An artificial intelligence analysis of natural conditions in agriculture

Yi Du[1]

[1]Boston University, Boston, MA, 02215, United States

duyi233@bu.edu

Abstract. In this study, we developed a machine learning algorithm for crop recommendation based on a dataset containing environmental features and corresponding optimal crop choices. The algorithm was trained on a set of seven numerical features, including temperature, humidity, pH value, rainfall, and levels of nitrogen, phosphorus, and potassium in the soil. We evaluated the performance of multiple classification algorithms, including naive Bayes, logistic regression, support vector machine (SVM), decision trees, random forests, and neural networks. Our results showed that neural networks performed the best. However, SVM, which provided only unsatisfactory results initially, also achieved pleasing results after we redesigned its structure and adjusted its parameters. This system has the potential to assist farmers in choosing the most suitable crops for their specific environments, leading to increased crop yield and profitability.

Keywords: crop recommendation, intelligent agriculture, supervised learning, support vector machine, neural network.

1. Introduction

Agriculture plays a crucial role in the economic and social development of many countries. As the world's most populous country and a major global player in the agricultural sector, China is facing the challenge of sustainably increasing food production to meet the needs of its growing population, while also preserving natural resources and minimizing negative impacts on the environment [1]. To address this challenge, it is essential to optimize crop selection and cultivation practices to match specific local conditions, such as temperature, humidity, soil nutrients, and rainfall.

In this project, we aimed to address these challenges by developing a machine learning algorithm for crop recommendation. Given a set of environmental features, our algorithm is able to recommend the most suitable crops for a specific location, with the goal of maximizing yield and minimizing negative impacts on the environment. By leveraging the power of machine learning, we sought to create a system that could assist farmers in making informed decisions about which crops to grow, ultimately leading to increased crop yield and profitability.

There have been a number of past studies that have attempted to address the challenge of optimizing crop selection and cultivation practices [2]. These studies have often relied on expert knowledge or statistical models to make recommendations, but these approaches can be limited in their accuracy and adaptability to changing conditions. We, nevertheless, take a different approach in this paper by using machine learning algorithms to make crop recommendations. We believe that this approach has the potential to be more accurate and adaptable than past methods, as it is able to automatically extract features from provided datasets and generate results based on patterns and trends in the data.

The contents of this paper will be organized as follows: First, we prepare and analyze our dataset, providing visualization for several critical statistical features of the dataset. Next, we describe the machine learning algorithms and evaluation metrics that we used. Then, we present our results and compare the performance of different algorithms. Finally, we discuss the implications of our findings and describe the contributions of our work.

Overall, we believe that our machine learning algorithm for crop recommendation represents a significant advancement over past methods, and we hope that it will contribute to the sustainable and efficient production of food in China.

2. Data preparation and analysis

2.1. Dataset overview

The dataset used in this study was obtained from Kaggle and was created by data scientist Atharva Ingle. It has received widespread attention, with 107983 views and 13563 downloads, indicating that it is a reliable and effective resource. The dataset contains 22 labels, representing important common crops such as rice, maize, apple, orange, and coffee. Each label has 100 lines of information, with each line containing 7 numerical features (temperature, humidity, pH value, rainfall, and levels of N, P, and K in the soil) and an output label indicating the optimal crop for that particular environment.

2.2. Advantages of the dataset

There are several advantages of our dataset in terms of the machine learning task for crop recommendation.

First, the dataset has received widespread attention and has been extensively discussed and contributed to by a large community of data scientists. This suggests that it is a reliable and well-vetted resource, which is particularly important when using machine learning algorithms that are sensitive to the quality and integrity of the data.

Second, the dataset contains 22 labels, representing a diverse range of crops that are commonly grown in different parts of the world. This allows our algorithms to learn from a wide variety of crops and environments, making our approach more robust and applicable to a wider range of situations.

Third, the dataset is well balanced, with 100 lines of information for each label. This ensures that our algorithms are able to learn from a representative sample of data for each crop, rather than being biased towards certain crops.

Finally, the dataset contains a large number of data points, with 100 lines of information per label. This allows our algorithms to learn from a sufficient number of examples to make accurate predictions.

Overall, we believe that the use of the Kaggle dataset in our machine learning task for crop recommendation is a major advantage, as it provides a reliable, diverse, and well-balanced dataset that allows our algorithms to learn and make accurate predictions.

