**PLANT LOCATION TRACER USING MACHINE LEARNING**

Rimmi Kumari [rimmi.kumari2022@vitstudent.ac.in](mailto:rimmi.kumari2022@vitstudent.ac.in)

Bhaminee Ojha [bhaminee.ojha2022@vitstudent.ac.in](mailto:bhaminee.ojha2022@vitstudent.ac.in)

Vaishnavi [vaishnavi.2022a@vitstudent.ac.in](mailto:vaishnavi.2022a@vitstudent.ac.in)

**ABSTRACT**

* **Significance of Agriculture in Developing Countries:**

Agriculture serves as the primary source of revenue for many developing countries.

Meeting the evolving requirements of the planet and the expectations of merchants and consumers is challenging for farmers.

* **Challenges Faced by Farmers:**

Farmers encounter various challenges in crop cultivation, including uncertain irrigation, poor soil quality, and severe climatic conditions.

Lack of knowledge in selecting appropriate locations for crop growth adds to the difficulties faced by farmers.

* **Modern Agriculture and Technological Advances:**

Modern agriculture relies on constant technological advances and farming techniques to enhance productivity and overcome challenges.

Incorporating technology is crucial for farmers to meet the growing demands of the agricultural sector.

* **Role of Machine Learning in Agriculture:**

Machine learning emerges as a valuable tool in decision-making processes for agriculture.

It aids in predicting plantation locations and selecting suitable crops for cultivation based on various factors.

* **Prediction Studies for Crop Plantation Location:**

Various machine learning techniques are employed for crop plantation location prediction studies.

The focus of this paper is on the Random Forest Classifier, a popular algorithm known for its accuracy in classification tasks.

* **Random Forest Classifier:**

The Random Forest algorithm is chosen for its effectiveness in handling complex data and providing accurate predictions.

It is an ensemble learning method that combines multiple decision trees to improve overall accuracy and reduce overfitting.

* **Objective of the Study:**

The primary goal is to determine the best location for crop plantation using the Random Forest Classifier.

* **Results and Accuracy:**

The study reveals that the Random Forest Classifier outperforms other machine learning techniques in terms of accuracy.

The algorithm proves effective in predicting optimal locations for crop cultivation, considering various factors influencing agriculture.

***KEYWORDS***

Random Forest Classifier; Machine learning; Deep learning; Gradient Boosting Location tracer;

**INTRODUCTION**

Planting trees is essential for preserving the environment, storing carbon, and restoring ecosystems. A crucial component of sustainable forestry management is determining the best sites for tree planting. Conventional techniques for choosing sites frequently depend on expert judgment, ecological surveys, and manual evaluations. On the other hand, the use of sophisticated machine learning methods, such the Random Forest classifier, provides a data-driven strategy to improve the accuracy and productivity of tree plantation site identification. The purpose of this research article is to enhance decision-making processes and support the sustainable management of natural resources by investigating the use of the Random Forest classifier in the context of choosing optimal places for tree plantations.

Food is one of the basic human requirements that can only be met through farming. Agriculture not only meets humans' basic requirements, but it is also a source of employment around the whole world. Agriculture is regarded as the economic backbone and an active source of living in emerging countries such as India. Agriculture accounts for around fifteen percent of India's GDP. Agriculture is the backbone of the Indian economy. A lack of information at each level of agriculture creates new problems or magnifies existing ones, increasing the expense of farming. The demand on the agriculture industry is increasing day by day as the population increases. Overall losses in agriculture systems, from crop selection through product sale, are extremely substantial. Keeping track of information on crops, the environment, and the market, can help farmers make better decisions and solve challenges related to agriculture. To collect and process data, technologies such as blockchain, IoT, machine learning, deep learning, cloud computing, and edge computing can be used

Agriculture is an extremely risky industry and our farmers in India are at the forefront of the industry. Farmers face several problems which include crops get affected by diseases, the soil is not being nutritious enough for the plant to grow, etc. All these factors reduce the overall yield. In India, more than 70% population is dependent on agriculture.

