ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, 2023

PREDICTION *OF CROP HARVESTS BASED ON WEATHER DATA USING ML*

[1] *S.Sravan Reddy,* [2] *Sai Phanindra Banala,* [3] *Bodhu Nithin ,* [4] *Nallapu Prashanthi*

[1] *Assistant Professor, Department of Information Technology, Teegala Krishna Reddy Engineering College Hyderabad, Telangana, India.*

[1] *sravanreddysathu@gmail.com*

[2,3,4] *UG Scholars Department of Information Technology, Teegala Krishna Reddy Engineering College , Hyderabad, Telangana, India.*

[2] *bspsaiphani@gmail.com* , [3] *bodhunithin5@gmail.com* , [4]*prashanthinallapu2002@gmail.com*

Abstract

Although agriculture remains the dominant economic activity in many countries around the world, in recent years this sector has continued to be negatively impacted by climate change leading to food insecurities. This is so because extreme weather conditions induced by climate change are detrimental to most crops and affect the expected quantity of agricultural production. Although there is no way to fully mitigate these natural phenomena, it could be much better if there is information known earlier about the future so that farmers can plan accordingly. Early information sharing about expected crop production may support food insecurity risk reduction. The study applies machine learning techniques to predict crop harvests based on weather data and communicate the information about production trends. The collected data were analyzed through Random Forest, Polynomial Regression, and Support Vector Regressor. Rainfall and temperature were used as predictors. The models were trained and tested.

1.INTRODUCTION

Agriculture is an economic activity that has a high dependency on weather conditions . This means that seasonal agriculture is dependent on natural weather conditions, also known as rainfed agriculture. Rainfed agriculture constitutes 80% of the cropland worldwide and generates good yields when crops have favorable weather conditions. In many lands where rainfall is scarce, rainfed

agriculture is supplemented by irrigation practices . The fact still remains that agricultural production is heavily reliant on rainfall and other weather variables. It is such the case that at times, farmers do not acquire the expected harvest due to the scarcity or abundance of rainfall and other weather parameters.

Climate change has a great impact on the productivity of agriculture and may lead to hunger or food insecurity. The latter is a crucial problem in the regions characterized by droughts or other weather-related disasters. Climate variables that affect crop production include precipitation, air temperature, humidity, and solar radiation . Different studies have shown that climate indices at both global and regional levels affect crop yields and food security . In their study, Damien et al. found that the reduced crop yields could be associated with either high temperature or abundant precipitation . Extreme temperature has negative effects on crop production due to various factors such as increased evapotranspiration and respiration of crops, and higher pest infestation . Increased precipitation intensity leads to increased runoff patterns that in turn cause floods and the risk of crop failure . Crop productivity can also be affected by the increased temperature that causes the increase in crop water demand . In all scenarios, climate change has a potential impact on agriculture in different ways.

Although the climate variables may be the same for a specific area, however, the needs of weather parameters are different from one crop to another according to their growing stage. This means that each crop has a different level of resilience to the atmospheric variables. When weather variables spike at an extreme level, a remarkable influence on crop production will be observed . The influence of climate change on agriculture can be observed everywhere. For example, from March to August 2018, a large portion of Europe experienced extreme temperatures, while the southern region of the continent experienced abundant rainfall.

Machine Learning based on prior crop prediction, soil quality analysis to achieve high crop yield throughout technology solution. The main objectives of this project are to predict crop-yield which can be extremely useful to farmers in planning for harvest and sale of grain harvest. Implement a machine learning algorithm that gives better prediction of suitable crop for the corresponding region and crop season in our country. This project

aims to predict yields based on location and weather data. The aim of this study is to look at the prediction of crops which will offer high yield within the given location considering the climate and soil parameters.

The primary objectives of this project are as follows:

Develop machine learning models for crop yield prediction.Analyze the performance of different machine learning algorithms in predicting crop yield. Identify the most influential factors affecting crop yield through feature selection and analysis. Provide insights and recommendations to farmers, policymakers, and agricultural stakeholders for improved decision-making.

2.LITERATURE SURVEY

Aruvansh Nigam, Saksham Garg, Archit Agrawal, conducted experiments on Indian government dataset and its been established that Random Forest machine learning algorithm gives the best yield prediction accuracy.

Leo Brieman, is specializing in the accuracy and strength & correlation of random forest algorithm. Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better the answer for the system..