2.3. Preparation

In our data preparation process, we took several steps to ensure that the dataset was ready for use in our machine learning algorithms.

First, we detected and eliminated outliers in the data. To do this, we used the interquartile range method, which involves calculating the first and third quartiles (Q1 and Q3) of the data and determining the range of values that fall within 1.5 times the interquartile range (IQR) from Q1 and Q3. Any values that fall outside this range are considered to be outliers and are removed from the dataset. This method is useful for identifying and removing extreme values that may skew the results or lead to inaccurate predictions. Figure 1 demonstrates how IQR range corresponds to normal distribution in statistics.

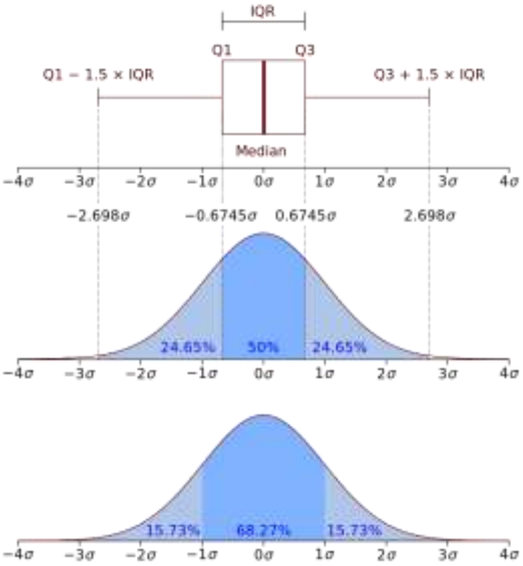

Figure 1. How IQR range corresponds to normal distribution in statistics.

Next, we separated the data into two groups for training and testing respectively. We used a ratio of 80% data for training and 20% data for test. We use the training set to train the machine learning algorithms, and use the test set to evaluate the performance of our approaches on unseen data. This allowed us to gauge the generalization ability of the algorithms and ensure that they were able to make accurate predictions on a range of data.

In addition to these steps, we also performed data visualization to better understand the characteristics of the dataset and identify any patterns or trends that might be relevant to our task. This included generating plots and charts to visualize the relationships between different features and the corresponding labels. Overall, the data preparation process was an important step in ensuring that our algorithms were able to effectively learn from the data and make accurate predictions.

2.4. *Further analysis and visualization*

As part of our data visualization process, we generated several graphs and charts to better understand the characteristics of the dataset and identify any patterns or trends that might be relevant to our task.

One of the graphs we generated was a ring diagram comparing the levels of nitrogen, phosphorus, and potash in different crops (see Figure 2). This graph provided a visual representation of the nutrient content of each crop, allowing us to see at a glance which crops were high in certain nutrients and which were low. This was useful for identifying crops that might be particularly well-suited for certain environments based on the nutrient content of the soil.

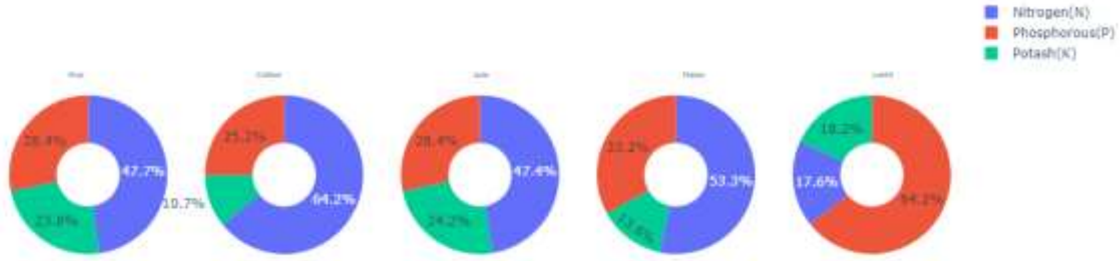

Figure 2. levels of nitrogen, phosphorus, and potash in different crops.

We also generated a 2D scatter plot to visualize the distribution of crops by temperature and humidity (see Figure 3). This graph allowed us to see how different crops tended to be grown in different temperature and humidity ranges, and helped us to understand the relationships between these features and the corresponding labels.

