ISSN:2320-0790

An International Journal of Advanced Computer Technology

COMPARISONS OF CLASSIFICATION ALGORITHMS ON SEEDS DATASET USING MACHINE LEARNING ALGORITHM

Divya Agrawal[1], Prof. Priyanka Dahiya[2]

[1]Department of Computer Science and Engineering,
Mody University College Of Engineering and Technology, Lachmangarh, Rajasthan
divagrwl1994@gmail.com[1], priyankadahiya@gmail.com[2]

Abstract: In history, agriculture has been the backbone of the economy. These agricultural activities remain undeveloped due to different factors. The current seed classification analysis is inefficient and has no validation mechanism. In this research, we have made an effort to present a predictive model to predict seed classes using machine learning algorithms. For this development, machine learning algorithm is used to learn from data which can be used to make predictions, to make real-world simulations. The developed model is experimented using seed dataset and then seed classes are predicted using the developed model. . The main machine learning methods used in this research is Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Neighbors Classifier (KNN), Decision Tree Classifier (CART), Gaussian NB (NB), and Support Vector Machine (SVM). This is a good combination of simple linear (LR and LDA), nonlinear (KNN, CART, NB and SVM) algorithm. In this research we are trying to present distinctive machine learning approaches, for classification of various seeds which provide opportunity to individuals or agriculturist to identify various seeds.

Keywords: Classification, CART, KNN, LDA, LR, Machine Learning, NB, Predictive Analysis, Seed Classification, SVM

I. INTRODUCTION

Agriculture is the most important economic sector of many developing countries. Most of the activities are done with a lack of modern technology. Currently, seed classification is done based on knowledge of human being. The purification of seed material becomes a significant part in this development and needs to be improved. Specifying the nature of wheat manually requires a specialist judgment and is time consuming. Sometimes the assortment of seeds looks so comparable that differentiating them becomes a very difficult task when carried out manually. To overcome this issue, machine algorithms can work to classify seeds as indicated by its quality. Machine Learning is widely used in the field of agriculture for differentiating the varieties of various crops and for identifying their quality as well.

The present paper deals with finding the accuracy of Seeds data. For doing this, six machine learning algorithms are used, that is, Logistic Regression (LR), Linear Discriminant Analysis (LDA),K Neighbors Classifier (KNN), Decision Tree Classifier (CART), Gaussian NB (NB), Support Vector Machine (SVM).

II. RELATED WORK

Studies have been done on classification of seeds using machine algorithms. These research have used different machine classifiers and have performed accuracy for carrying out their work.

A study by [1] used Discriminant Analysis and K Nearest Neighbor for classification of wheat and barley grain kernels. The framework preparing was performed with only morphological features, only color features and combination of morphological features, color features as well as textural features. It was reasoned that accuracies higher than 99% can be accomplished when morphological, color and textural feature types are used together as compared to using them alone.

This paper [2] presents the capability and potential of machine vision with the well- trained multilayer neural network classifiers for shapes, sizes, and varietal type. In order to classify the seeds, they used Weka classification tools; Function, Bayes, Meta and Lazy methods. Classifiers they used from these methods are Logistics, SMO, Naïve Bayes Updateable, Multilayer Perceptron, Naïve Bayes, Bayes Net and Classifier Multi Class. The Multilayer Perceptron classifier that gives the most accuracy value 97.6% using 5 fold Cross Validation. This exploration is concluded as the unsupervised artificial neural network gives better execution with 79% accuracy as compared to the supervised artificial neural networks which give 73% accuracy.

According to this paper [3], the researchers trying to use K-means clustering algorithm and the default Euclidean distance metric to cluster seed dataset. For this clustering, the researcher uses MATLAB as a programming environment. K-means function is used from statistics toolbox which is given two arguments. Those two arguments are the dataset and the number of the cluster the data going to be classified. In this study, the authors propose the system which is capable of clustering approximately seeds and the profiting K-means algorithm leads to the operation.

This Research [4], states Neural Networks to classify varieties of rice which contain a total of 9 different rice verities. To classify these varieties the authors uses image acquisition of seeds. They also developed to extract thirteen morphological features, six color features and fifteen texture features from color images of individual seed samples. Results of the paper is just designing and developing neural network models with two hidden layers in all networks using Matlab toolbox.