Balamurugan, have implemented crop yield prediction by using only the random forest classifier. Various features like rainfall, temperature and season were taken into account to predict the crop yield.

Mishra, has theoretically described various machine learning techniques that can be applied in various forecasting areas. However, their work fails to implement any algorithms and thus cannot provide a clear insight into the practicality of the proposed work.

3. PROBLEM STATEMENT

Most of the existing system are hardware based which makes them expensive and difficult to maintain also they lack to give accurate results. Some systems suggest crop sequence counting on yield rate and market value.

4. PROPOSED SYSTEM

The system proposed tries to overcome these drawbacks and predicts crops by analysing structured data. Its a totally software solution, it does not allow maintenance factor to be considered much. the accuracy

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, 2023

would be high when compared to hardware based solutions, because components like soil composition, soil type, Ph value, and all inherit picture during the prediction process.

5. METHODOLOGY

5.1.Linear Regression:

Linear regression is a statistical modelling technique used to establish a relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the dependent variable and independent variables, which means that any change in the independent variable(s) will result in a proportional change in the dependent variable. In linear regression, the dependent variable is denoted by 'y', while independent variables are denoted by 'x'. The model attempts to find the best-fit line or curve that represents the relationship between the dependent variable and independent variable(s). The best-fit line or curve is determined by minimizing the sum of squared errors between the predicted and actual values of the dependent variable.

There are two types of linear regression models: simple linear regression and multiple linear regression. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are two or more independent variables.

The simple linear regression model can be represented mathematically as:

$$y = a + bx$$

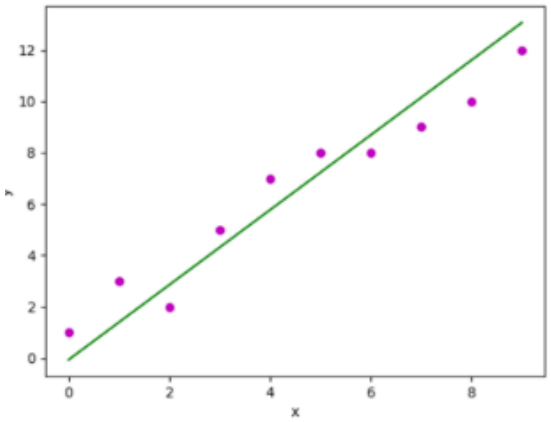
Where 'a' is the intercept of the line, 'b' is the slope of the line, 'x' is the independent variable, and 'y' is the dependent variable. The multiple linear regression model can be represented mathematically as:

$$y = a + b1x1 + b2x2 + ... + bnxn$$

Where 'a' is the intercept of the line, 'b1', 'b2', ..., 'bn' are the slopes of the line for independent variables x1, x2, ..., xn, respectively, and 'y' is the dependent variable. Linear regression is a commonly used modelling technique in various fields such as economics, finance, engineering, and science. It can be used to predict future values of the dependent variable based on the values of the independent variable(s).

However, it has limitations and assumptions, such as the linearity of the relationship between the dependent variable and independent variable(s), homoscedasticity of errors, and independence of errors.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, 2023



5.2.Multi linear Regression:

Multiple linear regression is a statistical modelling technique used to establish a relationship between a dependent variable and two or more independent variables. It is an extension of simple linear regression, which involves only one independent variable.

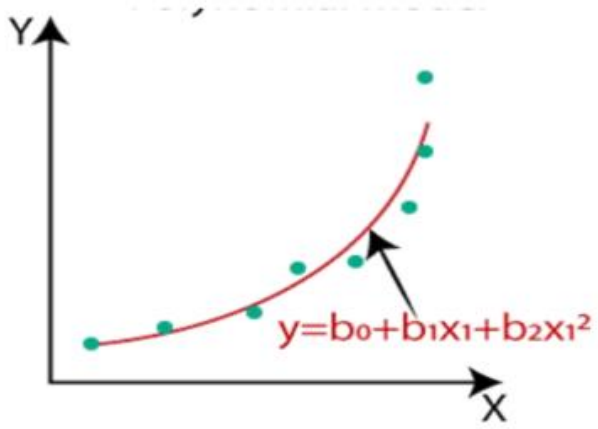
In multiple linear regression, the relationship between the dependent variable and independent variables is assumed to be linear. The model estimates the coefficients of the independent variables that best predict the dependent variable. The coefficients represent the slope of the line for each independent variable, indicating the amount of change in the dependent variable for a unit change in the corresponding independent variable, while holding all other independent variables constant. The multiple linear regression model can be represented mathematically as:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon$$

Where 'y' is the dependent variable, x1, x2, ..., xn are the independent variables, β0 is the intercept, and β1, β2, ..., βn are the coefficients of the independent variables, and ε is the error term.

