

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

WINTER SEMESTER 2023-24

Programme Int M Tech (MIS)

Course Title Data Mining Techniques

Course Code SWE2009

Slot C2

**Title: Fertilizer Prediction**

**Team Members:**

P. Thanmai,22MIS1030

S. Harshitha Reddy,22MIS1139

Dogiparthi Harini,22MIS1174

**Faculty:** Dr Pattabiraman V.

**TABLE OF CONTENTS:**

|  |  |
| --- | --- |
| **S. No** | **Content** |
| 1 | Abstract |
| 2 | Scope |
| 3 | Objective |
| 4 | Introduction |
| 5 | Literature Review |
| 6 | Data Description |
| 7 | Architecture |
| 8 | Proposed Methodology |
| 9 | Novelty |
| 10 | Results and Discussion |
| 11 | Results Conclusion |
| 12 | References |

**1.Abstract:**

Our work focuses on predicting chemical options with minimal negative impacts on the environment. To address the increasing pressure to implement sustainable practices in agriculture to reduce environmental damage from intensive agriculture, we use a variety of data mining techniques such as random forest, K-nearest neighbour (KNN), logistic regression, decision tree, and support vector machine (SVM) were used) to analyse historical trends in coal consumption and environment. We evaluated the efficiency of each data mining approach to determine its predictive effectiveness for environmentally friendly coal options. The reliability of the predictions was assessed using metrics such as accuracy, precision, recall, and F1-score. Among the tested methods, Logistic Regression, Decision Tree, and Random Forest emerged as the most promising because of their high accuracy in predicting more environmentally friendly coal options. By combining these efficient methods, our stacked model aims to provide more accurate and reliable forecasts than individual algorithms alone. Our stacked model represents a significant advance in agricultural decision making. We harness the power of data mining and machine learning to provide reliable solutions to recommend coal balance and environmental sustainability. Our model has the capability to help farmers, agricultural experts to make informed decisions that support sustainable agricultural methods.

**Keywords:** Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Support Vector Machine (SVM), Stacked Model

**2.Scope:**

The target of the project "Prediction of Fertilizer with Least Harmful Effects" is to create a predictive model for fertilizer recommendations with the least amount of adverse effects possible by applying data mining techniques, including random forest, KNN, logistic regression, decision tree, and SVM. The major goal is to tackle the problem of choosing fertilizers optimally to reduce environmental effect while preserving agricultural output. In order to improve the accuracy of fertilizer recommendations, the project intends to produce a stacked model that combines logistic regression, decision tree, and random forest approaches. This model will take advantage of the methods' high accuracy. A comprehensive report describing the model development process, performance analysis, and recommendations for coal selection based on model predictions includes deliverables Data collection, preprocessing, feature selection, model training, validation, and testing all part of project's work. The project also includes developing a simple interface for the fertilizer suggestion system. The boundaries of the project include the creation of the predictive model and the corresponding software interface, but they do not include the actual application of the fertilizer recommendations in farming environments. Limitations on accessible data, computational resources, and the assumption that the input data adequately depicts the real-world scenario are examples of constraints and assumptions.

**3.Objective:**

The number one objective of the "Prediction of Fertilizer with Least Harmful Effects" venture is to leverage facts mining techniques to create a predictive version that can propose the most effective fertilizer to use based totally on diverse environmental and soil elements. The goal is to offer accurate and informed fertilizer pointers that not simplest meet the agricultural wishes but additionally minimize the harmful environmental influences associated with fertilizer software.

The key objectives include:

• Build a predictive model: The most important purpose is to use facts mining techniques, along with logistic regression, choice bushes, and random forests, to broaden a version so one can forecast which fertilizer is first-rate for a sure set of soil and environmental variables. The intention of this model is to correctly recommend fertilizer choices depending on enter characteristics including temperature, humidity, moisture, crop type, soil type, and levels of phosphate, nitrogen, and potassium.

• Minimize unfavourable results: By suggesting fertilizers with the fewest poor impacts, the initiative ambitions to reduce the environmental impact of fertilizer application. The goal is to maximise fertilizer choice to support sustainable farming practices by integrating environmental parameters and soil properties into the predictive version.

