SCHOLARLY DEBUT
by Scholarly Review

Fall 2025

Application of Machine Learning to Classify Wheat Seeds

By Xinning Zhang

AUTHOR BIO

Xinning Zhang is a high school student who loves coding, drawing, and running. He wishes to explore more about machine learning, deep learning, and AI multi-modal.

ABSTRACT

We identify the type of wheat seeds into three categories: Kama, Rosa, and Canadian, different species of wheat in the real world. To do this, we analyze features including the wheat seed's perimeter, area, compactness, length, width, asymmetry coefficient, and groove length. Then, we transform the variables of the features into machine-readable numbers. With the data being correctly altered, Machine learning methods, including Standard Scaler, train_test_split, prepare the data by normalizing the data units, and split the data into training and testing sets to train and test a model. Finally, Categorical Classification is used to create a model that classifies wheat seed varieties. The five hundred times of iterations and tests of the model resulted in a 97% accuracy and a 0.13 loss, indicating a good performance. This model allows farmers without a huge experience in wheat identification to guess if they are examining Kama, Rosa, and Canadian wheat seeds.

SCHOLARLY DEBUT
by Scholarly Review
Fall 2025

INTRODUCTION

The accurate classification of agricultural products, like wheat seeds, is important for sorting, packaging, and improving crop quality. Our dataset contains morphological features of three wheat varieties: Kama, Rosa, and Canadian, which correspond to type 1, type 2, and type 3 wheat seeds.

This information includes area, perimeter, compactness, length, width, asymmetry coefficient, and groove. The Kama variation is usually the smallest in size, whereas the Canadian variation is the largest. This is shown through the factors like perimeter, the length of the shape; area, the size it occupies; kernel length; and tip-to-tip distance. However, the Canadian wheat seeds are known for their long, narrow grain. Thus, it usually has the least kernel width, distance across the grain, and the highest kernel groove length, distance of the central line that appears on the seed. On the other hand, the Kama wheat seed has the highest compactness, how circular it is, and Kernel width due to its round look. Figures 1 - 3 show the wheat seed varieties.

Figure 1: Kama wheat grains (Istockphoto)

Figure 2: Rosa Grains (Dmcc)

Figure 3: Scout 66 Hard Red Winter Wheat; Canadian Seeds (*Scout)*

Machine learning is the field of computer science that uses data to make predictions and decisions. Specifically, we will use supervised learning to predict new data based on the existing data. We will input training data in a table format with many inputs and the output, the seed variation. The computer will analyze the trend from the given data and create a model for further data to be tested in. By creating a model that identifies the three types of wheat seeds, wheat farmers will be able to use mobile devices to identify the seeds without taking a closer look and relying on experience. Later, this model can potentially be integrated into machinery that allows larger scale classification

DATA

The problem we are trying to solve is identifying wheat seed types. The initial data contains information about the seed we're looking at. This information results in 3 possible outcomes - type 1, 2, and 3-equivalent to Kama, Rosa, and Canadian wheat seeds. A sample of the data, which consists of 200 wheat seeds, is shown in Table 1 below. The table shows two examples from each type.

Area	Perimeter	Compactness	Kernel Length	Kernel Width	Asymmetry Coefficient	Kernel Groove	Type
15.26	14.84	0.8710	5.763	3.312	2.221	5.220	1
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1

17.63	15.98	0.8673	6.191	3.561	4.076	6.06	2
16.84	15.67	0.8623	5.998	3.484	4.675	5.877	2
11.84	13.21	0.8521	5.175	2.836	3.598	5.044	3
12.30	13.34	0.8684	5.243	2.974	5.637	5.063	3

Table 1: Sample dataset (Jmcaro)

Each of our wheat seeds represents a sample data point used to train our machine learning model. Our goal of the machine learning technique is to try to identify the classification of the wheat seeds based on the features in the 200 data points in the dataset.