The error term represents the unexplained variation in the dependent variable and is assumed to follow a normal distribution with a mean of zero and constant variance. The model assumes that the error terms are independent of each other and independent of the independent variables. Multiple linear regression can be used to predict the values of the dependent variable based on the values of the independent variables. It is commonly used in various fields such as economics, finance, social sciences, and engineering. It has several assumptions, including linearity, homoscedasticity, normality, independence, and absence of multicollinearity among the independent variables. Violations of these assumptions can lead to biased and unreliable results.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, 2023



5.3.Support Vector Regression:

Support Vector Regression (SVR) is a machine learning algorithm used for regression analysis. It is based on the Support Vector Machine (SVM) algorithm and is designed to handle continuous variables. The aim of SVR is to find the best fitting line that can explain the relationship between the dependent variable and the independent variables.

The main idea behind SVR is to identify the hyperplane that maximizes the margin between the data points and the hyperplane. The data points closest to the hyperplane are called support vectors, and they are used to define the margin. The hyperplane can be linear or non-linear, and it can be found by solving a constrained optimization problem. In SVR, the data is transformed into a high-dimensional feature space using a kernel function. This kernel function maps the original data to a higher dimensional space where a linear relationship between the dependent and independent variables can be found. Once the hyperplane is found in the higher-dimensional space, the solution is projected back into the original space.

The SVR algorithm can handle non-linear relationships between the dependent and independent variables using non-linear kernel functions such as the radial basis function (RBF) or the polynomial kernel. These kernel functions map the data to a higher-dimensional space where a linear relationship between the variables can be found. To train an SVR model, we need to define a loss function that measures the difference between the predicted values and the actual values. The loss function is minimized using an optimization algorithm such as gradient descent. The optimal parameters of the model are then found by minimizing the loss function.

One of the advantages of SVR is that it can handle a large number of features and can generalize well to new data. However, the choice of the kernel function and the hyperparameters can have a significant impact on the performance of the model. Therefore, careful selection of these parameters is important in order to obtain an

accurate and reliable model. Overall, SVR is a powerful and flexible regression algorithm that can be used in a wide range of applications, including finance, engineering, and biology.

5.4.Rondom Forest Classifier:

Random Forest Classifier is a popular machine learning algorithm that can be used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a powerful model that is highly accurate and robust.

The basic idea behind Random Forest Classifier is to create a large number of decision trees, each trained on a subset of the data and a subset of the features. This process is called bagging or bootstrap aggregating. The subsets are randomly selected, and each decision tree is trained independently of the others. During the training process, each decision tree is trained to predict the class of a data point based on a random subset of the features. This randomness helps to reduce overfitting and increases the diversity of the trees in the forest. The final prediction of the Random Forest Classifier is then made by combining the predictions of all the decision trees.

The algorithm can handle both categorical and continuous data, and it can handle missing values and outliers. Random Forest Classifier is known to be highly accurate and robust, and it can handle large datasets with high-dimensional feature spaces. One of the advantages of Random Forest Classifier is that it can provide feature importance rankings, which can help to identify the most important features in the dataset. This can be useful for feature selection and feature engineering. Another advantage of Random Forest Classifier is that it can be used for both classification and regression tasks. In regression, the algorithm uses the average of the predictions from all the decision trees in the forest to make the final prediction.

In summary, Random Forest Classifier is a powerful machine learning algorithm that can be used for a wide range of applications, including image recognition, natural language processing, and predictive analytics. Its ability to handle high-dimensional feature spaces and provide feature importance rankings make it a popular choice for many data scientists and machine learning practitioners.

5.5.XG-BOOST Classifier:

XG-Boost (Extreme Gradient Boosting) Classifier is a popular machine learning algorithm that is based on the gradient boosting technique. It is a fast and efficient algorithm that can be used for both classification and regression tasks. The basic idea behind XG-Boost Classifier is to combine the predictions of multiple weak models (usually decision trees) to create a more accurate and powerful model. The algorithm works by iteratively adding decision trees to the ensemble, where each tree corrects the errors made by the previous trees.

