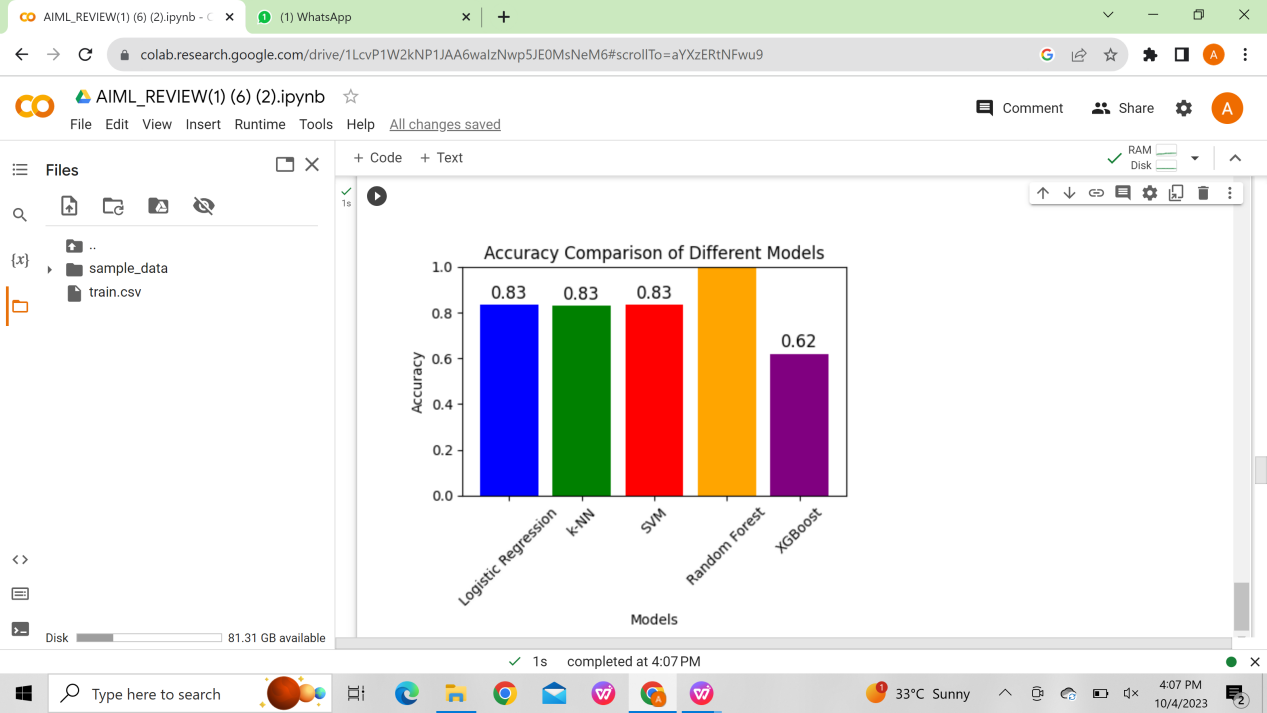
This table reflects the accuracy values which are obtained when different models are applied

|  |  |  |
| --- | --- | --- |
| **MACHINE LEARNING ALGORITHMS** | **ACCURACY BEFORE BALANCING THE DATA SET** | **ACCURACY AFTER BALANCING THE DATA SET** |
| LOGISTIC REGRESSION | 83% | 51% |
| K-NEAREST NEIGHBORS | 83% | 85% |
| SUPPORT  VECTOR MACHINE | 83% | 83% |
| RANDOM FOREST | 82% | 95% |
| XGBOOST | 62% | 92% |