These study[5], put forwards crop prediction by identifying various parameter like the type of soil, PH, phosphate, potassium, calcium, nitrogen, magnesium, sulfur, manganese, copper, iron, organic carbon depth, temperature, rainfall, humidity and parameter associated to the atmosphere. The authors design a network which correctly learns associations of effective climatic factors on crop yield, it can be used to estimate crop production in long or short term. This paper shows the ability of artificial neural network technology for the approximation and prediction of crops. In this paper, we shall examine one of the most common neural network architectures. They analyze the result by using feed forward back propagation ANN model for each area and finds the most effective factors on crop yield.

III. PYTHON LIBRARIES

This Paper used Python version 3.6. For implementing this research you need to install this are 5 key libraries. Below is a list of the Python libraries which is used in this Research

- scipy
- numpy
- matplotlib
- pandas
- sklearn

IV. METHODOLOGY

The Methodology system workflow is shown in above figure 1. It uses different steps which are as follows:

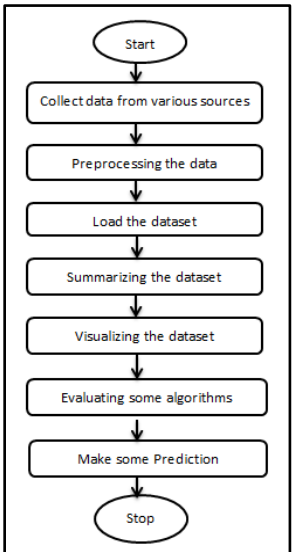

Fig 1: Work Flow of System

4.1. Preprocessing of data

Data collection, data pre-processing (data cleaning, attribute selection, data formatting and transformation, dimensionality reduction and the like) are the most important activities under data preparation, which finally resulted in creating target data set. After data collection, data pre-processing procedures were conducted to train more efficiently and then coordinate systems of the all the data were transformed to the same coordinate system so that all the data fit the model.

4.2. Summarizing the dataset

Data of seeds can be collected from various sources. In this research we have collected seeds data fromkaggle. Data is stored in Excel file in .csv file format. Here, pandas library is used for load the data. Pandas are also used to explore the data both with data visualization and descriptive statistics.

Now, we can take an overlook of data in few more ways:

4.2.1. Dimensions of Dataset

A Dimension is used to get quick idea of how many instances (rows) and how many attributes (columns) the data contains. Shape property is used to check the Dimensions. In this research there are 210 rows and eight attributes that is Area, Perimeter, Compactness, Length, Width, Asymmetry, groove, and Class.

```
>>> print(dataset.shape)
(210, 8)
```

Fig 2: Shape of Data

4.2.2. Peek at the data itself

It is always a good idea to really eyeball the information. In this research Area, Perimeter, Compactness, Length, Width, Asymmetry, groove, and Class are parameters.

Here, Figure 3 shows the top five format of data.

```
>>> print(dataset.head(5))
     area  perimeter  compact  length  width  asymmetry  groove  class
0   15.26      14.84   0.8710   5.763  3.312     2.2210   5.220      1
1   14.88      14.57   0.8811   5.554  3.333     1.0180   4.956      1
2   14.29      14.09   0.9050   5.291  3.337     2.6990   4.825      1
3   13.84      13.94   0.8955   5.324  3.379     2.2590   4.805      1
4   16.14      14.99   0.9034   5.658  3.562     1.3550   5.175      1
5   14.38      14.21   0.8951   5.386  3.312     2.4620   4.956      1
```

Fig 3: Head Dataset

Figure 4, shows the lowest five data values in our dataset. Generally we check head and tail of data.

```
>>> print(dataset.tail(5))
      area  perimeter  compact  length  width  asymmetry  groove  class
205  12.19      13.20   0.8783   5.137  2.981      3.631   4.870      3
206  11.23      12.88   0.8511   5.140  2.795      4.325   5.003      3
207  13.20      13.66   0.8883   5.236  3.232      8.315   5.056      3
208  11.84      13.21   0.8521   5.175  2.836      3.598   5.044      3
209  12.30      13.34   0.8684   5.243  2.974      5.637   5.063      3
```