• Enhance accuracy with ensemble modeling: The mission intends to create a stacked model that consists of logistic regression, choice tree, and random wooded area processes to boom the accuracy of fertilizer tips. This is finished through taking advantage of the high accuracy of these methods. By combining the blessings of many algorithms, the aim is to produce a prediction model that is greater accurate and strong.

• Utilization of applicable attributes: In order to offer correct fertilizer suggestions, the assignment intends to utilize critical parameters like temperature, humidity, moisture, crop kind, soil type, and nutrient degrees (phosphorus, nitrogen, and potassium). Making positive the predictive version takes into consideration each important issue that influences the environmental effect and efficiency of fertilizers is the purpose.

• Deliver actionable insights: By using the prediction version that has been hooked up, the venture intends to provide agricultural practitioners the facts they want to choose fertilizer accurately. The intention is to offer users with a device that enables powerful and sustainable manage of fertilizer relying on specific soil and environmental conditions.

**4.Introduction:**

The present-day agricultural area need to strike a careful balance between heading off the terrible environmental outcomes of farming operations and satisfying the arena's increasing need for meals. The prudent use of fertilizers, that are critical for increasing agricultural yields but can potentially have terrible influences on the surrounding ecology if improperly dealt with, is one of the predominant troubles in this discipline.

Our research crew commenced a facts mining undertaking to create a prediction model which can recommend the pleasant fertilizers at the same time as reducing their negative effects because they realized how crucial this hassle became. The venture, "Prediction of Fertilizer with Least Harmful Effects," uses modern analytical methods to provide farmers and agricultural experts a honest tool for making the high-quality decisions regarding fertilizer selection and application.

Our initiative intends to close the distance among environmental sustainability and agricultural productiveness via making use of statistics mining and predictive modeling. Through using an intensive dataset encompassing variables like temperature, humidity, moisture, crop kind, soil type, and nutrient levels (phosphorus, nitrogen, and potassium), we've got created a robust predictive model that may advise the satisfactory fertilizers for a given set of circumstances.

Our methodology is based totally at the combination of 3 extraordinarily correct records mining techniques: random forest, choice tree, and logistic regression. These algorithms had been selected because of their respective advantages and capacity to paintings well together in a stacked version. We have developed a extra correct and dependable fertilizer recommendation system that can modify to the diverse agricultural landscapes and environmental conditions observed in numerous locations with the aid of combining the effects of these strategies.

Our project's emphasis on proscribing the terrible affects of fertilizers is one in all its fundamental benefits. Our prediction algorithm considers the environmental effect of various fertilizer formulas in place of most effective optimizing crop output. With this approach, we will each mitigate any harm to the encompassing ecology and find out and advocate fertilizers which could decorate agricultural yield.

Deliverables for the challenge consist of an extensive file that describes the version improvement process, evaluates the model's overall performance, and indicates a fertilizer choice strategy primarily based on the version's predictions. Furthermore, we have created an intuitive interface that enables farmers and agricultural professionals to input their particular soil and environmental variables and reap custom designed fertilizer tips.

Our initiative intends to support ongoing efforts to create sustainable agriculture by way of fusing the power of data mining, predictive modeling, and environmental sense of right and wrong. By putting our fertilizer recommendation gadget into practice, we need to enable farmers and selection-makers to make nicely-knowledgeable choices that strike a stability between environmental stewardship and agricultural wishes, that allows you to in the end result in a more resilient and environmentally friendly meals manufacturing gadget.