PREPROCESSING

We first check if any of our values are missing. We start with the first column, the Area, and sum the missing data to produce a value for a given row. Null represents a missing data value. However, if the sum is zero, it means that none of the data values are missing. We repeated the same for all other columns until there was no missing data.

The dataset contains measurements of 200 samples of wheat kernels. Each sample belongs to one of three varieties of wheat: Kama (1), Rosa (2), or Canadian (3). Since all of our variables are numerical, continuous data, this dataset is well-suited for classification tasks due to the clear separation of classes in feature space. However, the outcome numbers are different. The numbers 1, 2, and 3 don't mean a numeral value but a categorical result, simply meaning what type of wheat seed it is. We then need to convert this classification column into binary or true-false outcomes. The type column will turn into 3 columns that are called type 1, 2, and 3, with number 1 for true and 0 for false. These changes are shown in Table 2 below.

Area	Perimeter	Compactness	Kernel Length	Kernel Width	Asymmetry Coefficient	Kernel Groove	Type 1	Type 2	Type 3
15.26	14.84	0.8710	5.763	3.312	2.221	5.220	1	0	0
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1	0	0

17.63	15.98	0.8673	6.191	3.561	4.076	6.06	0	1	0
16.84	15.67	0.8623	5.998	3.484	4.675	5.877	0	1	0
11.84	13.21	0.8521	5.175	2.836	3.598	5.044	0	0	1
12.30	13.34	0.8684	5.243	2.974	5.637	5.063	0	0	1

Table 2: Preprocessed dataset

Now that we have our data ready, we have more steps to prepare for the test. The features in our data were scaled using StandardScaler, which standardizes each feature to have a zero mean and unit variance. This step is crucial for algorithms like Support Vector Machine (SVM) that rely on distances between feature vectors. The standard score, similar to the z score, is calculated by the following, where u is the mean and s is the standard deviation.

$$Z = (x-u)/s$$

We split the data into training and testing sets using a train-test split with an 80:20 ratio, respectively. This ensures that the model is trained on a majority of the data while still being evaluated on unseen data to assess generalization performance.

MACHINE LEARNING

Again, our goal is to categorize the wheat seeds into three different types (Kama, Rosa, and Canadian). Thus, the selected model for our project was the Support Vector Classification. It draws lines that separate the discrete categories, which is also called categorical classification.

The method mod.fit trains the classifier by finding the optimal decision boundary (hyperplane) that best separates the classes in feature space. In this project, we allowed 500 optimization steps, iterations, for the machine to find the hyperplane. It then returns the validation of the data.

After the loss and accuracy are calculated, a table is returned that consists of the 500 iterations and a graph of the iteration vs. loss, as shown in Figure 4 below.

	accuracy	loss	val_accuracy	val_loss
0	0.275168	1.332488	0.28	1.262893
1	0.288591	1.294369	0.30	1.230052
2	0.302013	1.266768	0.30	1.201555
3	0.315436	1.242586	0.32	1.176517
4	0.328859	1.220172	0.32	1.153168
...	...	...	...	...
495	0.966443	0.135163	0.94	0.154248
496	0.966443	0.134997	0.94	0.154175
497	0.966443	0.134917	0.94	0.153965
498	0.966443	0.134698	0.94	0.153834
499	0.966443	0.134566	0.94	0.153641

Figure 4: Iteration vs. Loss

The accuracy and loss on the left side represent the calculation regarding the training data while the val_accuracy and val_loss represent the validation data, which means testing of new data.