Fig 4: Tail Dataset

4.2.3. Statistical summary of all attributes

Statistical summary includes the count, mean, std, the min and max values as well as many percentiles values.

```
>>> print(dataset.describe())
             area  perimeter  compact   length    width
count  210.000000  210.000000  210.000  210.000  210.000
mean    14.848000   14.559286    0.871    5.629    3.259
std      2.910000    1.305959    0.024    0.443    0.378
min     10.590000   12.410000    0.808    4.899    2.630
25%     12.270000   13.450000    0.857    5.262    2.944
50%     14.355000   14.320000    0.873    5.524    3.237
75%     17.305000   15.715000    0.888    5.980    3.562
max     21.180000   17.250000    0.918    6.675    4.033

       asymmetry      groove      class
count    210.000  210.000000  210.000000
mean       3.700    5.408071    2.000000
std        1.503    0.490480    0.818448
min        0.765    4.519000    1.000000
25%        2.561    5.045000    1.000000
50%        3.599    5.045000    2.000000
75%        4.769    5.877000    3.000000
max        8.456    6.550000    3.000000
```

Fig 5: Descriptions of Data

4.2.4. Class Distribution

In class Distribution is to understand that number of instances (rows) that belong to each class. It can view as an absolute count.

In figure 6 observed that each class has same number of instances that is 70 or 33% of the dataset.

```
>>> print(dataset.groupby('class').size())
class
1    70
2    70
3    70
```

Fig 6: Class Distribution

V. MACHINE LEARNING ALGORITHM

5.1. Logistic Regression(LR)

Logistic regression was produced by statistician David Cox in 1958. Logistic regression is is utilized as a part of different fields, including machine learning, most medical fields, and social sciences. LR is a statistical technique for analyzing a dataset in which there are at least one or more independent variables that decide an outcome [5].

5.2. Linear Discriminant Analysis(LDA)

Ronald A. Fisher planed the Linear *Discriminant* in 1936. LDA is "supervised" and processes the directions ("linear discriminants") that will represent the axes that maximize the separation between various classes[6].

5.3. Support Vector Machine (SVM)

SVM algorithm was develop by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. Support Vector Machine is a supervised machine learning algorithm which can be utilized for both classification and regression challenges. However, it

is most utilized in classification problems. SVM is a discriminative classifier formally defined by a separating hyperplane[7].

VI. DATA VISUALIZATION

Visualization is for interactions between the variables. Visualization have two basic type to plot Univariate and scatter plots. The above figure 2 shows the scatter plots of all pairs and attributes.

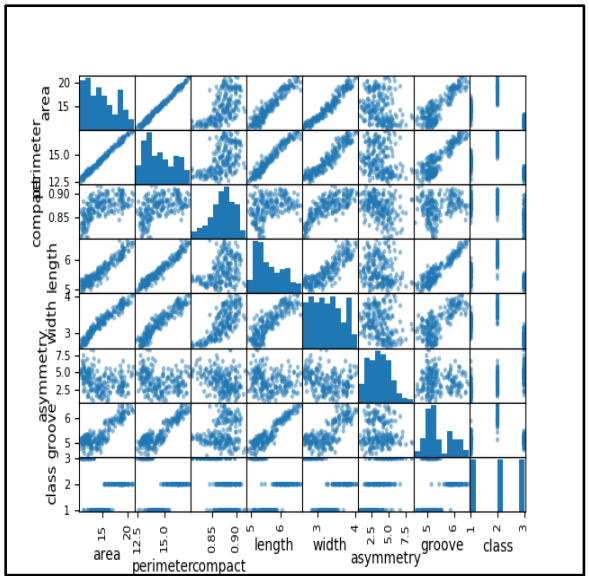

Fig 7: Scatter plot matrix

In scatter plot input variables structured relationship can spot easily. The diagonal groupings are some pairs of attributes. This suggests a high correlation and a predictable relationship.