During the training process, the algorithm calculates the gradient of the loss function with respect to the model's predicted values. It then uses this gradient to update the weights of the trees in the ensemble, so that the next tree focuses on the areas where the previous trees made the largest errors. XG-Boost Classifier uses a variety of techniques to reduce overfitting and improve the performance of the model. These techniques include regularization, early stopping, and pruning. Regularization helps to prevent overfitting by adding a penalty term to the loss function. Early stopping stops the training process when the model's performance on the validation set starts to degrade. Pruning removes branches of the decision tree that do not improve the model's performance.

One of the advantages of XG-Boost Classifier is that it can handle missing data and imbalanced classes. The algorithm can also handle both categorical and numerical data, and it can automatically handle feature scaling. XG-Boost Classifier is known to be highly accurate and robust, and it is widely used in many real-world applications, including image recognition, text classification, and fraud detection. It is also known to be computationally efficient and can handle large datasets with high-dimensional feature spaces.

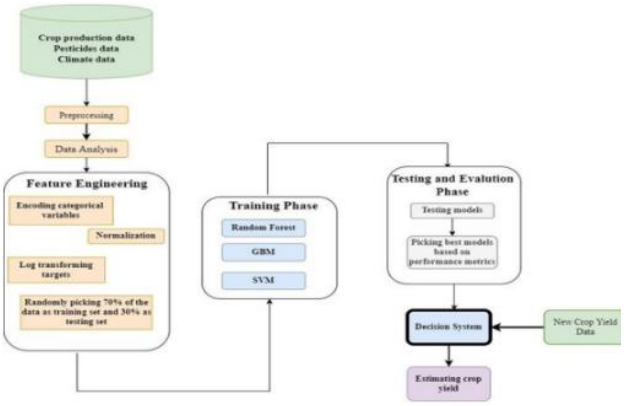
Overall, XG-Boost Classifier is a powerful and flexible machine learning algorithm that can be used in a wide range of applications. Its ability to handle missing data, imbalanced classes, and high-dimensional feature spaces make it a popular choice for many data scientists and machine learning practitioners.

6. DESIGN

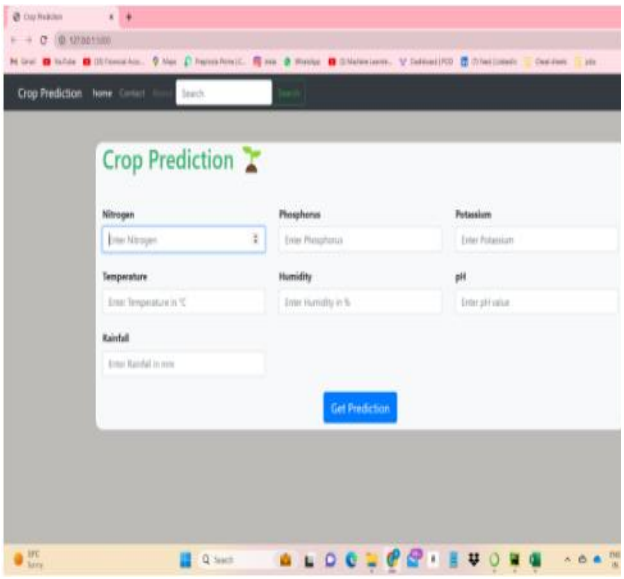
System design is transition from a user-oriented document to programmers or data base Personnel. The design is a solution, how to approach to the creation of a new system. This is composed of several steps. It

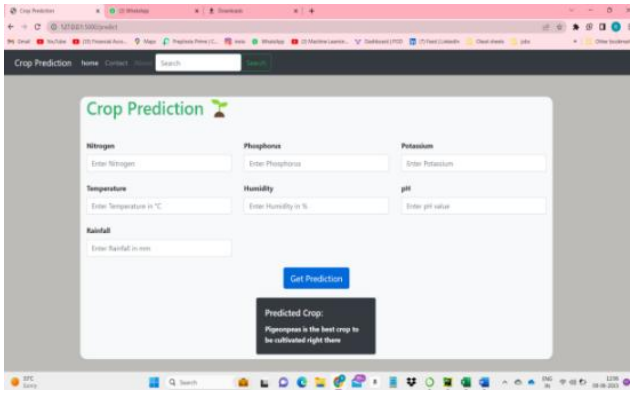
ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 1, 2023

provides the understanding and procedural details necessary for implementing the system recommended in the feasibility study. Designing goes through logical and physical stages of development, logical design reviews the present physical system, prepare input and output specification, details of implementation plan and prepare a logical design walkthrough.



