Research on Intelligent Planting Optimization of Soil Environment Based on Machine Learning Method

Yu Cao
School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, Henan Province, 450000, China

Keywords: Multiple Logistic Regression, Random Forest Model, Support Vector Machine Model, Machine Learning, Intelligent Planting

Abstract: Crops not only provide nutrients for human growth but also are the raw materials for many industries. There are a lot of food supply and food security issues. Crops play an important role in people's lives. This study trained an intelligent model to judge the high-yield crops in this region based on the content of some elements in the soil and the characteristics of temperature and humidity. Training multiple logistic regression, random forest, and support vector classification (SVC) models provide an initial model for this investigation. The parameters of this initial model were optimized and adjusted to find the best model based on the soil intelligent planting problem. Overall accuracy, precision, and recall were selected as the evaluation indexes of the three models. The overfitting of the three models is alleviated by grid search cross-validation. Finally, it is found that the accuracy of the SVC algorithm reaches 0.99 on both the training set and the test set. This model has outstanding performance in intelligent planting problems. Training the smart planting model based on the soil environment can help farmers better select the best crops in different soils to achieve high yields. Dramatic improvements in land use efficiency and yield could ease today's global food supply problems. The subsequent use of more accurate climate mitigation data collection equipment is expected to train an intelligent planting model that combines land mitigation and climate environment.

1 INTRODUCTION

Crops provide people with essential nutrients such as carbohydrates, proteins, fats, vitamins, and minerals. The world's population, estimated at 7.7 billion, consumes 2.5 billion tons of food annually, the majority of which is derived from crops. Crops are also important raw materials for various industries. Crops can provide different kinds of raw materials for textile, paper, chemical, pharmaceutical, energy, and other industrial sectors, such as cotton, linen, bamboo, wood, oil, sugar, starch, cellulose, ethanol, and so on. Companies can create significant economic value by further processing crops (Qikai, 2016). However, global food supply and security are threatened by multiple factors, such as climate change, geopolitical conflicts, the COVID-19 pandemic, and volatile food and fuel prices. The crisis has put hundreds of millions of people around the world at risk of hunger and malnutrition, especially in some low-income countries and regions (World Agric, 2022). According to the United Nations World Food Programme, around 345 million people in 82 countries are currently facing severe hunger (World Agric, 2019). On the other hand, with the improvement of medical technology, the death rate of the population is decreasing year by year, and the population is constantly growing. This factor has also led to an increasing number of people, and the phenomenon of food shortage is increasing. With the advancement of data acquisition technology, people can collect the desired data through various sensors (Fengshao, 2023). AI data-based machine learning algorithms have also been improved. These algorithms have been widely used in finance, biology, chemistry, and other fields. In agriculture, many researchers have made a lot of crop yield prediction techniques based on climate characteristics, environmental conditions, and other characteristics. For example, Lontsi Saadio Cedric et al. collected climate data, weather data, agricultural yield, and chemical data in West Africa. They used this data to build their system based on decision trees, multiple logistic regression, and K-nearest neighbor models.

Cao, Y.
Research on Intelligent Planting Optimization of Soil Environment Based on Machine Learning Method.
DOI: 10.5220/0012824000004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 481-487
ISBN: 978-989-758-690-3

They then adjusted the hyperparameters to obtain a model that did not overfit to predict the yield of six local crops (Saadio et al, 2022). He collected satellite meteorological data from 98 districts and counties in Jilin and Shanxi provinces. Based on these data, he used regression algorithms from random forests and support vector machines, as well as deep learning neural network models to predict local corn yields. Ultimately, each model's prediction effect is assessed and examined using the mean square error, mean absolute percentage error, and root mean square error. The study confirms that the neural network model is the most effective one, and it also demonstrates the viability of machine learning techniques in the agricultural sector (Junxiu, 2021). There is not much research on machine learning classification algorithms in agriculture, but these algorithms can use satellite images or images collected by aircraft to classify different crops in farmland. This helps to monitor the health of the field and to manage irrigation and fertilization promptly. In this study, based on the classification algorithm of machine learning, a model is designed to judge which crops can produce high yields in a given soil environment based on soil conditions. This model not only improves the efficiency of land use but also produces more food. The application of this model can help alleviate the problem of food shortage.

