Show How Wheat Types Can Be Classified Using Geometric Measurements From Wheat Seeds.

1. ***Abstract***

*Data mining methods are now greatly needed to produce useful information due to the availability of enormous volumes of data. An essential supervised learning strategy in this new age of digitisation is categorisation, which is covered in-depth information regarding data mining techniques. To execute classification analysis on various types of accessible data, an open-source machine learning algorithm and java-based application (WEKA software) are also considered a tool of choice. With the help of WEKA, it is feasible to use data to draw out relevant information and choose the best technique for creating precise prediction models. Tasks involving categorisation and analysis comprise several machine learning techniques.*

*Additionally, it includes tools for preparing data, regressing, classifying, applying association rules, clustering, and selecting features in a general-purpose environment, and visualisation. This paper, which offers several machine learning techniques for data mining, will have categorisation as its primary emphasis, and the algorithms may be applied to the prepared wheat dataset. The classification approach differentiates the seed types Canadian, Rose, and Kama according to their geometrical characteristics. We use intelligent system methods to distinguish between the wide varieties of wheat kernels. From the UCI machine learning database, we selected the wheat dataset. Specifically, there are seven geometrical characteristics (area, perimeter, compactness, length of kernel, width of seed, asymmetry coefficient, length of kernel groove and one class). In WEKA classification, the following algorithm will be considered, Decision Tree(DT), Neural Network(NN)/Multi-Layer Perceptron(MLP), Support Vector Machine(SVM), Naïve Bayes(NB), and K-Nearest Neighbors(KNN). After that, the training test option that considers predicting model accuracy is Use Training Set(UTS), Cross-Validation(Folds-10)(CV-10), and Percentage Split(66%)(PS-66). The accuracy percentage for all the algorithms are; DT: UTS 100%, CV-10 89.5238%, PS-66 91.5493%; NN/MLP: UTS 99.5238%, CV-10 95.2381%, PS-66 95.7746%; SVM: UTS 94.2857%, CV-10 93.8095%, PS-66 91.5493%; NB: UTS 90.9524%, CV-10 91.4286%, PS-66 91.5493%; KNN: UTS 100%, CV-10 94.2857%, PS-66 91.5493%. Finally, It was observed that during the classification phase, the Use Training Set approach provides more accuracy than Cross-Validation.*

1. ***Introduction***

*In the twenty-first century, digitalisation and Data mining have made managing the exponential rise of data much more accessible. Businesses built data warehouses with millions of entries and characteristics, but they still await a return on their investment. They cannot create enough because they need more employees, expertise, and equipment. The automated categorisation of instances using data patterns gleaned from a dataset is known as data mining* [1]*. Various algorithms have been created and used to extract information and find knowledge patterns that can be helpful for decision assistance. Data preparation, pattern recognition, clustering, and classification are standard data mining technologies known as KDD (knowledge discovery in databases). The classification components of data mining will be the main emphasis of this study. Models are taught to produce categories in classification using a training data set. Creating a data collection, dimension reduction, feature selection, model selection, training the model, and predicting unidentified input samples are all steps in the classification process. There is a wealth of literature on data mining classification methods in the biological sciences and agriculture* [2]*. One of the most popular cereal grains eaten worldwide is wheat. Sorting wheat makes it more competitive. Manual sorting, which farmers use to separate wheat using just their eyes, is exceedingly labour-intensive, necessitates a larger workforce, and produces inaccurate results. It is a laborious, time-consuming, expensive, and sometimes erroneous process to analyse and manually sort wheat using the human eye and to sort wheat using machines that consider just a few aspects. As a result, Sorting the grain is crucial, and erroneous sorting may be prevented by not categorising the grain. Machine learning has shown to be a valuable tool for analysis across a range of fields and applications. Since digitalisation is being utilised in the project to categorise wheat, it does away with antiquated, manual processes and human error* [3]*.*

1. ***Dataset Description***

*The UCI website is an excellent dataset repository where the wheat seed data of size 13.2 kilobytes was collected* [4]*. Two hundred and ten (210) wheat seed samples from the three wheat classes, Kama, Rosa, and Canadian, are gathered for the categorisation procedure. Seeds are divided into three types of wheat based on seven geometrical or morphological characteristics (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove).*