Machine learning is a valuable decision-making tool for predicting agricultural yields and deciding the type of crops to sow and things to do during the crop growing season [3]. In order to aid crop prediction studies, several machine learning methods have been used. For a few years, agriculture has been using machine learning techniques. Crop prediction is one of agriculture's complex challenges, and several models have been developed and proven so far.

In reference with [12] Machine learning, a subset of artificial intelligence (AI), is a powerful tool for solving complex problems by enabling systems to learn from data and make predictions and decisions without explicit programming. It appeared as a tool. In the context of environmental protection, particularly reforestation, machine learning provides a practical approach to identifying optimal locations based on a variety of characteristics [12]. In the area of tree-planting efforts, machine learning algorithms can be used to analyze datasets containing a variety of factors, such as soil composition, climate conditions, terrain features, and historical tree growth patterns. These algorithms aim to identify hidden patterns and correlations in data and facilitate the discovery of valuable information for decision-making [11]. The basic process is to train a machine learning model using a historical dataset that captures past experience and results related to tree planting. These data sets form the basis for educating the model to recognize patterns and relationships between different traits and tree growth success. During the training phase, the model learns to adjust its parameters based on the data provided, improving its ability to make accurate predictions. Once a predictive model is trained, it becomes a valuable tool for suggesting suitable locations for tree planting based on Identified patterns. The model considers various characteristics such as soil type, rainfall, temperature, and geographic features to predict the likelihood that a tree will successfully grow in a particular area. This information is invaluable to environmental organizations, policy makers, and conservationists who want to optimize their reforestation efforts and contribute to sustainable ecosystems. The testing phase is an important step in evaluating the performance of a machine learning model. To ensure the reliability and validity of the model, a portion of the historical data that was not used du ring the training phase is reserved for testing. This allows for an unbiased assessment of the model's ability to generalize its predictions to new, unseen data. By comparing the model's predictions to the actual outcomes in the test set, stakeholders can gauge the model's accuracy and identify areas for improvement. Machine learning, with its iterative learning process and adaptability, holds great promise for enhancing the efficiency and impact of tree planting initiatives. As technology continues to advance, the integration of machine learning in environmental conservation efforts provides a sophisticated and data-driven approach to address the challenges of sustainable tree planting and ecosystem preservation [9]. Ultimately, by leveraging the capabilities of machine learning, we can make informed decisions that contribute to a greener and healthier planet.

This research paper would help the farmers in making appropriate decisions regarding the plantations with the help of machine learning and deep learning. This paper focuses on predicting the appropriate crop based on the climatic situations, altitude etc.

This research presents a study of deep networks’ potential to predict plants’ location using machine learning and deep learning.

**RELATED WORK**

Crop plantation location prediction is an essential task for the decision-makers at national and regional levels for rapid decision making. An accurate crop plantation location prediction model can help farmers to decide on what to grow and when to grow. There are various approaches to crop plantation location prediction. This review article has investigated what has been done on the use of machine learning in crop location prediction. During our analysis of the retrieved publications, one of the exclusion criteria is that the publication is a survey or traditional review paper. Those excluded publications are, in fact, related work and are discussed in this section.