VII. RESULTS

7.1. Build the Model

This is a good combination of simple linear (LR and LDA), nonlinear (KNN, CART, NB and SVM) algorithms. It ensures that the results are directly comparable.

```
>>> models = []
>>> models.append(('LR', LogisticRegression()))
>>> models.append(('LDA', LinearDiscriminantAnalysis()))
>>> models.append(('KNN', KNeighborsClassifier()))
>>> models.append(('CART', DecisionTreeClassifier()))
>>> models.append(('NB', GaussianNB()))
>>> models.append(('SVM', SVC()))
>>> results = []
>>> names = []
>>> for name, model in models:
        kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
        cv_results = cross_validation.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
        results.append(cv_results)
        names.append(name)
        msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
        print(msg)
```

Fig 8: Build the Model

7.2. Select the Best Model

We have 6 models and accuracy estimations for each. Next is to compare the models to each algorithm and select the most accurate. Running the code of figure 8 we get the following results:

```
LR: 0.916912 (0.047658)
LDA: 0.958456 (0.037853)
KNN: 0.875000 (0.049680)
CART: 0.868382 (0.065122)
NB: 0.880515 (0.060035)
SVM: 0.887132 (0.067951)
```

Fig 9: Accuracy of Algorithm

7.3. Plot the Model

We have 6 models and accuracy estimations for each algorithm. Presently, comparing each model to other and select the most exact. After implementing the algorithms and the results of their evaluated performance were obtained. The outcome demonstrate that the accuracy of Logistic Regression (LR) comes out to be 91.6%, Linear Discriminant Analysis (LDA) comes out to be 95.8%, K Neighbors Classifier (KNN) gives 87.5%, Decision Tree Classifier (CART) gives 88%, Gaussian NB (NB) comes to be 88.05%, and Support Vector Machine (SVM) turns out to be 88.71%. This implies that Linear Discriminant Analysis (LDA) Performed superior than all the algorithms.

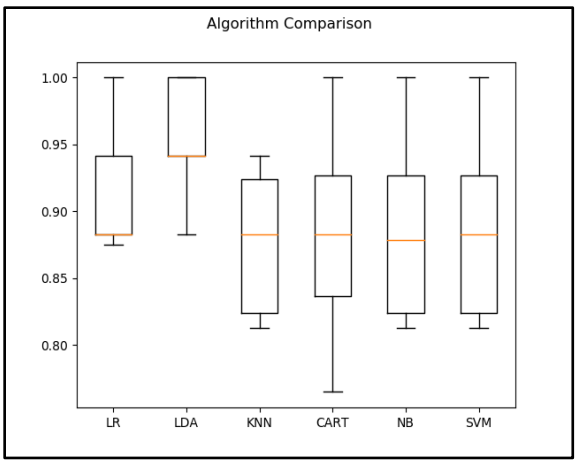

Fig 10: Algorithm Comparison

Now, by using concept of pipelines we have implemented the algorithms and the results of their evaluated performance were obtained. The outcome show that accuracy of Logistic Regression (LR) comes out to be 92.86%, Linear Discriminant Analysis (LDA) comes out to be 95.8%, K Neighbors Classifier (KNN) gives 90.4%, Decision Tree Classifier (CART) gives 88%, Gaussian NB (NB) comes to be 88.05%, and Support Vector Machine (SVM) comes out to be 92.83%. As compared to previous fig we got more accuracy by pipeline. This means that Logistic Regression (LR) Performed better than all the algorithms.

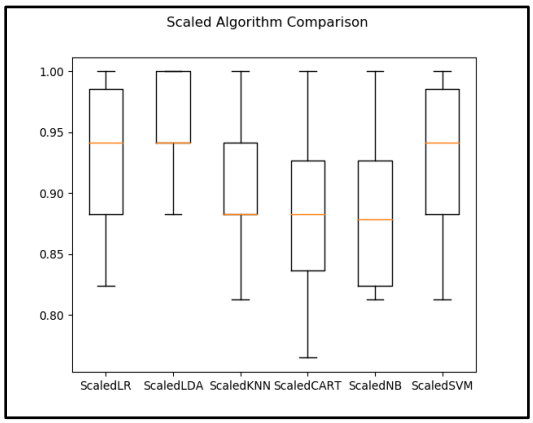

Fig 11: Scaled Algorithm Comparison

7.4. Predictions

In standard scalar prediction we will do prediction on different algorithm at many parameters, mean score, and standard score. SVM algorithm was the most accurate model that is 0.93 means 93%.

```
>>> model = SVC()
>>> kfold = cross_validation.KFold(n=num_instances, n_folds=num_folds, random_state=seed)
>>> grid = GridSearchCV(estimator=model, param_grid=param_grid, scoring=scoring, cv=kfold)
>>> grid_result = grid.fit(rescaledX, Y_train)
>>> print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_))

Best: 0.934524 using {'C': 1.5, 'kernel': 'linear'}
```