This study will use three machine learning methods, which are multiple logistic regression, support vector machine classification, and random forest. This study will first train the model of each algorithm separately based on the data. Secondly, the optimal model of each algorithm is obtained by cross-verifying the selection of hyperparameters. Finally, the overall accuracy, precision, recall, and other indicators of the best model of each algorithm are compared to get the best model in this study. This paper aims to investigate machine learning-based intelligent planting optimization of soil environment. encourage tillering, strengthen roots against cold stress, and frequently speed up ripening (Miaomiao et al, 2023).

Potassium: A nutrient obtained by plants from soil and fertilizers. Potassium can improve disease resistance, stem strength, drought tolerance, and winter survival of plants (Hashim & Mohammed, 2023).

Soil temperature: The average soil temperature for biological activity is between 50 and 75 degrees Fahrenheit, similar to the temperature of the human body. These values are conducive to the normal life functions of the Earth's biota, such as decomposing organic matter, mineralizing nitrogen, absorbing soluble substances, and metabolism (Huiyong et al, 2008).

pH: The pH value indicates the acidity or alkalinity of the soil. The pH range of 5.5 to 6.5 is ideal for plant growth, as this range guarantees the availability of nutrients (Chenxiao, 2022).

Rainfall: Rainfall also affects how fast crops grow from seed, including when they are harvested. Balanced rainfall and proper irrigation can accelerate plant growth, which can reduce germination times and the time between planting and harvesting (Podzikowski et al, 2023).

2 METHOD

2.1 Data Set Description

Table 1 shows some of the data features and labels in the data set. This study is based on soil nitrogen content, phosphorus content, potassium content, temperature, humidity, soil pH, rainfall data, and high-yield crop categories. To make the subsequent data analysis more beneficial, all table data will be reserved for 3 decimal places. This data comes from Kaggle. Kaggle is an online platform for data scientists, machine learning engineers, and other professionals in data-related fields.

Nitrogen: One of the basic elements of plants, the main component of chlorophyll. Chlorophyll is the compound that enables plants to photosynthesize (Yanxiao et al, 2022).

Phosphorus: Phosphorus is essential for both the growth of new tissues and cell division. In plants, phosphorus is also involved in the intricate process of energy conversion. When phosphorus is added to soils with low levels of accessible phosphorus, it can

Table 1: Partial data on factors affecting crop growth and corresponding high-yielding plants.

	Nitrogen	Phosphorus	Potassium	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879	82.002	6.502	202.935	rice
1	85	58	41	21.770	80.319	7.038	226.655	rice
2	60	55	44	23.004	82.320	7.840	263.964	rice
3	74	35	40	26.491	80.158	6.980	242.864	rice
4	78	42	42	20.130	81.604	7.628	262.717	rice

Table 2: Statistical overview of factors affecting crop growth.

	Nitrogen	Phosphorus	Potassium	temperature	humidity	ph	rainfall
mean	50.551	53.362	48.149	25.616	71.481	6.469	103.463
std	36.917	32.985	50.647	5.063	22.263	0.773	54.958
min	0	5	5	8.825	14.258	3.504	20.211
25%	21	28	20	22.769	60.261	5.971	64.551
50%	37	51	32	25.598	80.473	6.425	94.867
75%	84.25	68	49	28.561	89.948	6.923	124.267
max	140	145	205	43.675	99.981	9.935	298.560

2.2 Feature Engineering

Table 2 shows a statistical summary of the factors affecting crop growth. The three elements Nitrogen, Phosphorus, and Potassium are necessary for the growth of crops. The content of each element required by plants has a large span and significant standard deviation, which can indicate that the amount required by each type of plant for these three elements is significantly different. Most of the temperatures required for crop productivity are concentrated around 25, with only a few crops at some extreme temperatures. The standard deviation of pH in the soil is the smallest. The pH required for high yield of crops is concentrated around neutral and slightly acidic. The difference between humidity and rainfall is also relatively large, which is also conducive to our division and practical application of crops. Because the climate is different around the world, the rainfall is also different, and different areas are suitable for different high-yielding crops.