**5.Literature Review:**

The problem of forecasting fertilizer application is maximum crop yields while minimizing waste. For example, the authors of [1] demonstrate how farmers can prevent plant toxicity and deficiency and obtain appropriate yields by forecasting fertilizer usage. The use of fuzzy logic systems to lower fertilizer use and boost agricultural output is suggested in Paper [2]. Moreover, [10] emphasizes that improving fertilizer efficiency on its own is not enough to address problems with soil compaction. As demonstrated in [11], a quantitative link between fertilizer consumption, agricultural productivity, nitrogen need, and nitrate residual levels must be established in order to improve fertilizer recommendations. This is in Paper [4], which offers a thorough method for calculating nutrient requirements and taking soil chemical characteristics into consideration. Temperature fluctuations and random rainfall patterns make agricultural yield prediction difficult. On the other hand, precise crop yield prediction can be achieved by utilizing the data mining methods suggested in [3]. No management can raise the danger of nitrogen leaching, which can result in crop loss, according to Laura J.T. Hess et al. in [5]. Furthermore, a new metric for assessing "soil health and quality" is presented in [7], which focuses on improving soil health. [4] et al., Sabri Arik (2016) In this study, we propose a method based on several observable spatial and spectral aspects of the soil samples to predict their functional properties. Extreme Learning Machine (ELM), a relatively new development of the single hidden-layer feed-forward network (SLFN) learning methodology, provides the basis for prediction in this method, which is based on the pre-processing. [5] Vaneesbeer Singh, et al. (2017)This article proposes a strategy that predicts the yield category based on the state of macro- and micronutrients in the dataset using various machine learning approaches. The dataset from Krishi Bhawan (Talab-Tillo) in Jammu was used to predict crop yield. Macronutrients (ph,Oc,Ec,N,P,K,S) and micronutrients (Zn,Fe,Mn,Cu) found in samples taken from various parts of the Jammu District are the parameters included in the data.Following examination, machine learning algorithms are used to forecast the yield category. As a result, the category will specify the crop yield. Several classifier algorithms are employed to solve the classification issue, which is the formulation used to predict crop yield.We also discovered a few more publications that concentrated on using machine learning algorithms to analyze crop health. After reviewing a few research publications, Karishma, M. et al. discovered that the most popular machine learning algorithms are SVM and ANN. The artificial intelligence utilized in this work to incorporate important current technology improved the crop's growth rate [18]. H. Al-Hiary et al. reported a method for identifying and categorizing leaf diseases that included the use of neural networks for recognition in addition to K-means clustering [19]. Sonali Jain used the SVM method in 2013 to extract features from content-based images [20]. Numerous algorithms, including the Random Forest, J48, JRip, and Support Vector Machine (SVM) are included in the Weka(3.8.0) tool [21]. Here, too, we utilized this program to investigate train data with a machine learning technique. Any algorithm can be applied with ease, and we have control over some of its characteristics.

**6.Data Description:**

The information this is provided includes a huge variety of agricultural parameters which are crucial for crop control and development. Temperature, humidity, moisture, crop type, soil kind, nitrogen degrees, potassium content, phosphorus awareness, and the encouraged fertilizer call are most of the vital variables it carries. With each access inside the collection, a awesome aggregate of those factors customized for certain vegetation is represented, providing insightful statistics for enhancing agricultural methods. Temperature affects several physiological methods in plant life, that's important for crop growth and development. Temperature versions are captured by the dataset, which gives vital insights into the variety of temperatures which can be perfect for various crops. Humidity is another important environmental factor that influences the transpiration, uptake of vitamins, and standard health of flowers. The dataset affords a complete understanding of the environmental factors required for ideal crop boom through include records on humidity. A predominant aspect influencing plant water availability and nutrient uptake is soil moisture content. Information on soil moisture ranges is protected in the dataset, which enables with dealing with soil fitness and determining while irrigation is vital. The dataset additionally includes soil kind classifications, including Sandy, Loamy, Black, Red, and Clayey, which have an effect on soil aeration, water retention, and nutrient availability. The dataset's Crop Type statistics identifies the specific crops which can be being tested, which consist of Tobacco, Barley, Wheat, and a number of other crops similarly to Maize, Sugarcane, Cotton, and Paddy. Since every crop has special desires in terms of nutrients and environmental situations, the Crop Type variable is essential for custom designed farming recommendation. The macronutrients nitrogen, potassium, and phosphorus are necessary for the increase, development, and productiveness of plants. Data on the quantities of phosphorus, potassium, and nitrogen wished for ideal crop nutrition are included inside the dataset. These nutrient ranges are important for growing fertilization plans and making sure the vegetation get hold of sufficient nutrients. Based on the included look at of crop nutrient requirements and environmental situations, guidelines for suitable fertilizers are given inside the Fertilizer Name column of the dataset. A form of fertilizers, which include urea, DAP (ammonium phosphate), and specifically formulated blends like 14-35-14 and 17-17-17, are endorsed in accordance with the particular necessities of individual crop varieties and the outside surroundings. All things taken into consideration, agricultural practitioners, teachers, and stakeholders involved in crop control and fertilizer utility can benefit significantly from the dataset. Users can enhance agricultural production and sustainability by way of using the dataset's numerous facts to make nicely-informed selections approximately crop cultivation, nutrient control, and fertilizer selection.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Temperature | Humidity | Moisture | Soil Type | Crop Type | Nitrogen | Potassium | Phosphorous | Fertiliser |
| 26 | 52 | 38 | Sandy | Maize | 37 | 0 | 0 | Urea |
| 29 | 52 | 45 | Loamy | Sugarcane | 12 | 0 | 36 | DAP |
| 34 | 65 | 62 | Black | Cotton | 7 | 9 | 30 | 14-35-14 |
| 32 | 62 | 34 | Red | Tobacco | 22 | 0 | 20 | 28-28 |
| 28 | 54 | 46 | Clayey | Paddy | 35 | 0 | 0 | Urea |
| 26 | 52 | 35 | Sandy | Barley | 12 | 10 | 13 | 17-17-17 |
| 25 | 50 | 64 | Red | Cotton | 9 | 0 | 10 | 20-20 |
| 33 | 64 | 50 | Loamy | Wheat | 41 | 0 | 0 | Urea |
| 30 | 60 | 42 | Sandy | Millets | 21 | 0 | 18 | 28-28 |
| 29 | 58 | 33 | Black | Oil seeds | 9 | 7 | 30 | 14-35-14 |