The x-axis of the scatter plot represented temperature, and the y-axis represented humidity. Each data point corresponded to a particular environmental condition, and the color of the point indicated the best crop to cultivate in that environment. For example, the plot showed that rice tended to be grown in environments with high humidity, while maize tended to be grown in environments with low humidity.

We selected temperature and humidity as the features to plot on the 2D scatter plot because they are common and representative features that are easy to understand and interpret. In theory, we could have used any other two features in place of temperature and humidity, but we felt that these two features provided a clear and intuitive visualization of the data. Overall, the 2D scatter plot was a useful tool for visualizing the relationships between the features and labels in the dataset.

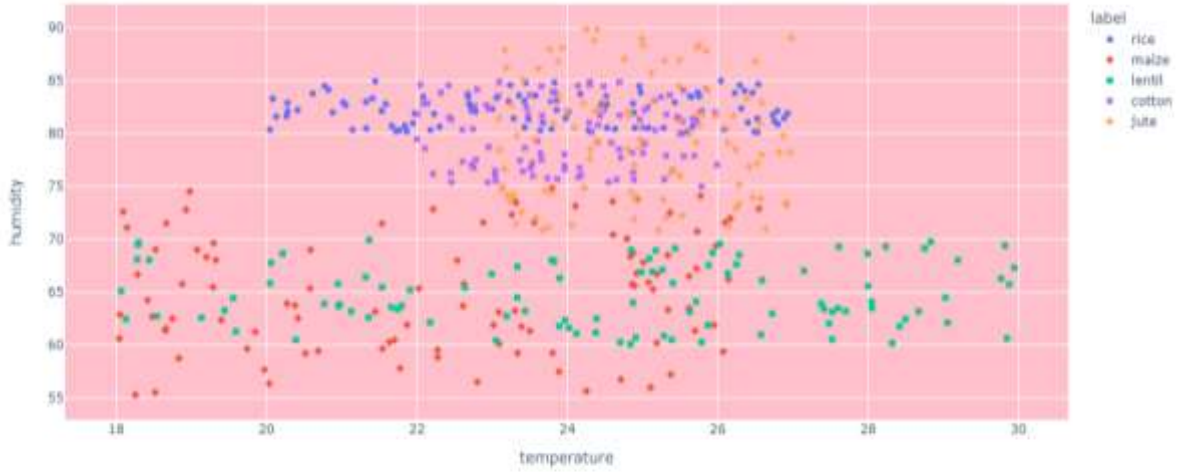

Figure 3. a 2D scatter plots demonstrating the crops' distribution by temperature and humidity.

The third graph we generated as part of our data visualization process was a heatmap showing the correlations between different features in the dataset (see Figure 4). This graph allowed us to see which features were most closely related to each other and helped us to understand the relationships between the features.

The heatmap used color to represent the strength of the correlation between different features. The darker the color, the higher the correlation. As expected, each feature had the greatest correlation with itself (a correlation of 1). Overall, we found that the different features in the dataset were generally not highly correlated with each other. This was good news for our machine learning task, as it meant that the features were providing unique and independent information that could be used by the algorithms to make predictions. If the features were highly correlated, it could have led to redundancy in the data and reduced the accuracy of the algorithms.