Fig 12: Prediction Result

It is valuable to keep a validation set just in case you made a slip during training. Such as over fitting to the training set or information leak both will result in an overly optimistic result. Run the SVM model on validation set and compress the outcomes as a final accuracy score, confusion, matrix and a classification report. The accuracy is 0.98 or 98% show in figure.

```
>>> print(accuracy_score(Y_validation, predictions))
0.9761904761904762
>>> print(confusion_matrix(Y_validation, predictions))
[[ 9  0  1]
 [ 0 20  0]
 [ 0  0 12]]
>>> print(classification_report(Y_validation, predictions))
             precision    recall  f1-score   support

        1.0       1.00      0.90      0.95        10
        2.0       1.00      1.00      1.00        20
        3.0       0.92      1.00      0.96        12

avg / total       0.98      0.98      0.98        42
```

Fig 13: Accuracy Result

The classification report provides a breakdown of each class by precision, recall, f1-score and support showing results.

VIII. CONCLUSION AND FUTURE SCOPE

In this study, our goal was to classify seed accuracy using machine algorithms. For this we used six different algorithms to implement. Eight independent variables were selected and examined in developing the model. The model is developed for prediction of determinant factors of seeds based on Machine Learning.

In this research we used Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Neighbors Classifier (KNN), Decision Tree Classifier (CART), Gaussian NB (NB), and Support Vector Machine (SVM). We send information in normal and scaled manner, and then prediction was performed. Based on all the outcomes we concluded that Support Vector Machine (SVM) perform better than all

algorithm. The accuracy of this algorithm emerged out to be more than that algorithm. The machine learning approach is recommended as flexible and precise way to solve the stated issue.

REFERENCES

[1] F. Guevara-Hernandez and J. Gomez-Gil, "A machine vision system for classification of wheat and barley grain kernels", *Spanish Journal of Agricultural Research*, vol. 9, no.3, pp. 672-680, 2011.

[2] TekalignTujo G., Dileep Kumar G., ElifeneshYitagesu D., MeseretGirma B. "Predictive Model to Predict Seed Classes using Machine Learning", *International Journal of Engineering Research & Technology (IJERT)*, vol. 6, no. 08, pp. 334-344, Aug. 2017.

[3] A. R. Parnian, "Autonomous Wheat Seed Type Classifier System", *International Journal of Computer Applications* , vol. 96 , no.12, pp. 14–17, 2014.

[4] C. S. Silva and U. Sonnadara, "Classification of Rice Grains Using Neural Networks", Proceedings of Technical Sessions, vol. 29, pp. 9–14, 2013.

[5] S. S. Dahikar and S. V Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach",*International Journal of Innovative Research in Electrical, Electronics, Instrumentation and* Control *Engineering*, vol. 2, no. 1, pp. 683–686, 2014.

[6] F. Guevara-Hernandez and J. Gomez-Gil, "A machine vision system for classification of wheat and barley grain kernels", *Spanish Journal of Agricultural and Research*, vol. 3, no. 9, 2011.

[7] Dreiseitl, Stephan and LucilaOhno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", *Journal of biomedical informatics*, vol. 35, pp. 352-359, 2002.

[8] Ye, Jieping, Ravi Janardan, and Qi Li. "Two-dimensional linear discriminant analysis" Advances in neural information processing systems, pp. 1569-1576, 2005.

[9] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines", ACM Transactions on Intelligent SystemsTechnology, vol. 2, pp. 1-27, 2011.

[10] Punn, Meesha, and NidhiBhalla. "Classification of wheat grains using machine algorithms" *International Journal of Science and Research*, vol. 2, no. 8, pp. 363-366, 2013.

[11] AltafSaeed, Muhammad Tariq, Muhammad Ibrahim, Nazir Ahmad Abu MazharAhmad,RaoSohailAftab, Syed Muntazir Mehdi, "Identification of Canola Seeds using Nearest Neighbor and K-Nearest Neighbor Algorithms", *Computer Engineering and Intelligent Systems*, Vol.6, no.10, pp. 36-42, 2015.