**7.Architecture:**

Data Visualization

SVM

Logistic Regressions

Proposed Methodology

Random Forest

KNN

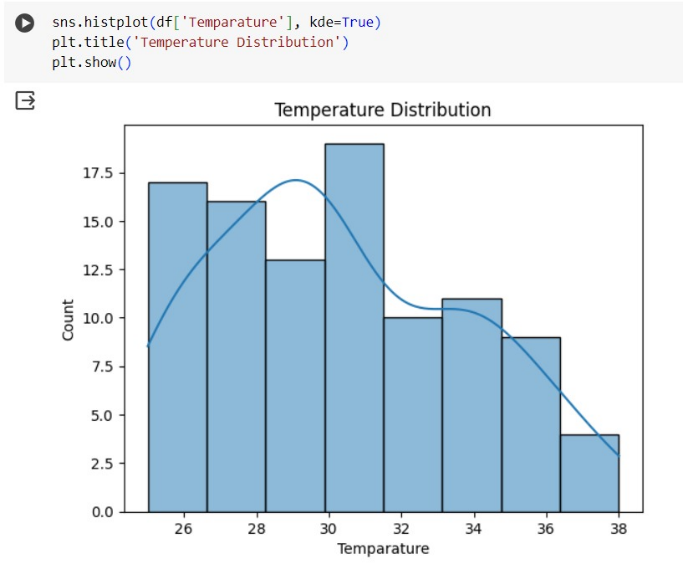
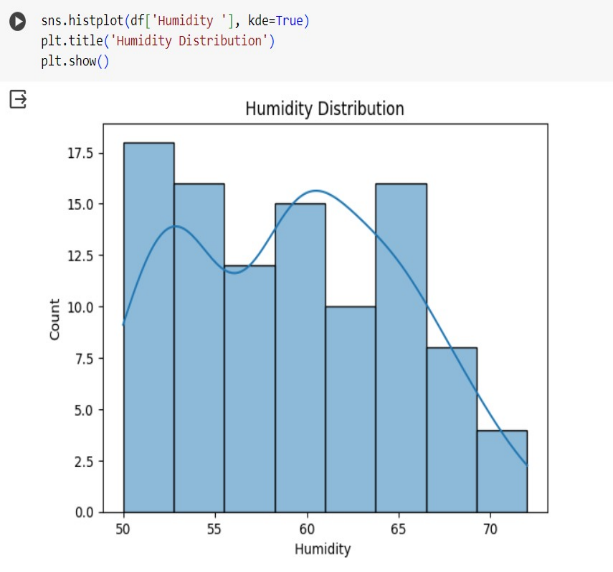
Decision Tree