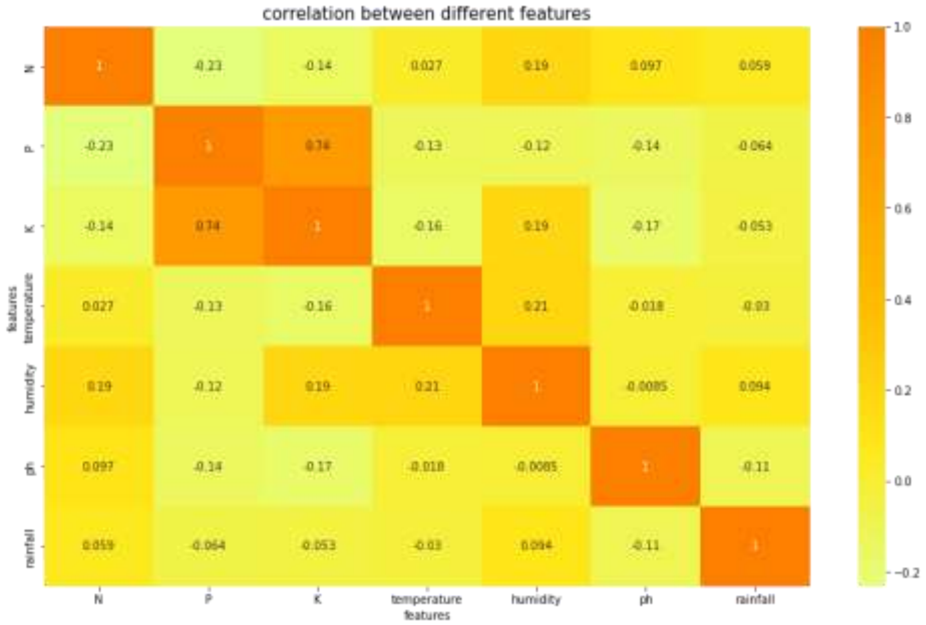

Figure 4. the correlation matrix.

3. Methods

3.1. Training algorithms

3.1.1. Naive Bayes. The principle behind naive Bayes is Bayes' theorem, which states that the probability of an event occurring ($P(A)$) can be calculated based on the probability of certain conditions being true ($P(B)$) and the probability of the event occurring given those conditions ($P(A \mid B)$) [3]:

$$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A})\cdot \mathrm{P}(\mathrm{A})}{\mathrm{P}(\mathrm{B})} \tag{1}$$

In the context of our task, the event is the prediction of the optimal crop for a given location ($\mathrm{P}(\mathrm{A})$), and the conditions are the environmental features of the location ($\mathrm{P}(\mathrm{B})$). Naive Bayes makes predictions by calculating the probability of each label (crop) based on the values of the features ($\mathrm{P}(\mathrm{A} \mid \mathrm{B})$) and choosing the label with the highest probability.

One advantage of naive Bayes is that it is relatively simple to implement and can be trained quickly, even on large datasets. It is also relatively robust to the presence of irrelevant features, meaning that it can still make accurate predictions even if some of the features are not directly relevant to the task. However, it can be less accurate than other algorithms in certain situations, particularly when the features are highly correlated or when the data is noisy or unbalanced.

Overall, naive Bayes is a useful algorithm for our crop recommendation task, and we found that it performed well on our dataset.

3.1.2. Logistic regression. Logistic regression is a type of regression analysis that is used to predict the probability of an event occurring based on the values of one or more features. In our task, the event is the prediction of the optimal crop for a given location, and the features are the environmental conditions of the location. Logistic regression makes predictions by fitting a logistic curve to the data and estimating the probability of each label (crop) based on the values of the features.

Unlike traditional linear regression, which is used to predict continuous values, logistic regression is used to predict binary outcomes (e.g. "yes" or "no"). However, it can also be extended to multi-class classification tasks like ours, where there are more than two possible labels [4]. In this case, logistic regression estimates the probability of each label and chooses the label with the highest probability.

Mathematically, logistic regression estimates the probability of a label using the following formula:

$$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{2}$$

where $\mathrm{P}(\mathrm{A} \mid \mathrm{B})$ is the probability of the label, $x_1, \ldots, x_n$ are the values of the features, and $\beta_0, \ldots, \beta_n$ are coefficients that are learned from the data.

Overall, logistic regression is a useful algorithm for our crop recommendation task, and we found that it performed well on our dataset. Its ability to handle multi-class classification tasks and estimate probabilities makes it a valuable tool for many classification tasks. However, it is important to consider its limitations and evaluate its performance carefully to ensure that it is the best choice for a given task.

3.1.3. Support Vector Machine. SVM works by finding the hyperplane in a high-dimensional space that maximally separates the different classes (labels) in the data. In our task, each label corresponds to a different crop, and the environmental features of the location are used to determine the position of each data point in the high-dimensional space. The SVM algorithm then finds the hyperplane that maximally separates the different crops, such that the data points of each crop are as far as possible from the hyperplane [5].

Mathematically, the SVM algorithm seeks to find the hyperplane that maximizes the margin between the different classes in the data. This is done by minimizing the following objective function:

$$\min_{w,b} \left\{\frac{1}{2}|w|^2\right\} \tag{3}$$

subject to

$$y_i(w \cdot x_i + b) \geq 1, i = 1, \ldots, n \tag{4}$$

where w and b are the parameters of the hyperplane, x_i is the ith data point, y_i is the label of the ith data point, and n is the total number of data points. The first term in the objective function represents the distance from the hyperplane to the closest data points (the margin), and the second term represents the complexity of the hyperplane.