Pearson's formula is used to study the multiple correlations among the parameters (Layeb, 2023). Figure 1 shows a hotspot map of correlations between features. There was a strong positive correlation between the content of Phosphorus and Potassium, which made different types of crops produce high yields. The above situation shows that the content of Phosphorus and Potassium in an environment of high yield should increase and decrease consistently. The Nitrogen element shows a certain negative correlation with the above two elements. Other features have relatively small dependencies.

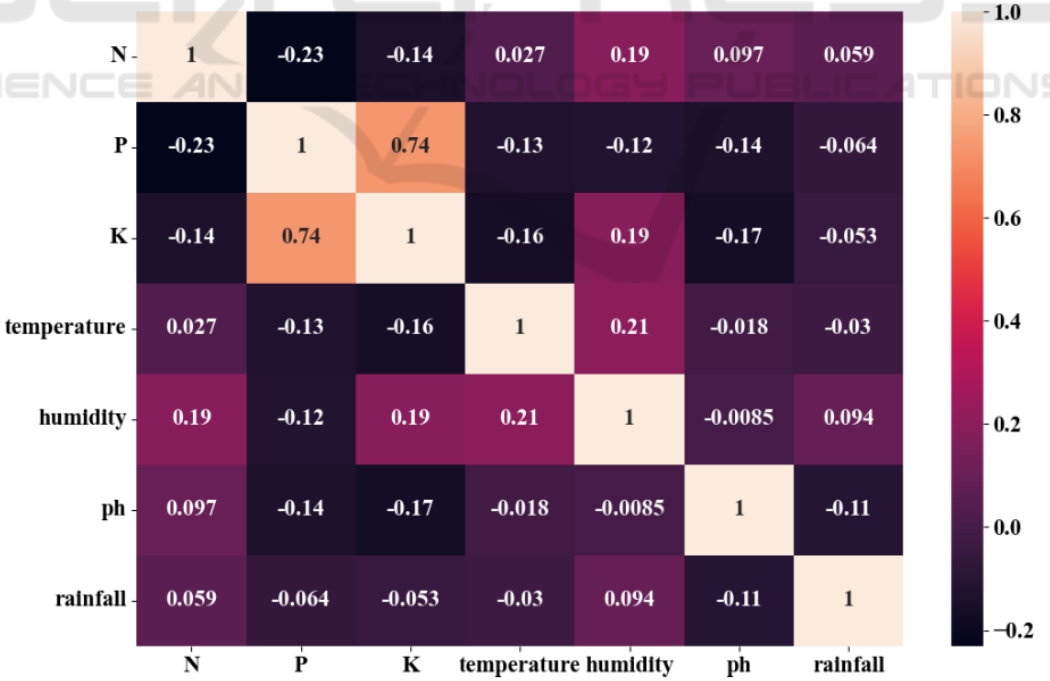

Figure 1. Pearson correlation between features (Photo/Picture credit: Original).

2.3 Proposed a Soil Environment Intelligent Planting Model Based on Machine Learning

2.3.1 Multiple Logistic Regression (LR)

A statistical technique for binary classification that forecasts the likelihood that a result will fall into one of two categories is called logistic regression. Multivariate categorization can also be accomplished with logistic regression. An important technique in statistics and machine learning, it simulates the link between the logarithmic probabilities of independent and dependent variables. (Leijia, 2023).