7. RESULTS





8. CONCLUSION

In conclusion, machine learning projects require careful planning, preparation, and execution to be successful. A successful machine learning project should start with a clear problem statement and a well-defined set of objectives. It should also include a thorough data analysis phase, where the data is pre-processed, cleaned, and prepared for training and testing.

Choosing the right machine learning algorithm is also critical for the success of the project. Depending on the nature of the problem, different algorithms may be more suitable than others. It is important to carefully evaluate the strengths and weaknesses of each algorithm and choose the one that best fits the requirements of the project.

Once the algorithm has been selected, it is important to tune its hyperparameters to optimize its performance. This can be a

time-consuming process that requires careful experimentation and evaluation.

Finally, a successful machine learning project should include a thorough evaluation phase, where the performance of the model is tested on a holdout dataset. The evaluation should include various metrics, such as accuracy, precision, recall, F1- score, and ROC curve analysis. In summary, machine learning projects require careful planning, preparation, and execution to be successful. A successful project should start with a clear problem statement and a well-defined set of objectives, include a thorough data analysis phase, carefully select the right algorithm, tune its hyperparameters, and include a thorough evaluation phase. By following these steps, data scientists and machine learning practitioners can create powerful and accurate machine learning models that can solve complex problems and provide valuable insights.

9. FUTURE SCOPE

While the documentation on crop yield prediction using machine learning provides valuable insights and contributes to the current understanding of the subject, there are several potential avenues for future research and development in this field. Here are some future scopes to consider:

1. Integration of Advanced Remote Sensing Techniques: Incorporating advanced remote sensing techniques, such as hyperspectral imaging or unmanned aerial vehicles (UAVs), can provide high-resolution and real-time data on crop health, vegetation indices, and other relevant variables. Future research can explore the integration of these techniques with machine learning models to enhance the accuracy and timeliness of crop yield predictions.

2. Fusion of Multi-Source Data: Combining data from multiple sources, such assatellite imagery, climate models, soil sensors, and farm management systems, can lead to more comprehensive and robust crop yield prediction models. Future studies can focus on developing techniques to effectively integrate and utilize data from diverse sources to improve prediction accuracy.

3. Incorporating Climate Change Considerations: Climate change poses significant challenges to agricultural productivity. Future research can focus on developing machine learning models that explicitly account for the impact of climate change on crop yield. This may involve integrating climate projections, historical weather data, and other relevant factors into the models to provide more accurate

predictions under changing climatic conditions.

10 REFERENCES

[1] Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., ... & Boote, K. J. (2013). Uncertainty in simulating wheat yields under climate change. Nature Climate Change, 3(9), 827-832.

[2] Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. Environmental Research Letters, 2(1), 014002.

[3] Miao, Y., Mulla, D. J., Robert P. B., & David A. L. (2006). Applying neural networks to agricultural data for precision management. Agronomy Journal, 98(6), 1289-1297.

[4] Qu, M., Ma, L., & Sun, Z. (2017). A review of crop yield prediction models based on machine learning. In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-6). IEEE.

[5] Reynolds, M. P., Quilligan, E., Aggarwal, P. K., Bansal, K. C., Cavalieri, A. J., Chapman, S. C., ... & Rebetzke, G. J. (2016). An integrated approach to maintaining cereal productivity under climate change. Global Food Security, 8, 9-18.

[6] Srivastava, A. N., Yadav, R., & Lal, S. K. (2019). Crop yield prediction using machine learning techniques: A review. Computers and Electronics in Agriculture, 162, 627-648.

[7] Tan, C. W., Boominathan, P., & Lee, B. S. (2018). Machine learning techniques for crop yield prediction. In 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1021-1026). IEEE.

[8] Thorp, K. R., White, J. W., French, A. N., & Hoogenboom, G. (2017). Assimilation of remote sensing data into crop models for yield prediction. Agronomy Journal, 109(2), 423-434.

[9] Wang, J., Cao, Z., Yao, X., Zhao, Y., & Wang, L. (2019). A survey on neural network-based crop yield prediction using remote sensing data. Remote Sensing, 11(15), 1822.