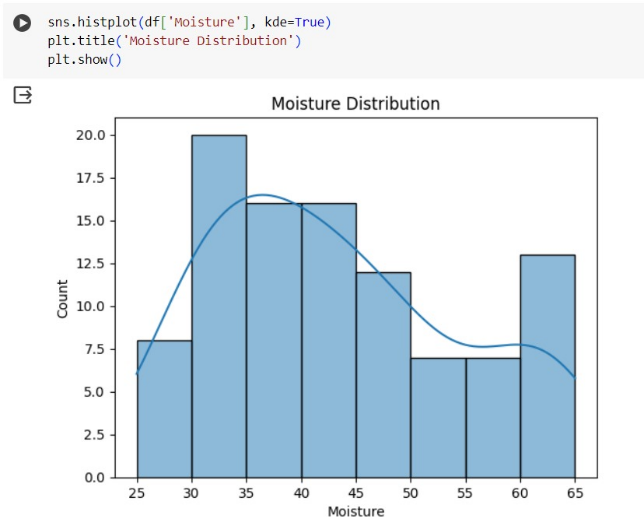
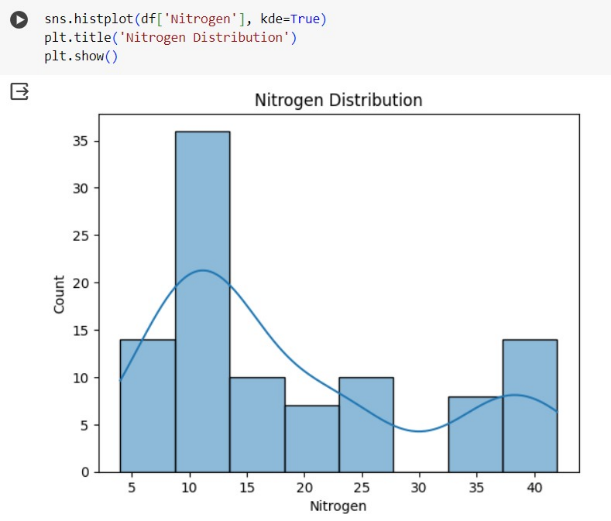
Model Training and Evaluation

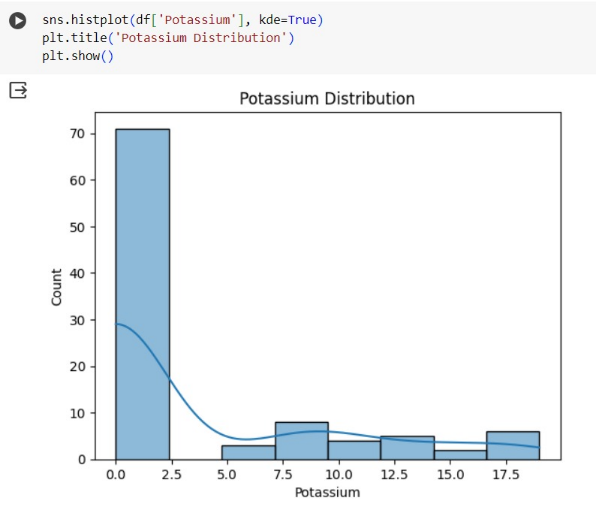
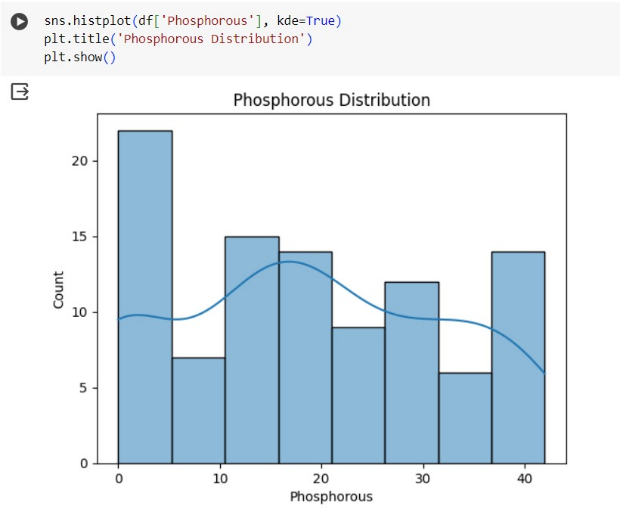
Data Collection