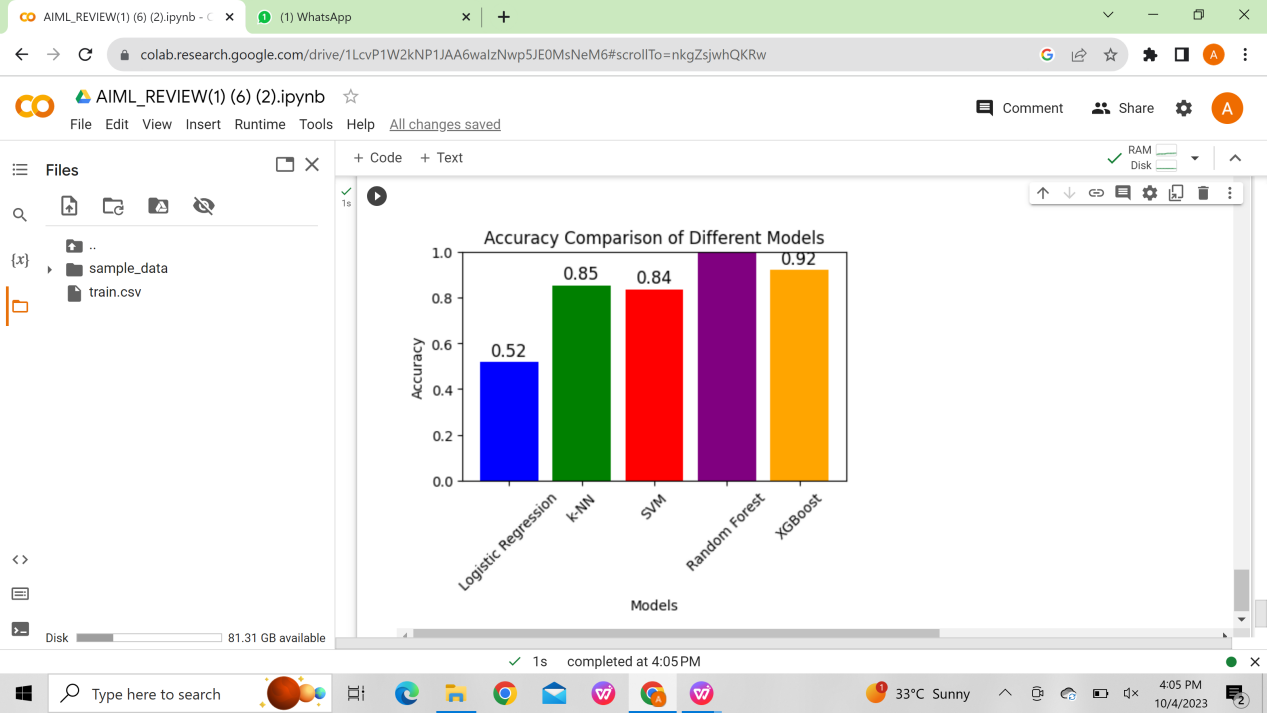
**Table-3 Accuracies of different models**

After balancing the data set we got better accuracy than before.Highest accuracy is acquired in Random Forest.

The following graphs shows the accuracy of 5 models before and after balancing the data



**Figure8 BEFORE BALANCING THE DATA**

****

**Figure9 AFTER BALANCING THE DATA**

Both graphs are representing the accuracies of different models when the provided data set is imbalanced and when it is balanced.The graphs suggests that the accuracy is more in the balanced data set and it is obtained when random forest model is applied.

The proposed system recommends whether the crop gets damaged or not by considering parameters as Estimated\_Insects\_Count, Crop\_Type, Soil\_Type,Pesticide\_Use\_Category, Number\_Doses\_Week, Number\_Weeks\_Used, Number\_Weeks\_Quit, and Season.In the output section the system displays whether the crop gets damaged or not along with its accuracy.The proposed system also balances the imbalanced data set for more accuracy.

All the code related to above results can be found at:https://github.com/archanakagithapu/DAFE\_PROJECT.git

By this we can conclude that it is better to balance the data to get most accurate results.This model is definitely going to help the farmers in a huge profitable way.

**CONCLUSION**

**T**his study has explored the impact of features like season , crop type and soil type etc on crop damage.This work advocates for early information sharing specifically on expected yield so one can ensure a proper planning before growing the crop . The entire aim of this project is to predict crop damage before getting yield . This can be done using some machine learning techniques like KNN , Logistic Regression , Support Vector Machine, Random Forest.This research will help the farmer to know about the crop production rating before the farmer gets crop yield. Depending on the parameters which have been provided by the farmer we can predict the output in the form of accuracy using some classification algorithms which will give the farmer, whether the farmer will get good yield or damaged crop.