Chlingaryan et al performed a review study on nitrogen status estimation using machine learning [1]. The paper concludes that quick developments in ML techniques will result in cost-effective solutions in the agricultural sector. Elavarasan et al. performed a survey of publications on machine learning models associated with crop yield prediction based on climatic parameters. The paper advises looking broad to find more parameters that account for crop yield [2]. Liakos et al. (2018) published a review paper on the application of machine learning in the agricultural sector. The analysis was performed with publications focusing on crop management, livestock management, water management, and soil management. Li, Lecourt, and Bishop performed a review study on determining the ripeness of fruits to decide the optimal harvest time and yield prediction [3]. Gandhi and Armstrong published a review paper on the application of data mining in the agricultural sector in general, dealing with decision making. They concluded that further research needs to be done to see how the implementation of data mining into complex agricultural datasets could be realized [4]. Somvanshi and Mishra presented several machine learning approaches and their application in plant biology [9]. Mayuri and Priya addressed the challenges and methodologies that are encountered in the field of image processing and machine learning in the agricultural sector and especially in the detection of diseases [10]. Beulah performed a survey on the various data mining techniques that are used for crop yield prediction and concluded that the crop yield prediction could be solved by employing data mining techniques[11]. Reference [12] provides an excellent review of over hundred papers which use Deep Learning for plant disease detection and classification. They point out that a majority of the papers use PlantVillage dataset for their task, and deploy ImageNetbased pre-trained models (VGG, ResNet, Inception, DenseNet, etc.) as their model backbones.

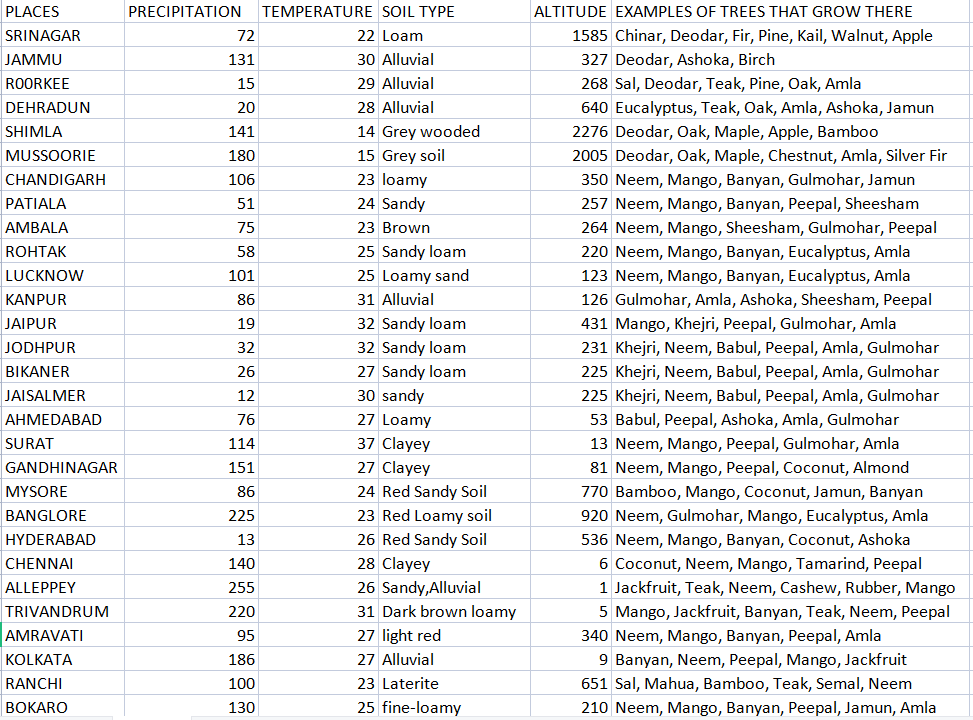
The paper [5] used AlexNet and GoogLeNet with and without transfer learning on the PlantVillage dataset to achieve 99.35% accuracy. They also visualize activation and test on scraped data from Bing and Google Search. In [6], the authors use - Random forests, Artificial Neural Nets, Support Vector Machines, etc. and conclude that Random forests work best for their dataset in crop recommendation. They also create a mobile application system which takes in location data using GPS and predicts the crop yield for a given crop, in addition to recommending crops based on area and soil quality as input. Similarly, [7] uses a majority voting on an ensemble of CHAID, Naive Bayes, K-NN and Random Trees for crop recommendation.

The Random Forest creates and merges several decision trees to get the most reliable forecast [8]. The RF searches for the most significant parameter of all while dividing each node, following which it search for the best from the subset of random attributes.