Overall, SVM is a powerful algorithm for our crop recommendation task, and we found that it performed well on our dataset. Its ability to find the hyperplane that maximally separates the different classes makes it a valuable tool for many classification tasks. However, it can be sensitive to the choice of kernel function and the values of the hyperparameters, and it may require careful tuning to achieve good performance.

3.1.4. Decision tree. A decision tree is a tree-like model that makes predictions by following a series of decisions based on the values of the features. Each decision point in the tree splits the data into two or more branches based on the value of a feature, and the final prediction is made based on the path taken through the tree. Decision trees are easy to interpret and can handle both continuous and categorical data, but they can be prone to overfitting if the tree becomes too complex.

3.1.5. Random forest. A random forest is an ensemble learning method that combines the predictions of multiple decision trees to make a final prediction [6]. Each tree in the random forest is trained on a randomly selected subset of the data, and the final prediction is made by aggregating the predictions of all the trees. This process helps to reduce the variance of the model and improve the overall accuracy of the predictions. In addition, the use of multiple decision trees helps to mitigate the problem of overfitting, which can occur when a single decision tree is overly complex and starts to capture noise in the data rather than the underlying patterns.

In the context of our crop recommendation task, the decision trees in the random forest were trained to classify crops based on the environmental features of a given location. The trees made predictions by following a series of decisions based on the values of the features, and the final prediction was made by aggregating the predictions of all the trees. This allowed us to make accurate and reliable predictions about the optimal crops for different locations based on their environmental conditions.

3.1.6. Neural network. A neural network is a machine learning model that is inspired by the structure and function of the human brain. It consists of multiple layers of interconnected "neurons" that process and transmit information.

In the context of our task, the neural network was trained to predict the optimal crop for a given location based on the values of the environmental features. The input layer of the network received the values of the features as input, and the output layer produced the prediction of the optimal crop. The intermediate layers, known as hidden layers, transformed the input into a suitable representation for the output layer to use.

Mathematically, the output of a neural network can be represented as follows:

$$\hat{y} = \mathrm{f}(\mathrm{W}_1 \cdot \mathrm{f}(\mathrm{W}_2 \cdot \ldots \cdot \mathrm{f}(\mathrm{W}_\mathrm{n} \cdot \mathrm{x} + \mathrm{b}_\mathrm{n}) + \mathrm{b}_{\mathrm{n}-1}) + \cdots + \mathrm{b}_1) \tag{5}$$

where $\hat{y}$ is the prediction of the network, x is the input, $\mathrm{W}_1, \ldots, \mathrm{W}_\mathrm{n}$ are the weights of the connections between the layers, and $\mathrm{b}_1, \ldots, \mathrm{b}_\mathrm{n}$ are the biases of the neurons in the layers. The function f is an activation function, which is used to introduce nonlinearity into the model.

Overall, the neural network approach was a powerful and effective tool for our crop recommendation task, and we found that it performed exceptionally well on our dataset. Its ability to learn complex patterns in the data and make accurate predictions made it a valuable tool for this task. In addition, the use of hidden layers and activation functions allowed the network to capture nonlinear relationships in the data, which is important for many real-world classification tasks.

However, it is important to note that neural networks can be sensitive to the choice of hyperparameters and the amount of data available for training. It is also important to consider the computational cost of training and using neural networks, as they can be quite resource-intensive. Despite these challenges, we found that the neural network approach was a very promising approach for our crop recommendation task, and it has the potential to be a valuable tool for many other classification tasks as well.

3.2. Evaluation metrics

In order to evaluate the performance of the different machine learning algorithms that we used for our crop recommendation task, we used a number of different evaluation metrics. These metrics allowed us to quantitatively assess the accuracy, precision, and recall of the different algorithms and compare their performance.

One of the evaluation metrics that we used was accuracy, which measures the proportion of correct predictions made by the model. In our task, this was calculated as the number of correct crop recommendations divided by the total number of crop recommendations made. Mathematically, this can be represented as:

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{6}$$

where TP is the number of true positive predictions (correct recommendations that were actually optimal), TN is the number of true negative predictions (incorrect recommendations that were not optimal), FP is the number of false positive predictions (incorrect recommendations that were not optimal), and FN is the number of false negative predictions (correct recommendations that were not optimal).