*Chart, bar chart, histogram

Description automatically generated*

*Fig 3.1. All attributes for the dataset.*

The above fig 3.1 shows that the numeric input attributes Area with 179(85%) unique values, 193 distinct values and statistical values (Mean-14.848, Standard deviation- 2.91, minimum- 10.59, and Maximum - 21.18). Perimeter with 137(65%) unique values, 170 distinct values and statistical values (Mean-14.559, Standard deviation- 1.306, minimum- 12.41, and Maximum - 17.25). Compactness with 163(78%) unique values, 186 distinct values and statistical values (Mean- 0.871, Standard deviation- 20.024, minimum- 0.808, and Maximum - 0.918). Length of the kernel with 168(80%) unique values, 188 distinct values and statistical values (Mean- 5.629, Standard deviation- 0.443, minimum- 4.899, and Maximum - 6.675). Width of the kernel with 159(76%) unique values, 184 distinct values and statistical values (Mean- 3.259, Standard deviation- 0.378, minimum- 2.63, and Maximum - 4.033). Asymmetry coefficient with 204(97%) unique values, 207 distinct values and statistical values (Mean- 3.7, Standard deviation- 1.504, minimum- 0.765, and Maximum - 8.456). Length of kernel groove with 109(52%) unique values, 148 distinct values and statistical values (Mean- 5.408, Standard deviation- 0.491, minimum- 4.519, and Maximum - 6.55) and nominal output data class ratio are completely and balanced distributed.

1. ***Methodology***

*The wheat dataset was read through Jupiter Notebook with the following pandas, NumPy sets of python codes below, and screenshots for executing the various codes that display the instances(row), attributes(column) and the nominal class as displayed in fig 4.1.*

*>>>import pandas as pd*

*>>>import numpy as np*

*>>>from pandas.core.arrays.sparse import SparseArray as \_SparseArray*

*>>>wheatData = pd.read\_csv('/Users/apple/Desktop/Data Science Foundation/Course Work/WheatData(1).csv')*

*>>>wheatData*

*Table

Description automatically generated*

*Fig 4.1. Two hundred ten instances, 7-attributes and 1 class.*

The descriptive statistical data type information(float) and memory size of the wheat dataset is displayed in fig 4.2. below with the sets of python codes.

*>>>wheatData.info()*

*>>>wheatData.shape*

*Table

Description automatically generated*

*Fig 4.2. The numbers of rows & columns, data-type float64-attributes and object class.*

*The python codes display the total number of attributes/column names and the class object in the wheat dataset, as shown in fig 4.3.*

*>>>wheatData.columns*

*Graphical user interface, text, application

Description automatically generated*

*Fig 4.3. Columns name.*

*The code below shows detailed statistical data of the wheat datasets with the quartiles, mean, minimum, maximum, count/instances, and standard deviation of all the class attributes, as shown in fig 4.4.*

*>>>wheatData.describe().transpose()*

*Table

Description automatically generated*

*Fig 4.4. The count, mean, stand deviation, minimum & maximum attributes value, and percentile attributes.*

*To analyse the link between structural activity and classification, Weka software is employed. After processing for categorisation, the data mentioned above was produced in CSV format. Four classification techniques—Function, Bayes, Meta, and Lazy—are tested on wheat datasets to assess their properties.*

1. ***Modelling***

*Naïve Bayes: NB is a collection of supervised learning techniques, and while the distribution assumption permits numerical inputs, the input values are often taken to be nominal. The posterior probability for each class is calculated, and the class with the most significant likelihood is predicted. As a result, it handles issues involving binary and multi-class classification* [5]*.*

Text

Description automatically generated with low confidence

***Fig 5.1. Bayes.NaiveBayes algorithm accuracy data***

*The NB algorithm with 10-fold cross-validation in fig 5.1 above shows that out of 210-instances, 192- instances were correctly classified with 91.4286 % accuracy, 18-instances incorrectly classified with 8.5714 % inaccuracy, and the class accuracy with figures for each class(Kama: 0.857 True Positive rates, 0.057 False Positive rates, and 0.806 Matthew Correlation coefficient, Rosa: 0.943 True Positive rates, 0.021 False Positive rates, and 0.925 Matthew Correlation coefficient, and Canadian: 0.943 True Positive rates, 0.050 False Positive rates, and 0.884 Matthew Correlation coefficient). The confusion matrix shows 70 instances for each class; 60 correct instances are classified as kama, 66 valid instances are classified as rosa, and 66 right instances are classified as Canadian. The Matthew Correlation coefficient is derived from the confusion matrix, which predicts the balance or imbalance of the dataset.*

*Decision Tree: DT/ Classification And Regression Trees (CART) support classification and regression problems. It analyses a data instance and builds a tree; until a forecast can be made, one should start from the base and work their way up to the leaves (roots). Until the tree reaches a certain depth, creating a decision tree requires continually selecting the best split point to utilise as a prediction tool* [6]*. The decision tree in fig 5.2.1. shows transparency, lowers the risks and makes it easier to validate the input attributes of the wheat datasets.*