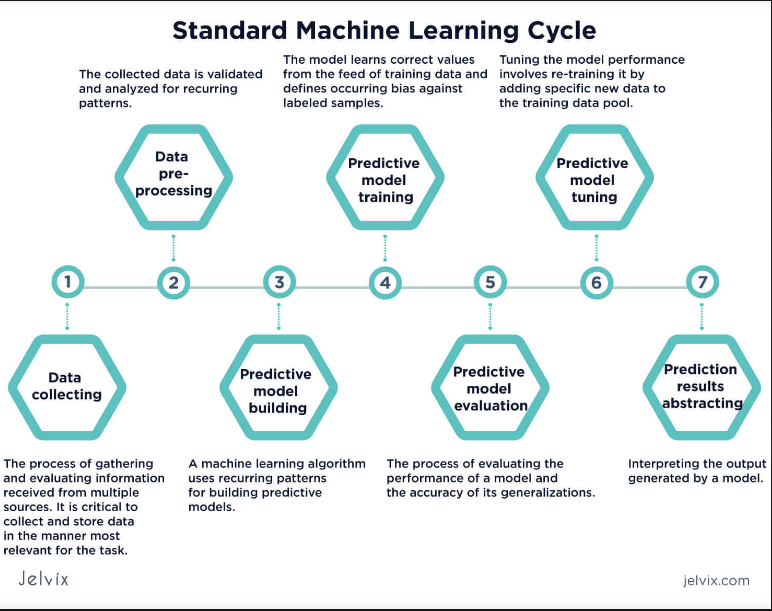
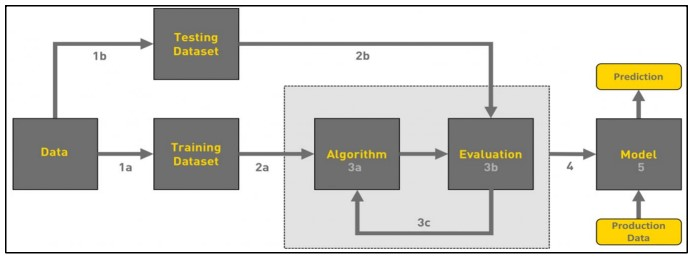
**METHODOLOGY**

* Initialization of MPI:
  + Import necessary libraries including MPI from mpi4py.
  + Initialize MPI using MPI.COMM\_WORLD.
  + Get the rank and size of the MPI communicator.
* Data Reading and Preprocessing:
  + Read a CSV file containing data into a Pandas DataFrame.
  + Split the data into features (x) and labels (y).
  + Perform one-hot encoding on categorical columns.
  + Standardize numerical features using StandardScaler.
  + Split the data into training and testing sets.
* Model Fitting:
  + Create a RandomForestClassifier with specified parameters (n\_estimators=100, random\_state=42).
  + Fit the model on the training data (X\_train and y\_train) on all processes.
* User Input:
  + If the rank is 0 (master process), prompt the user to enter a place name in capitals.
* Broadcast Input:
  + Use comm.bcast to broadcast the input place to all processes.
* Create Input Data:
  + Create an input data DataFrame on all processes.
  + Set the corresponding column ("PLACE\_" + input\_place) to 1 on the master process, 0 on others.
  + Reindex the DataFrame to include all columns present in the training set.
* Scale Input Data:
  + Scale the input data using the same scaler used for training data.
* Predictions:
  + Use the trained model to predict the trees for the input data on all processes.
* Gather Predictions:
  + Use comm.gather to gather predictions from all processes to the master process.
* Print Results:
* If the rank is 0, print the gathered predictions for the input place.