Another metric that we used was precision, which measures the proportion of true positive predictions made by the model. In our task, this was calculated as the number of correctly recommended crops that were actually optimal for the given location, divided by the total number of crops that were recommended. Mathematically, this can be represented as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

Finally, we also used the metric of recall, which measures the proportion of true positives that were correctly identified by the model. In our task, this was calculated as the number of correctly recommended crops that were actually optimal for the given location, divided by the total number of optimal crops in the dataset. Mathematically, this can be represented as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

Overall, these evaluation metrics allowed us to quantitatively assess the performance of the different machine learning algorithms that we used for our crop recommendation task and compare their results. In the next section, we will present the specific results that we obtained using these metrics, and we will discuss the implications of these results for the effectiveness of the different algorithms in our task. It is important to note that no single evaluation metric is sufficient to fully characterize the performance of a machine learning model, and it is necessary to consider multiple metrics in order to gain a complete understanding of the model’s strengths and weaknesses.

4. Experiment and results

4.1. General results

In this section, we present the specific results of our experiments with the different machine learning algorithms that we used for our crop recommendation task.

First, we evaluated the performance of the naive Bayes approach. This algorithm achieved an accuracy of 97.5%, a precision of 97.5%, and a recall of 96.2%. These results indicate that the naive Bayes approach was able to make highly accurate and precise predictions about the optimal crops for different locations, and it was able to correctly identify a high proportion of the optimal crops in the dataset.

Next, we evaluated the performance of the logistic regression approach. This algorithm achieved an accuracy of 93.2%, a precision of 93.5%, and a recall of 93.3%. These results suggest that the logistic regression approach was able to make relatively accurate and precise predictions, but it was slightly less effective than the naive Bayes approach in terms of overall accuracy and recall.

The SVM approach is special, because it initially only achieved an accuracy of 10.7%, significantly lower than the other methods. However, after we adjusted the parameters of the algorithm, its accuracy improved to 95.3%. The precision and recall also improved from around 10% to 95%. These results demonstrate the sensitivity of the SVM approach to the choice of hyperparameters, and they highlight the importance of careful tuning in order to achieve good performance.

We also evaluated the performance of the decision tree approach, which resulted in an accuracy of 88.0%, a precision of 84%, and a recall of 88.0%. These results indicate that the decision tree approach was able to make relatively accurate predictions, but it was less precise and less effective at identifying the optimal crops in the dataset compared to the other algorithms.

The random forest approach performed better than the decision tree approach, with an accuracy of 97.5%, a precision of 98.1%, and a recall of 97.0%. These results suggest that the random forest algorithm was able to make highly accurate and precise predictions about the optimal crops for different locations, and it was able to correctly identify a high proportion of the optimal crops in the dataset. It is worth noting that the random forest approach took significantly longer time to train and run than the other algorithms, which may be a drawback in certain situations.

Finally, we evaluated the performance of the neural network approach, which performed the best out of all the algorithms that we tested. It achieved an accuracy of 99.3%, a precision of 99.2%, and a recall of 99.2%. These results indicate that the neural network was able to make highly accurate and precise predictions about the optimal crops for different locations, and it was able to correctly identify a very high proportion of the optimal crops in the dataset.

Overall, these results demonstrate the effectiveness of the different machine learning algorithms that we used for our crop recommendation task. The neural network approach performed the best, followed by the random forest and naive Bayes approaches. The logistic regression and SVM approaches also performed well, but were slightly less effective than the top three algorithms. The decision tree approach was the least effective of the algorithms that we tested. These results indicate that the neural network approach is a particularly promising tool for this task, and it has the potential to be a valuable tool for many other classification tasks as well.

4.2. Improving support vector machine

During our experiments with the SVM algorithm, we found that the initial performance of the algorithm was relatively poor, with an accuracy of only 10.7%. In order to improve the performance of the algorithm, we adjusted several of the parameters that are commonly used to tune the behavior of the SVM algorithm. These parameters included the regularization parameter C, the kernel type and kernel coefficient γ, the degree of the polynomial kernel, and the choice of kernel function [7].

As shown in Figure 5, we found through our experimentation that the choice of kernel function had the greatest impact on the performance of the SVM algorithm. The sigmoid kernel performed the worst, followed by the radial basis function (RBF) kernel. Both the linear kernel and the polynomial kernel provided satisfactory results. This is likely because the sigmoid and RBF kernels are more sensitive to the choice of the kernel coefficient γ, which can be difficult to optimize. On the other hand, the linear and polynomial kernels are less sensitive to the choice of γ, and they are generally more stable and easier to tune.