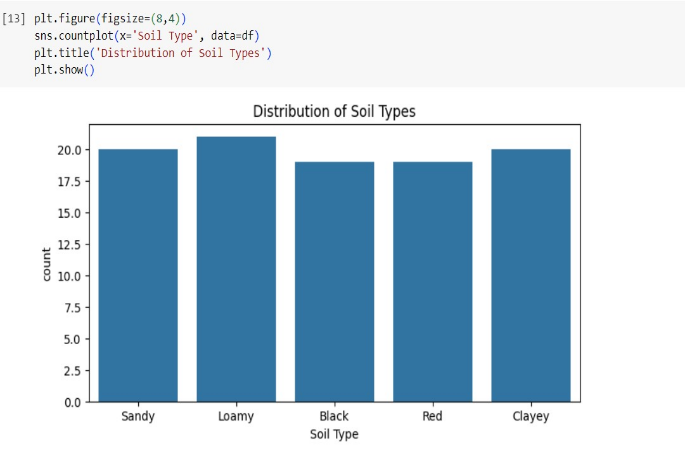
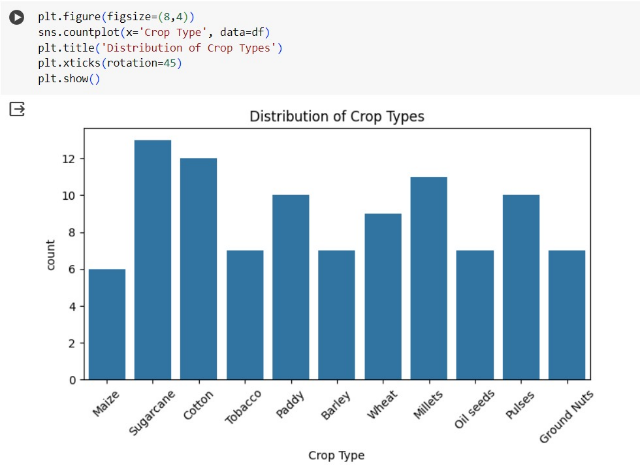
**Data Collection:**

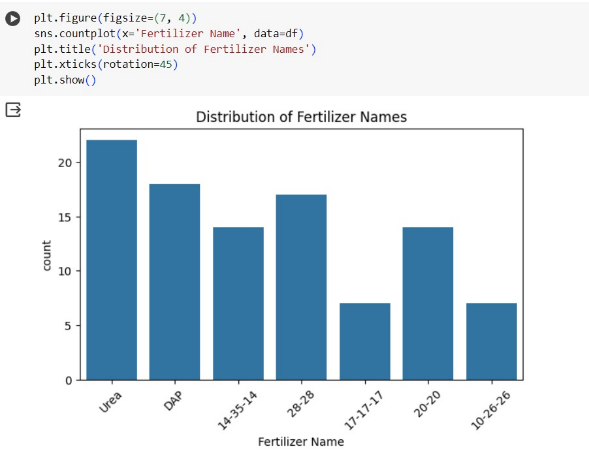
**Data Visualization:**

** **



**Evaluation:**

Metrics: Various evaluation metrics, such as accuracy, precision, recall, support, and the F1 score, were employed to comprehensively assess model performance.

**Comparison:**

Model Strengths and Weaknesses: A comparative analysis highlighted the strengths and weaknesses of each model. The stacking classifier model made provides the best accuracy as compared to the other models with least harmful effects.

**8.Proposed Methodology:**

**Models Used:**

**Logistic Regression:**

Logistic Regression is a effective gadget studying algorithm used for binary type responsibilities. It is a sort of supervised studying algorithm that fashions the chance of a binary based variable based totally on one or more impartial variables.

The key concept behind Logistic Regression is to apply the logistic sigmoid characteristic to map the linear aggregate of the impartial variables to a opportunity value among zero and 1, which represents the likelihood of the established variable being 1 (or "True"). The logistic regression equation is given via P(Y=1in which P(Y=1opportunity of the dependent variable being 1 given the impartial variables X, β₀ is the intercept term, and β₁, β₂, ..., βₙ are the coefficients of the unbiased variables X₁, X₂, ..., Xₙ.

The version is skilled using maximum chance estimation, which targets to discover the values of the version parameters that maximize the chance of the determined statistics. Once the model is trained, it could be used to make predictions on new data by using evaluating the output possibility to a chosen threshold (normally 0.5) to classify the instances as zero or 1. The version's performance is evaluated using metrics such as accuracy, precision, recollect, and F1-rating.