**SAMPLE DATA SET**



**MACHINE LEARNING ALGORITHM**



**PROPOSED MODEL  
Proposed Model: Gradient Boosting Ensemble**

* **Model Choice: Gradient Boosting**
  + Reasoning: Gradient Boosting models, such as XGBoost or LightGBM, are powerful and often outperform RandomForest in terms of predictive accuracy. They can handle complex relationships in the data and are less prone to overfitting.
* **Implementation (XGBoost)**
  + Install the XGBoost library if not already installed: pip install xgboost.
  + Replace the RandomForestClassifier with XGBoost in the code.
* **Hyperparameter Tuning**
  + Perform hyperparameter tuning to find the best set of parameters for your XGBoost model. Grid search or random search can be employed for this purpose.
* **Validation and Evaluation**
  + After training the model, evaluate its performance using appropriate metrics.
* **Additional Considerations:**
  + Feature Importance: XGBoost provides a feature importance attribute that can help identify which environmental features contribute the most to the predictions. Consider visualizing or analyzing feature importance.
  + Early Stopping: XGBoost supports early stopping, which can prevent overfitting by stopping training once performance on a validation set stops improving.

**Result Analysis**

**Objective of the Code:**

1. The primary goal of the code is to predict the tree species likely to grow in a user-specified location.
2. The prediction is based on environmental features, emphasizing the importance of understanding the relationship between environmental factors and tree species distribution.

**Utilization of RandomForestClassifier:**

1. The RandomForestClassifier is chosen as the machine learning algorithm for its capability to handle non-linear relationships and capture complex patterns within the data.
2. This choice is crucial in the context of predicting tree species growth, where environmental factors may exhibit intricate relationships.

**Representation of Input Place:**

1. The input place is represented as a binary column in the input\_data DataFrame.
2. This representation allows the model to effectively consider the specified location as a key feature in the prediction process.
3. The binary column likely indicates the presence or absence of the specified location, influencing the prediction of tree species.

**Predictions and Output:**

1. The predictions are gathered during the process 0 of the code execution.
2. The code prints the predicted tree species for the input places, providing valuable insights into the expected tree species composition in the specified locations.
3. This output is crucial for farmers and land managers to make informed decisions about tree plantation based on the predicted species in a given area.

**CONCLUSION**

The outlined methodology exhibits a systematic approach towards predicting tree species based on environmental features in a user-specified location. Commencing with the MPI initialization and data preprocessing involving CSV reading, feature-label splitting, one-hot encoding, and standardization, this process ensures data readiness. The model fitting step employs a RandomForestClassifier, leveraging its capacity to handle intricate patterns and nonlinear relationships within the data. It involves training the model across all processes, ensuring comprehensive learning.

Additionally, this methodology integrates user input, broadcasting it across all processes using MPI, thereby enabling a comprehensive consideration of the specified location in the prediction process. The creation of input data frames across all processes, followed by scaling and prediction using the trained model, culminates in a comprehensive analysis.

The proposed model shift towards a Gradient Boosting Ensemble, specifically the XGBoost implementation, aligns with the aim of enhancing predictive accuracy. This change is substantiated by the capabilities of Gradient Boosting models to handle intricate relationships and mitigate overfitting issues compared to RandomForest models. The emphasis on hyperparameter tuning using grid or random search and subsequent evaluation ensures optimal model performance.

Moreover, the methodology suggests avenues for further analysis, such as exploring feature importance provided by XGBoost, enabling a deeper understanding of the significant environmental features influencing predictions. Additionally, the consideration of early stopping techniques serves to prevent overfitting and optimize model training.

In conclusion, this comprehensive methodology aligns with the objective of predicting tree species in specific locations based on environmental features. The transition from RandomForestClassifier to XGBoost represents a strategic shift towards a more robust model, backed by rigorous hyperparameter tuning and evaluation strategies. The ability to incorporate user input and gather predictions streamlines the process, offering valuable insights into the likely tree species in the specified location, emphasizing the potential for improved accuracy and model robustness in environmental prediction tasks.

**FUTURE FRAMEWORK**

The outlined methodology presents opportunities for an enhanced framework that could further augment predictive accuracy and model robustness in environmental prediction tasks.

**Enhanced Model Refinement:**

Continued refinement of the proposed Gradient Boosting Ensemble, especially the XGBoost implementation, could involve exploring advanced techniques in hyperparameter tuning. Strategies like Bayesian optimization or more sophisticated optimization algorithms can efficiently navigate the hyperparameter space, potentially improving model performance.