For each category j, the weight of each feature is represented by formula (1), The number of features is represented by the superscript, the number of categories by the subscript, the total number of features by n, and the total number of categories by N.

$$Z_j = w_j^0 x^0 + w_j^1 x^1 + w_j^2 x^2 + \ldots + w_j^n x^n \quad j = 1,\ldots,N \tag{1}$$

The probability that the classification result is a class for the input point x is given by formula. (2).

$$P(y = j|Z_j) = \frac{e^{Z_j}}{\sum_{k=1}^{N} e^{Z_k}} \tag{2}$$

Create training and test sets from the data set. Thirty percent of the whole data set is the test set, while the remaining seventy percent is the training set. Each category of data in the training set will be calculated by formula (1) and formula (2) to get its probability in each category, and the maximum probability will be selected as its final predicted value. The model's performance is assessed using the cross-entropy loss function (formula (3)), which converges to a minimum value. Using gradient descent algorithm to reduce the loss function. Finally, the best multiple logistic regression classification model is obtained.

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_w\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_w\left(x^{(i)}\right)\right)\right] \tag{3}$$

2.3.2 Random Forest Model (RF)

Compared to a single decision tree, this potent tree-based ensemble learning system outperforms it. By using retractable sampling to generate multiple different training subsets, we can train an independent decision tree for each subset. We can then combine these decision trees into an integrated model, and for the prediction task, have the trees vote to get the final prediction result. To make the decision tree in tree integration more diverse, we can further introduce randomness in the selection of partition features for each node. To be more precise, we can choose a subset of size K (K less than N) at random from all N features at each node. From this subset, we can choose the feature that has the maximum information gain to be the partitioning feature. (Ruiyan, 2022).

Steps to create a random forest:

Given a training set size N and the number of trees to be created n, for each tree, N samples are taken from the total sample to form the training set of the tree. Repeat n times to form n numbers to form the forest

For M features of each sample, m features are selected as the training features (m < M).

Use these m features to make the tree extend as far as possible

Enter the new data you want to predict into n trees and vote for the result.

An indicator of a data set's purity is entropy. Higher entropy indicates greater impurity, whereas lower entropy indicates higher purity. When building a decision tree, we can use entropy to decide which feature should be divided at the node. In general, we choose the feature that will reduce entropy the most. This means that you want to make the subset as pure as possible by splitting it. In classification problems, there are two formulas for calculating entropy: information entropy (formula (4)), and Gini impurity (formula (5)).

$$H(D) = \sum_k p_k \log_2 p_k \tag{4}$$

$$H(D) = \sum_k p_k(1 - P_k) \tag{5}$$

2.3.3 Support Vector Machine Model (SVM)

Finding a hyperplane to divide the sample into segments is the goal of the binary classification algorithm known as support vector machine, where the idea behind segmentation is to maximize the interval. Let the hyperplane be $\omega^T x + b = 0$, given a sample x, $\omega^T x + b$ Represents the distance to the hyperplane. Suppose that after extending the space point to N-dimensional space, the distance from the point to the hyperplane is formula (6) (Saadio et al, 2022):

$$d = \frac{y*|\omega^T x + b|}{||\omega||} \tag{6}$$

According to the Lagrange multiplier method, KKT condition, and duality problem, we can calculate a set of w values of the optimal SVM objective function according to the following steps.

$$\min \frac{1}{2}||\omega||^2 \text{s.t } 1 - y * (\omega^T x + b) \leq 0 \quad (7)$$

Since the original objective function is a downward convex function, according to the duality problem, we can see that a) must be a strong duality problem, which can be converted to

$$L(w, b, \lambda) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{n} \lambda_i [1 - y_i(\omega^T x_i + b)] \quad s.t \lambda_i \geq 0 \quad (8)$$

The process of building an SVM model consists of three stages: feature selection, feature transformation, and hyperplane creation

3 RESULTS AND DISCUSSION

3.1 Model Evaluation

In this investigation, the model's quality was assessed using three indicators: recall, accuracy, and precision. The training of the model on the training set and the verification on the test set can well judge the accuracy of the training model and whether there is overfitting and generalization ability through these three indexes

Table 3: Three model evaluation indicators using default parameters.