**Decision Tree:**

Decision trees are a popular machine learning algorithm for classification and regression tasks. It is a tree-like decision model of a supervised learning process based on inputs.

The algorithm works by iteratively dividing the data into smaller units based on which feature provides the greatest information gain or the greatest entropy reduction Each internal node in the decision tree represents a feature, . and the branches represent possible values ​​for that object. The leaf node represents the final classification or regression result.

The main advantage of decision trees is that they are easy to interpret and understand, because the decision process is visually represented in a tree structure where both statistical and classification data can be handled, and it is robust against outliers and values of the loss of. Decision trees can also capture nonlinear relationships in the data, making them an alternative to machine learning problems.

**KNN:**

K-Nearest Neighbors (KNN) is a easy and intuitive system mastering set of rules used for each category and regression responsibilities. It is a kind of example-primarily based learning, where the model makes predictions based on the similarity among the enter statistics and the schooling information.

The way KNN works is by finding the K closest (maximum comparable) information factors to the input information factor after which creating a prediction based on the bulk class (for category) or the average fee (for regression) of these K friends.

The key benefit of KNN is that it is simple to recognize and enforce, and it does not require any assumptions about the underlying data distribution. It is also effective in managing non-linear relationships and complicated choice barriers. However, KNN can be computationally steeply-priced for massive datasets, and it is able to be sensitive to the choice of the K parameter and the space metric used.

**Random Forest:**

Random Forest is a effective ensemble mastering algorithm used for each class and regression responsibilities in system getting to know. It works by means of creating a couple of choice trees, every skilled on a random subset of the education information and a random subset of the functions.

The key idea in the back of Random Forest is that by way of combining the predictions of a couple of decision bushes, the algorithm can improve the overall accuracy and balance of the version. Each person selection tree within the Random Forest is a susceptible learner, but while blended, they form a strong and strong version.

The most important benefit of Random Forest is its capacity to address a wide range of facts types, its robustness to outliers and noise, and its capacity to offer function importance statistics. It is also highly easy to use, as it has few hyperparameters to tune, and it is able to take care of each numerical and express records effectively.

**SVM:**

Support Vector Machine (SVM) is a powerful machine getting to know set of rules used for category and regression responsibilities. It is especially effective for binary category troubles. The main goal of SVM is to discover an most excellent hyperplane that separates the records points of various training, maximizing the margin among them.

SVM works through transforming the statistics into a higher-dimensional function area and finding the hyperplane that fine separates the training. It uses help vectors, which are the records points closest to the hyperplane, to determine the selection boundary. SVM can cope with both linearly separable and non-linearly separable records by way of the usage of different kernel features, inclusive of linear, polynomial, or radial basis function (RBF).

The key benefits of SVM consist of its capacity to handle excessive-dimensional records, its effectiveness in dealing with complex datasets, and its resistance to overfitting. SVMs are extensively used in diverse packages, inclusive of text type, photo category, junk mail detection, gene expression analysis, and anomaly detection.

However, SVMs can be computationally pricey, in particular with massive datasets. Additionally, selecting the ideal kernel function and tuning the hyperparameters may be tough. Nevertheless, SVM stays a famous and powerful algorithm within the field of gadget studying, presenting correct and strong solutions to type and regression problems.

**9.Novelty:**

The "Prediction of Fertilizer with Least Harmful Effects" project is a noteworthy example of a forward-thinking initiative that skillfully combines cutting-edge data mining methods with a strong dedication to environmental sustainability. The creative method of creating a layered predictive model by merging several effective algorithms—logistic regression, decision tree, and random forest—is the fundamental innovation of this project. The initiative has produced a more reliable and flexible decision-support system that can offer precise fertilizer recommendations catered to various agricultural settings by utilizing the special capabilities of these particular methodologies. The initiative differs from traditional fertilizer management tactics, which frequently rely on a single analytical tool, in that it uses an integrated data mining approach.The project's unique quality is its unwavering focus on limiting the detrimental environmental repercussions of fertilizer application, which goes beyond its technical components. In contrast to conventional fertilizer recommendation systems, which put an emphasis on optimizing crop yields, this study adopts a holistic approach by taking into account the possible effects of various fertilizer formulations on the local ecology. Utilizing an extensive dataset that includes variables like temperature, humidity, wetness, crop type, type of soil, and nutrient levels, the predictive model of the project can determine fertilizers that minimize environmental degradation while still meeting agricultural needs. In the realm of sustainable agriculture, this emphasis on striking a balance between sustainability and productivity sets a new benchmark. The creation of an intuitive interface that enables farmers and agricultural specialists to quickly input their unique soil and environmental variables and obtain customized fertilizer recommendations is another example of the project's inventive spirit. The project is further distinguished by the technology's practical usability and accessibility, which help to close the gap between research and practical application. The "Prediction of Fertilizer with Least Harmful Effects" project marks a revolutionary step forward in the pursuit of a more resilient and ecologically conscious food production system through the smooth integration of data-driven insights and environmental consciousness.