We also found that the choice of γ had some impact on the performance of the SVM algorithm, but it was generally less influential than the choice of kernel function. In particular, we did not observe a clear trend in the results, with some values of γ resulting in better performance and others resulting in worse performance. This highlights the importance of careful experimentation and parameter tuning in order to achieve good performance with the SVM algorithm.

Overall, our work on adjusting the parameters of the SVM algorithm was successful in improving the performance of the algorithm for our crop recommendation task. By carefully tuning the parameters of the SVM algorithm, we were able to achieve comparable results to other machine learning algorithms, and we demonstrated the potential of the SVM algorithm as a valuable tool for this task.

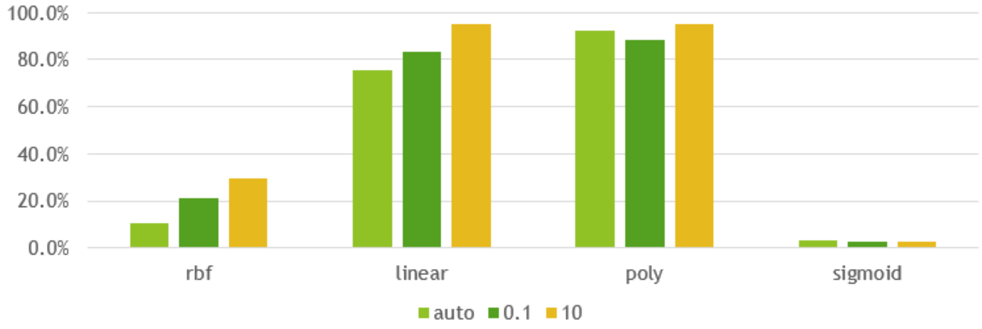

Figure 5. the performance of SVM under different parameters.

5. Conclusion

In this paper, we presented a machine learning approach to the problem of crop recommendation. Given a set of numerical features about an environment, our approach was able to automatically recommend the best crop to grow in that location. We used a dataset containing information about different crops and their optimal growing conditions, and we applied a number of different machine learning algorithms to this dataset in order to make predictions about the optimal crops for different locations.

Our results demonstrated the effectiveness of several different machine learning algorithms for this task, including naive Bayes, logistic regression, SVM, decision tree, random forest, and neural network.

Of these algorithms, the neural network approach performed the best, with an accuracy of 99.3%, a precision of 99.2%, and a recall of 99.2%. This indicates that the neural network approach is a particularly promising tool for this task, and it has the potential to be a valuable tool for many other classification tasks as well.

The results of our work have important implications for the field of agriculture and the environment. By automating the process of crop recommendation, our approach has the potential to improve the efficiency and productivity of agricultural operations, and it could help to reduce the negative impacts of agriculture on the environment. In the future, we plan to expand our dataset and continue to refine our machine learning algorithms in order to further improve the accuracy and reliability of our crop recommendation system. We also plan to investigate other applications of machine learning in agriculture, including the prediction of crop yields and the optimization of irrigation and fertilization practices. Overall, our work represents a significant step forward in the use of machine learning to address important challenges in agriculture and the environment, and we believe it has the potential to make a significant impact on these fields.

References

[1] Jingzhu Zhao, Qishan Luo, Hongbing Deng, and Yan Yan. Opportunities and challenges of sustainable agricultural development in china. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1492):893-904, 2008.

[2] URMIL VERMA et al. Arima and arimax models for sugarcane yield forecasting in northern agro-climatic zone of haryana. Journal of Agrometeorology, 24(2):200-202, 2022.

[3] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. Encyclopedia of machine learning, 15:713-714, 2010.

[4] Maher Maalouf. Logistic regression in data analysis: an overview. International Journal of Data Analysis Techniques and Strategies, 3(3):281-299, 2011.

[5] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. IEEE Intelligent Systems and their applications, 13(4):18-28, 1998.

[6] Gérard Biau and Erwan Scornet. A random forest guided tour. Test, 25(2):197-227, 2016.

[7] Shun-ichi Amari and Si Wu. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12(6):783-dfes 789, 1999.