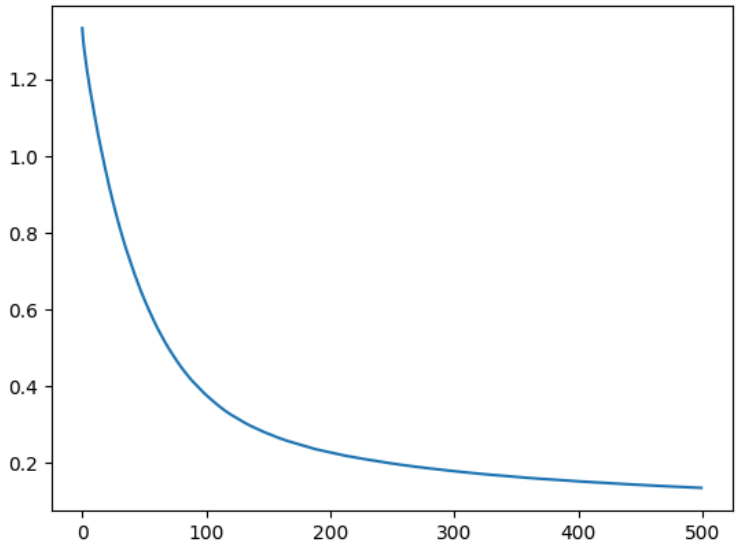

Figure 5: Graph of Iteration vs. Loss

The graph, in Figure 5 above, shows that as the iteration goes closer towards 500, the loss of the data goes closer to 0.13.

RESULTS

The result from Figure 4 shows outputs about the model created from the 200 data values we input. We see an increasing trend for accuracy and a decreasing trend for loss as the iteration continues to 500.

The higher the accuracy and the lower the loss, the better the performance of the model is. Accuracy tells us the percentage of seeds the model classified correctly out of the 200 seeds we input above. The accuracy of 0.97 or 97% from training data and 0.94 or 94% from the validation data shows a good performance of the model. Loss is a score that measures how far off the model's predictions are from the true labels. It is a numerical value calculated by categorical cross-entropy that the model tries to minimize. A value of 0.13 from the training data and a value of 0.15 from the validation data show a fair confidence and prediction of the data. Overall, the results indicate that the model is doing very well.

However, both the training data, normal accuracy and loss, are higher than the validation data, accuracy and loss with "val". This difference between the training and validation data shows that the data is slightly overfitting. It happens when the model learns the training data too well, meaning it's memorizing the data instead of generalizing from it. It is like a student memorizing the exam word for word instead of applying knowledge. To fix it, we would have to drop hidden layers (not directly visible layers that enhance learning through functions, improving the prediction accuracy), but this will lower the accuracy values. Since the model already performs fairly good, the issue of overfitting can be neglected.

CONCLUSION

With the goal of classifying each seed into Kama, Rosa, and Canadian wheat varieties, we used 200 data points that consist of seven features and the variety. The varieties were assigned to 3 outcomes and given a true or false indicator. Then the data was preprocessed using two steps. The features of the data were standardized by Standard Scalar, and the dataset was split into training and testing sets.

A Support Vector Machine classifier was used to train the model. It finds the best boundary that separates the seed types. The model achieved high training, test accuracy, and low loss, indicating strong performance. Though some overfitting was observed, training and validation accuracy were both high.

Thus, the model successfully classified wheat seeds with high accuracy, showing that SVMs are effective for small and well-separated datasets. By using careful preprocessing and a strong but simple model (SVM), excellent classification performance was achieved on a small but real-world dataset, highlighting the power of classic machine learning techniques for structured data problems.

REFERENCES

jmcaro. (2018, October 23). *Machine Learning Classifiers - Wheat Seeds*. Kaggle.com; Kaggle. https://www.kaggle.com/code/jmcaro/machine-learning-classifiers-wheat-seeds/input (2025).

Istockphoto.com. https://www.istockphoto.com/photos/kamut-wheat

Dmcc, R. G. (2016). *Rosa Grains Dmcc(Distributors & Wholesalers) in Jumeirah Lake Towers (Al Thanyah 5), Dubai - HiDubai*. Hidubai.com. https://www.hidubai.com/businesses/rosa-grains-dmcc-b2b-services-distributors-wholesalers-jumeirah-lake-towers-al-thanyah-5-dubai

Scout 66 Hard Red Winter Wheat. (2023). Canadianvalleyseeds.com. https://www.canadianvalleyseeds.com/product/scout-66-hard-red-winter-wheat/