	Accuracy		Precision		Recall	
Model	Train	Test	Train	Test	Train	Test
LR	0.975	0.948	0.976	0.951	0.976	0.950
RF	1.000	0.995	1.000	0.995	1.000	0.995
SVC	0.975	0.978	0.978	0.982	0.978	0.981

In the case of the logistic regression model, there is an overfitting issue and an excessive variance between the accuracy on the training and test sets.When compared to the other two models, its prediction accuracy is not as good, and its precision and recall rate are also evident. For the random forest model, its score on the training set has reached an astonishing 1, and there is a high possibility of overfitting. However, the results on the test set are very good, indicating that after some adjustments and optimization, the model has a great accuracy to solve this problem. The SVC algorithm's training set score and test set score are fairly similar, and the features acquired in the training set are practically applicable.

Hyperparameter selection and cross-validation are employed to address the overfitting issue and improve the model training's predictive accuracy in order to address the aforementioned issues.

3.2 Training Optimization

To ensure that the model does not learn too much from the training set. The author uses grid search cross-validation to identify the model's ideal hyperparameters. The grid search is a parameter tuning technique where the parameters are adjusted within a specified range with a defined step size. These modified parameters are used to train the learner, and the parameter that produces the best accuracy on the validation set is chosen. This process involves training and comparing multiple models. Using the GridsearchCV library, takes the machine learning model, the hyperparameter selection grid, and the number of cross-validations as inputs. The best estimator (machine learning model) of the best parameters is given.

The maximum number of method convergence iterations and the reciprocal of the regularization coefficient are the linked hyperparameters for the logistic regression algorithm. The regularization coefficient can assist the model in overcoming overfitting, significantly increase the model's accuracy when applied to unknown data, and streamline the intricate model by getting rid of mistakes and noise. If the number of algorithm iterations falls within a suitable range, the model's efficiency can be raised. After using grid search and cross-validation, the best model parameters of the multiple logistic regression algorithm in solving this problem are {C = 0.01, max_iter = 700, cv = 5}. The details of the model are shown in Table (4).

For the random forest algorithm, the associated hyperparameter is selected as the number in the created forest.The quantity of appropriate trees in the forest makes sure the model picks up the features of the data set without becoming overfit. After using grid search and cross-validation, the best model parameters of the multiple logistic regression algorithm in solving this problem are {'max_depth'= 6, 'n_estimators'= 100, cv = 5}. The details of the model are shown in Table (5).

Table 4: Multiple logistic regression to optimize details.

mean_fit_time	param_C	param_max_iter	params	mean_test_score	rank_test_score
0.3864604	0.01	700	{'C': 0.01, 'max_iter': 700}	0.977	1

Table 5: The random forest algorithm optimizes details

mean_fit_time	param_max_depth	param_n_estimators	params	mean_test_score	rank_test_score
0.190494919	6	100	{'max_depth': 6, 'n_estimators': 100}	0.990	1

For the SVM classification algorithm, the related hyperparameters are the penalty coefficient C of the error term and the type of kernel function. The size of C has a strong correlation with the accuracy and generalization ability of training results Choosing the appropriate penalty coefficient has a great influence on the fitting ability and generalization ability of the model. After using grid search and cross-validation, the best model parameters of multiple logistic regression algorithms in solving this problem are {'C'= 10, 'kernel'= 'rbf', gamma='auto', probability=True}. The details of the model are shown in Table (6).

Table 6: Support vector machine classification optimization details.

mean_fit_time	param_C	param_kernel	params	mean_test_score	rank_test_score
0.12366333	10	rbf	{'C': 10, 'kernel': 'rbf'}	0.985	1

Table (7) shows the accuracy rate, precision rate, and recall rate of the three optimized models after training. Compared with the default parameters before optimization, the data detection on the test set is more accurate. Furthermore, there is very little score difference between the training and test sets, and the overfitting issue has also been resolved. The LR model has an overall accuracy of approximately 0.98, the RF model has an overall accuracy of approximately 0.99, and the SVC model has an overall accuracy of approximately 0.99. The model construction of the three algorithms for this problem is relatively good, but by comparing the time of model fitting in Tables 4, 5 and 6, Out of the three approaches, the support vector machine's classification model is the most effective. In general, all three models show their ability to solve this problem after being optimized.