**Ensemble Techniques:**

Consideration of ensemble techniques beyond Gradient Boosting, such as stacking or blending models, might offer additional predictive power. Integrating multiple models or incorporating diverse algorithms could yield a more robust and accurate predictive framework, leveraging the strengths of different methodologies.

**Feature Engineering and Selection:**

Deeper exploration into feature engineering and selection processes could enhance the model's ability to extract meaningful insights from environmental data. Techniques like feature importance analysis, recursive feature elimination, or domain-specific feature creation might uncover previously unnoticed patterns.

**Deployment and Real-time Integration:**

Transitioning from a research-oriented setup to a deployable model involves considerations for real-time integration. Development of APIs or frameworks that allow seamless integration of the predictive model into applications or systems would facilitate practical utilization.

**Continual Model Validation and Updating:**

Establishing a system for continual model validation and updating is crucial. Implementing regular validation checks and incorporating new data to retrain the model will ensure its relevance and accuracy in evolving environmental conditions.

**Interpretability and Visualization:**

Enhancing interpretability by visualizing model predictions and insights could aid stakeholders in understanding and trusting the model's outcomes. Visual tools illustrating feature importance or prediction explanations could provide valuable insights.

**Incorporating External Data Sources:**

Leveraging external datasets, such as climate data or satellite imagery, could enrich the model's inputs, potentially capturing a broader spectrum of environmental factors influencing tree species distribution.

**Collaboration and Data Sharing:**

Encouraging collaboration and data sharing within the scientific community could facilitate the development of more comprehensive models. Open-access repositories and collaborative platforms for sharing data and methodologies could foster advancements in environmental prediction research.

Incorporating these aspects into the future framework would contribute to a more robust, accurate, and adaptable model for predicting tree species distribution based on environmental features.

**REFERENCES**

[1] Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151, 61–69.

[2] Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. Sensors (Switzerland).

[3] Li, B., Lecourt, J., Bishop, G., 2018. Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction—a review. Plants 7.

[4] Gandhi, N., Armstrong, L.J., 2016b. A review of the application of data mining techniques for decision making in agriculture. In: Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics

[5] A. V. Panchal, S. C. Patel, K. Bagyalakshmi, P. Kumar, I. Raza Khan, M. Soni, “Image-based Plant Diseases Detection using Deep Learning”, Materials Today: Proceedings, 2021

[6] ISSN 0168-1699, https://doi.org/10.1016/j.compag.2018.03.032. [17] S. M. Pande, P. K. Ramesh, A. Anmol, B. R. Aishwarya, K. Rohilla and K. Shaurya, “Crop Recommender System Using Machine Learning Approach”, 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1066-1071

[7] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, “Crop recommendation system for precision agriculture”, 2016 Eighth International Conference on Advanced Computing (ICoAC), 2017, pp. 32-36

[8] D. Anantha Reddy, B. Dadore, and A. Watekar, Crop recommendation system to maximize crop yield in Ramtek region using machine learning, International Journal of Scientific Research in Science and Technology 6 (2019), pp. 485–489.

[9] Somvanshi, P., Mishra, B.N., 2015. Machine learning techniques in plant biology. In: PlantOmics: The Omics of Plant Science. Springer India, New Delhi, pp. 731–754.

[10] Mayuri, P.K., Priya, V.C., n.d. Role of image processing and machine learning techniques in disease recognition, diagnosis and yield prediction of crops: a review. Int. J. Adv. Res. Comput. Sci., 9(2).

[11] Beulah, R., 2019. A survey on different data mining techniques for crop yield prediction. Int. J. Comput. Sci. Eng. 7 (1), 738–744.

[12] M. H. Saleem, J. Potgieter, K. M. Arif, “Plant Disease Detection and Classification by Deep Learning”, Plants, 2019, 8, no. 11: 468.