**12.References:**

[1] Jahan, N. and Shahariar, R., 2020. Predicting fertilizer treatment of maize using decision tree algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 20(3), pp.1427-1434.

[2] Bondre, D.A. and Mahagaonkar, S., 2019. Prediction of crop yield and fertilizer recommendation using machine learning algorithms. International Journal of Engineering Applied Sciences and Technology, 4(5), pp.371-376.

[3] Chimkode, S., Kanjikar, A., Patil, A., Kumar, K. and Bemelkedkar, A., 2023. SMART USE OF FERTILIZER IN AGRICULTURE USING MACHINE LEARNING. Int. Res. J. Mod. Eng. Technol. Sci, 6200.

[4] Kanuru, L., Tyagi, A.K., Aswathy, S.U., Fernandez, T.F., Sreenath, N. and Mishra, S., 2021, January. Prediction of pesticides and fertilizers using machine learning and Internet of Things. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[5] Pandiarajaa, P., 2021. A survey on machine learning and text processing for pesticides and fertilizer prediction. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), pp.2295-2302.

[6] Meng, L., Liu, H., L. Ustin, S. and Zhang, X., 2021. Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods. Remote Sensing, 13(18), p.3760.

[7] Krutika Hampannavar, Vijay Bhajantri, Shashikumar G. Totad “Prediction of Crop Fertilizer Consumption,” Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),2018, PP.1-5

[8] Yulong Yin, Hao Ying, Huifang Zhen, Qingsong Zhang, Yanfang Xue, Zhenling I, “Estimation of NPK requirements for rice production in diverse Chinese environments under optimal fertilization rate,” Agricultural and Forest Meteorology, vol-279, 2019, PP. 1-6

[9] Laura J.T. Hess, Eve-Lyn S. Hinckley, G. Philip Robertson, Pamela A. Matson, “Rainfall intensification increases nitrate leaching from tilled but not no-till cropping systems in the U.S. Midwest,” Agriculture, Ecosystems & Environment,vol-290,2020, PP.1-10

[10] Usman Ahmed, Jerry Chun-Wei Lin, Gautam Srivastava, Youcef Djenouri, “A nutrient recommendation system for soil fertilization based on Evolutionary Computation,” Computers and Electronics in Agriculture, vol-189, 2021, PP. 1-7

[11] Rubby Aworka, Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Charles Lebon Mberi Kimpolo, Tarik Nahhal, Moez Krichen, “Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries,” Smart Agricultural Technology, vol-3, 2022, PP. 1-9

[12] Ruchita T, Shreya B, Prasanna D, and Anagha C (2017). Crop Yield Prediction using Big Data Analytics, Volume 6, Issue 11, IJCMS.

[13] Rohit Kumar Rajak, Ankit Pawar, Mitalee Pendke, Pooja Shinde, Suresh Rathod, and Avinash Devare (2017). Crop Recommendation System to maximize Crop yield using Machine Learning. Issue 12 IRJET.

[14] Sadia A, Abu Talha K, Mahrin Mahia, Wasit A, and Rashedur M.R.(2018). Analysis of Soil Properties and Climatic Data To Predict Crop Yields and Cluster Different Agricultural Regions of Bangladesh, IEEE ICIS 2018 (pp.80-85).

[15] Dutta, S., Chakraborty, S., Goswami, R., Banerjee, H., Majumdar, K., Li, B. and Jat, M. L., “Maize yield in smallholder agriculture system-An approach integrating socio-economic and crop management factors,” PloS one, vol. 15, no. 2, e0229100, 2020.