Table 7: Three model evaluation indexes using optimization parameters

	Accuracy		Precision		Recall	
Model	Train	Test	Train	Test	Train	Test
LR	0.989	0.983	0.989	0.983	0.989	0.983
RF	0.989	0.990	0.989	0.991	0.989	0.991
SVC	0.989	0.992	0.989	0.992	0.989	0.992

4 CONCLUSION

This study is based on soil nitrogen content, phosphorus content, potassium content, temperature, humidity, soil pH, and rainfall data to intelligently select the most suitable for high-yield crops in a specific environment. The data in the data set is preprocessed first to make the formation of the model faster. Second, to find the association between the features and hidden information inside the data, feature engineering is used to the data set. Three models were employed in this study to train the data in an initial stage. It is evident from the experimental data that the three models are not very convincing in the training and test sets and have certain flaws. Finally, a more comprehensive model for this problem is generated by applying the grid search cross-validation method to optimize the above model. The SVC model has a considerable advantage in fitting time and not only has an accuracy rate of 0.99 in the validation of the test set, but it also has a precision and recall rate that are close to 0.99. Due to the large amount of data and the significant

classification of each crop feature, the three models have significant performance in the three indicators after the final optimization. The difference in overall accuracy lies in the difference in the division of individual crops. The SVC model has a more significant effect than the above two models on the environmental high yield of aconitum bean.

Machine learning is more and more widely used in today's life, and many researchers in agriculture have conducted in-depth research on the yield prediction of a single or several crops. This research can help people to know what kind of soil in their country or region is suitable for growing crops. This practice can prevent the loss of yield and income caused by random planting, and help people to maximize their profits in farming. If some areas are not suitable for planting some crops, the soil environment can be adjusted according to the range values of various characteristics in the above study to achieve the purpose. This study can also be combined with the other yield models described above to not only identify the best crops to grow in the area, but also to predict the yield. In future development, the model can be further improved to add more characteristics such as climate, the region where the soil is located, the amount of sunshine and other atmospheric environments. Combining the soil environment with the atmosphere further improves the smart planting model, allowing more accurate classification of which crops are more productive in a given region.

REFERENCES

Y. Qikai, the Ningxia Hui Autonomous Region, Ningxia Southern Chemical Technol. Co. (2016).

International food and agriculture trends, World Agric. (05):119-121 (2022).

Global Food Crisis Report, World Agric. (05): 96 (2019).

Z. Fengshao, Wirel. Interconn. Technol. 19(02): 109-111 (2023).

L. C. Saadio, Y. W. A. Hamilton, A. Rubby, et al, Smart Agric. Technol. 2 (2022).

H. Junxiu, Jilin Agric. Univ. (2021).

F. Yanxiao, J. Yongda, L. Weiwei, For. Sci. Technol. Inf. 54(02): 109-111 (2022).

C. Miaomiao, Y. Bin, R Guangqian, et al, Flora. 309 (2023).

B. A. Hashim, H. A. Mohammed, IOP Conf. Ser. Earth Environ. Sci. 1262(8) (2023).

Y. Huiyong, L. Meixue, S. Jieya, et al, Priv. Sci. Technol. (03):102 (2008).

G. Chenxiao. Shanxi Agric. Univ. Agricultural University (2022).

L. Y. Podzikowski, M. M. Heffernan, J. D. Bever, Front. Ecol. Evol. 11(2023)

A. Layeb, Int. J. Intell. Syst. Appl. 15(2):37-42 (2023).

L. Leijia, Southwest Univ. (2023).

L. Ruiyan, Nanfang Agric. Machinery, 53(22):63-65+87 (2022).