A picture containing text, sky, map

Description automatically generated

***Fig 5.2.1 trees.RandomTree algorithm visualise tree for accurate data.***

A picture containing text

Description automatically generated

***Fig 5.2.2. trees.RandomTree algorithm accuracy data***

The DT algorithm with 10-fold cross-validation in fig 5.2.2 above shows that out of 210-instances, 188- instances were correctly classified with 89.5238 % accuracy, 22-instances incorrectly classified with 10.4762 % inaccuracy, and the class accuracy with figures for each class(Kama: 0.871 True Positive rates, 0.093 False Positive rates, and 0.768 Matthew Correlation coefficient, Rosa: *0.957 True Positive rates, 0.021 False Positive rates, and 0.936 Matthew Correlation coefficient, and Canadian: 0.857 True Positive rates, 0.043 False Positive rates, and 0.827 Matthew Correlation coefficient). The confusion matrix shows 70 instances for each class; 61 correct instances are classified as kama, 67 right instances are classified as rosa, and 60 valid instances are classified as Canadian.*

*K-Nearest Neighbors: KNN is a straightforward method that allows regression and classification. The training dataset is stored and queried to locate the k most comparable training patterns before creating a forecast. Therefore, the only computation required to make a prediction is to query the training dataset. Other than the primary training dataset, there is no model. It is a simple strategy, but it only relies on one fundamental premise—that the distance between data instances affects the accuracy of predictions. It often results in an outstanding performance as consequence. When making predictions for classification problems, KNN will utilise the mode (most common class) of the k most similar examples in the training dataset* [7]*.*

Text

Description automatically generated with medium confidence

***Fig 5.3.*** ***lazy.IBk algorithm accuracy data***

The KNN algorithm with 10-fold cross-validation in fig 5.3 above shows that out of 210-instances, 198- instances were correctly classified with *94.2857* % accuracy, 12-instances incorrectly classified with *5.7143* % inaccuracy, and *the class accuracy with figures for each class(Kama: 0.886 True Positive rates, 0.029 False Positive rates, and 0.870 Matthew Correlation coefficient, Rosa: 0.957 True Positive rates, 0.014 False Positive rates, and 0.946 Matthew Correlation coefficient, and Canadian: 0.986 True Positive rates, 0.043 False Positive rates, and 0.928 Matthew Correlation coefficient). The confusion matrix shows 70 instances for each class; 62 correct instances are classified as kama, 67 valid instances are classified as rosa, and 69 valid instances are classified as Canadian.*

*Neural Network: NN/MLP is a machine learning model that mimics the functions of the human brain by taking inspiration from biological neural networks. NN comprises three layers: the input layer, the hidden layer, and the output layer* [8] *as shown in the fig 5.4.1. below.*

*1. Input Layer indicates the input variables to be supplied into the network.*

*2. Hidden Layers are the training-relevant computation layers (or parameters).*

*3. Output Layer refers to the model's output, using the class label in a*

*classification task or the actual number in a regression task as*

*examples.*

Diagram

Description automatically generated with low confidence

*Fig 5.4.1 Describes input, hidden, and output layer mesh.*

*Text

Description automatically generated with medium confidence*

***Fig 5.4.2*** ***functions.SMO algorithm accuracy data***

The NN algorithm with 10-fold cross-validation in fig 5.4.2 above shows that out of 210-instances, 197- instances were correctly classified with *93.8095* % accuracy, 13-instances incorrectly classified with *6.1905* % inaccuracy, and *the class accuracy with figures for each class(Kama: 0.900 True Positive rates, 0.043 False Positive rates, and 0.860 Matthew Correlation coefficient, Rosa: 0.971 True Positive rates, 0.021 False Positive rates, and 0.947 Matthew Correlation coefficient, and Canadian: 0.943 True Positive rates, 0.029 False Positive rates, and 0.914 Matthew Correlation coefficient). The confusion matrix shows 70 instances for each class, 63 correctly samples classified as kama, 68 correctly classified as rosa, and 66 correctly classified as Canadian.*

*Support Vector Machine: SVM were created for binary classification issues using numerical input variables. The method has since been extended to cover multi-class classification and regression issues. Additionally, it automatically transforms nominal values into numerical values; before being utilised, input data is standardised* [9]*. SVM identifies the optimal line to divide the data into two groups. An optimisation method is used; however, it only considers data instances from the training dataset most closely related to the line that best demarcates the classes. The examples are referred to as support vectors, therefore the technique's name.*

*Text

Description automatically generated with medium confidence*

***Fig 5.5.*** ***functions.MultilayerPerceptron algorithm accuracy data***

The SVM algorithm with 10-fold cross-validation in fig 5.5 above shows that out of 210-instances, 200- instances were correctly classified with 95.2381 % accuracy, 10-instances incorrectly classified with 4.7619 % inaccuracy, and the class accuracy with figures for each class(Kama: 0.929 True Positive rates, 0.036 False Positive rates, and 0.893 Matthew Correlation coefficient, Rosa: 1.000 True Positive rates, 0.021 False Positive rates, and 0.969 Matthew Correlation coefficient, and Canadian: 0.929 True Positive rates, 0.014 False Positive rates, and 0.925 Matthew Correlation coefficient). The confusion matrix shows 70 instances for each class, 65 valid samples classified as kama, 70 correctly classified as rosa, and 65 correctly classified as Canadian.