**FUTURE SCOPE**

**I**n future, new features from the fields can be gathered to get a perfect image of the crop damage using other machine learning algorithms and deep learning algorithms such as ANN or CNN to get more accurate predictions.

**I**n coming years, this approach can be developed for any type of users like low level farmers who are uneducated , by adding some additional features into the machine learning model which can be learned by any type of user in hardly 1 hour . This research work can be enhanced to higher level by availing it to whole India for country’s development.

**REFERENCE SECTION**

[1] Yelamarthi.K,FF Haque,A Abdelgawad,VP Yanambaka(2020).Crop Yield Analysis Using Machine Learning Algorithms.IEEE

[2]A Priyadharshini,S Chakraborty(2021).Intelligent Crop Recommendation System using Machine Learning.IEEE

1. G Fenu,FM Malloci(2020,January 21).An Application of Machine-Learning Technique-in Forecasting-Crop Disease.Association for Computing Machinery New York,NY,United States

[4]A.Sharma,A.jain,P.Guptha,V.Chowdhary(31 December 2020)Machine Learning Applications for Precision Agriculture.IEEE

[5].MK.Dharani,R.Thamilselvan,P.Natesan,S.Santhoshkumar,PCD.Kalaaivaani(2021).Review on Crop Prediction Using Deep Learning Techniques.IOP

[6]A.Nigam,S.Garg,A.Agarwal(2020).Crop Yield Prediction Using Machine Learning Algorithms.IEEE

[7]R.Medar,VS.Rajpurohit(2019).CropYieldPredictionusingMachineLearningTechniques.IEEE

[8]YJN.Kumar,V.Spandana,VS.Vaishanavi,K.Neha,V.Devi(2020).Supervised Machine Learning Approach for Crop Yield Prediction in Agriculture Sector.IEEE

[9]D.Thomas(2020).Crop Yield Prediction Based On Indian Agriculture Using Machine Learning.IEEE

[10]Akshay Dhande,Rahul malik(10 August 2022).Design of a highly efficient crop damage detection ensemble learning model using deep convolutional networks.SPRINGER

[11]Nischitha K (08 August 2020).Crop prediction using machine learning approaches.ijert

[12]Tiago Domingues,Tomas Brandao,Joac Ferreira(5 September 2023).Machine learning for detection and prediction of crop diseases and pests:A comphrehensive survey.MDPI

[13]Donggeun Han,Donghee Yoo,Tae youngkim(5 July 2023).Analysis of social welfare impact of crop pest and disease damages due to climate change:a case study of dried red peppers.Springer

[14]Thuanha,Yanbenshen,Hemabuddu,Eric Johnson,Steven J(16 February 2022). Quantifying hail damage in crops using sentinel-2 imagery.MDPI

[15] Asif Ali ,Jens C.streibig,Christian,Andreasen(31 July 2013).Yield loss prediction models based on early estimation of weed pressure.ELSEVIER

[16]G.Belanger,Y.Castonguay,A.Bertrand,C.Dhont(January 2006).Winter damage to perennial for age crops in eastern canada.Canadian Journal of Plant Science

1. Sathwik Gara,Pradyumn Pundir,Himanshu Jindal,Somya Garg( 2021). Towards a multimodal system for precision agriculture reusing IOT &ML.IEEE
2. Johnathnshook,TryambakGangopadhyay,Soumiksarkar,Ganapathysubramanian(2021)Crop yield integrating genotypes & weather variables using deep learning.plos.org
3. Guiping hu,Mohsen shahhosscin,Martinez Feria(4 December 2019).Maize yield & nitrate loss prediction with ML algorithms.IOP Publishing Ltd
4. Hemantkumarawan,Nilima ashtanka(7 January 2017 )An appropriate model predicting pest/disease of crops using ML algorithms.IEEE

[21] Data set link:https://www.kaggle.com/code/gauravduttakiit/crop-damage-prediction-flaml-accuracy