1. ***Results And Discussions***

*The Weka classifiers that were selected had the best accuracy compared to the other classifiers. Area, compactness, perimeter, the width of the kernel, length of the kernel, length of kernel groove, and asymmetry coefficient are the characteristics utilised for categorisation. Each seed, including Canadian, Kama, and Rosa, had 70 samples collected each for the classes.*

*Table

Description automatically generated*

***Fig 6. Wheat Dataset Classification Results.***

*Fig 6 above depicts the accuracy values against each of the classifiers. Fig 6 shows that the classifier Multilayer Perceptron(SVM) and Lazy IBK(KNN) give the highest accuracy, 95.2 % and 94.2 %, among all other classifiers when using 10-Fold Cross Validation. The third highest accuracy was obtained from the Functions SMO Classifier(NN), which is 93.8 %, while the Decision Tree classifier gives lower classification accuracy of 89.5 %. Similarly, the highest MMC and ROC values were obtained for the classes kama(89.3% | 99.3%), Rosa(96.9% | 99.9%) and Candian(92.5% | 99.6%), respectively) from the Multilayer Perceptron(SVM), the second highest is Lazy IBK(KNN).*

1. **Conclusion**

*Using Weka classification techniques and methodologies, the Kama, Rosa, and Canadian seeds are categorised in the current study. The database on the UCI website provided the data set. 10-fold cross-validations, user training data, and percentage splits (66%) were used to quantify performance. SVM (Multilayer Perceptron) with 10-fold cross-validation provides the most remarkable accuracy performance of any Weka classifier (95.2%). In comparison, Decision Tree and K-Nearest Neighbors give the highest accuracy performance of any Weka classifier (100.0%) when we utilise the training set technique. The cross-validation approach performs less well than the training set. In the future, classifiers other than the Weka classifiers would have been developed, and classifiers would also be combined to increase accuracy performance further. In addition, unsupervised machine learning methods exist, including clustering, which classifies seeds. Additionally, after extracting these characteristics, these can be further utilized by using the same techniques and classifiers to categorise the additional seed categories.*

References

[1] S. Singhal and M. Jena, "A study on WEKA tool for data preprocessing, classification and clustering," *International Journal of Innovative Technology and Exploring Engineering (IJItee),* vol. 2, *(6),* pp. 250-253, 2013.

[2] J. Madhavan *et al*, "Wheat seed classification using neural network pattern recognizer," *Materials Today: Proceedings,* 2021.

[3] R. H. Ajaz and L. Hussain, "Seed classification using machine learning techniques," *Seed,* vol. 2, *(5),* 2015.

[4] D. a. G. Dua C., "UCI Machine Learning Repository," *Irvine, CA: University of California, School of Information and Computer Science.,* vol. 11, *(11),* pp. 1639-1646, 2019. Available: <http://archive.ics.uci.edu/ml.>

[5] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," *arXiv Preprint arXiv:1302.4964,* 2013.

[6] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news," *International Journal of Innovative Science, Engineering & Technology,* vol. 2, *(2),* pp. 438-446, 2015.

[7] S. B. Aher and L. Lobo, "Data mining in educational system using weka," in *International Conference on Emerging Technology Trends (ICETT),* 2011, .

[8] S. Shedole and B. J. Sowmya, "A Convolution Neural Network-Based Wheat Grain Classification System," *Journal of Scientific Research,* vol. 66, *(2),* 2022.

[9] M. Olgun *et al*, "Wheat grain classification by using dense SIFT features with SVM classifier," *Comput. Electron. Agric.,* vol. 122, pp. 185-190, 2016.

[10] E. G. Kulkarni and R. B. Kulkarni, "Weka powerful tool in data mining," *International Journal of Computer Applications,* vol. 975, pp. 8887, 2016.

[11] S. K. David, A. T. Saeb and K. Al Rubeaan, "Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics," Computer Engineering and Intelligent Systems, vol. 4, (13), pp. 28-38, 2013.

[12] S. S. Keerthi et al., "Improvements to Platt's SMO algorithm for SVM classifier design," Neural Comput., vol. 13, (3), pp. 637-649, 2001.

[13] M. I. Jordan, M. J. Kearns and S. A. Solla, Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference. 199810.

stylefix