Prediction of Black Sigatoka Disease in Banana Plants By Data Mining Classification Techniques using Scikit for Python

Srivalli Devi S, A Geetha

Abstract*: Agriculture has been evolving since humans started cultivating plants for food consumption. As the agriculture field evolves, the disease control measures too have evolved. Now in this modern era, disease in plants can be easily identified using computers. Data mining is the process of obtaining the useful information from the data. Before the electronic era, diseases in plants are identified just by seeing the symptoms of the plants. Similarly, we can identify the diseases in plants using data mining by supplying the disease symptoms data and classify them accordingly. The purpose of this paper is focusing on the prediction of the diseases from images of black sigatoka disease and uses the following methods: MultilayerPerceptrons, SVM,KNeighborsClassifier,K-NeighborsRegressor, Gaussian Process Regressor, Gaussian Process Classifier, GaussianNB, Decision Tree Classifier, Decision Tree Regressor, linear models such as Linear Regression, RidgeCV, Lasso, ElasticNet, Logistic RegressionCV, SGD Classifier, Perceptron and Passive Aggressive Classifier and ensemble models of the above classifiers. The results are compared, and multilayer perceptron model is seen to give better results for individual classifiers and ensemble of week classifiers gives better results when ensembled. In future, a new hybrid algorithm would be used from the above algorithms for attaining better accuracy. The scikit is a library used for classification, clustering, regression, dimensionality reduction,model selection and preprocessing. Our paper discusses various classifiers used in scikit-learn library for Python and their ensembling is done. This can be applied to all the classification tasks. Classification is done for classifying the black sigatoka disease in banana from healthy leaves.This disease is the most vulnerable one among banana plants.

***Keywords* : agriculture, black sigatoka, classification, data mining, scikit.**

I. INTRODUCTION

Data mining is the study of gathering, cleansing, handling, exploring, and getting beneficial perceptions from data [1]. Such analysis can be used for research purposes and may lead in obtaining new knowledge. Agricultural applications such as seed and crop-growth analysis and crop yield prediction with various other factors soil condition, temperature, rainfall, fertilizer, pesticides, irrigation. Two important factors helping agriculturist and government in decision making are providing the historical crop yield record for farmers and making crop insurance policies for government [2]. Data mining techniques can be categorized predictive, descriptive and prescriptive [3]. Descriptive analytics gives insight into the past, predictive analytics helps in understanding the future and prescriptive analytics advises on possible outcomes.

Common challenges faced in agriculture are disease occurrence in cultivated plants, poor water management, improper use of fertilizers, pesticides and fungicides. Diseases can be identified only after the visible symptoms occur. The reasons of disease attacks are abiotic, mesobiotic and biotic factors. Plant diseases can be categorized as endemic, epidemic and sporadic. The different modes of spreading are through seed, soil, air and insects [4].

Pseudocercospora fijiensis is the fungus that causes the black sigatoka disease (black streak leaf disease) is a disease that occurs in banana plants. This occurs on the leaves of the banana plant. India, Pakistan, South Africa, Israel and mainland Australia are free of this disease [5]. The initial symptoms of this diseases are very small chlorotic spots on bottom of 3rd or 4th open leaf which then grows into brown streaks running parallel to leaf veins. The streaks forms clusters and spots merge forming a chlorotic halo and streak's colour becomes darker visible on the top side of the leaf. This lesion then becomes big and darken giving black streaks [6].

Black Sigatoka disease is a threat for global exports and food consumption [25]. And therefore, it's better to identify the disease as early as possible to avoid the fast spreading of the diseases.is is an International reputed journal that published research articles globally.

II. RELATED WORK

Decision Trees, Random Forest, Gaussian Naïve Bayes, Support Vector Machine, Neural Networks, K-Nearest Neighbours, ensemble models of the above algorithms are used for the grass grub damage prediction. Combination of Decision Trees, Random forest and Support Vector Machine has produced better results [8]. The metrics considered are accuracy, mean accuracy, precision, recall and F-1 score. Data mining techniques such as Naïve Bayes, Logistic Regression, Random forest, Support Vector Machine, Adaboost, J48 are used [9].

Revised Manuscript Received on January 5, 2020
***** Correspondence Author
Srivalli Devi S*, PG & Research Department of Computer Science, Chikkanna Government Arts college, Tirupur, India. Email: ssrivallidevi@gmail.com
Dr.A.Geetha, PG & Research Department of Computer Science, Chikkanna Government Arts college, Tirupur, India. Email: gee_sam@yahoo.com

Spray application decision prediction is their focus under study. The authors concluded that Naïve Bayes and AdaBoost were the best classifiers for predicting spray decisions. Metrics used are precision, recall, true positive and false positive. The black sigatoka disease can be predicted using image processing and data mining techniques. Various classification algorithms used are Decision tree, Random forest, Extremely Randomized Trees, Naive Bayes and support vector classifier (Linear SVM and RBF SVM), Nearest Neighbours [10]. The performance was measured using AUC (area under curve). The Extremely Randomized Trees performed well when compared to other classifiers.

The back propagation neural networks was used for the classification of cotton leaf diseases with pattern recognition techniques. This kind of neural network solves multiple class problems [11]. Artificial neural network is used as the classifier in the detection of leaf disease along with some computer vision and pattern recognition techniques with fuzzy logic for grading the diseases [12]. Pomegranate disease classification is done using back propagation neural network[13]. The sugarcane rust disease identification is done using SVM (support vector machine) [14] and also SVM used to predict the leaf diseases in wheat plant, supporting multiple classes [15].

Banana is one of the key foods along with fruit crops that have potential to give ample revenue to the farmers and country's financial system. Machine learning techniques like minimum distance classifier, support vector machines, artificial neural networks, Principal component analysis and k-nearest neighbour are used to find the diseases in extracted area of the image using image processing techniques [21].Expert system was constructed for the identification of different banana diseases using Delphi and CLIPS [22]. But only nine different banana diseases were identified.

Images were pre-processed using image processing techniques and for the process of classification machine learning algorithms such as nearest neighbours, Decision Tree, Naïve Bayes,Random forest and Support Vector Machine trained and used to classify between healthy and diseased leaves of banana. The experimental results suggested that randomized trees produced good results [23]. Banana leaves diseases are identified using Adaptive Neuro Fuzzy Systems and Support Vector Machine. In [24] the work comprises of detecting and rating the affected segment of banana Black Sigatoka disease and Panama wilt disease. It uses video as input. Image processing techniques used for segmentation.

III. METHODS

In this paper research focuses on the classification of banana plant disease black sigatoka using data mining. This approach helps in agriculturist to easily identify using images from robots or drones. The data mining techniques used here are Multilayer Perceptrons,SVM, K-Neighbors Classifier, K-Neighbors Regressor, Gaussian Process Regressor, Gaussian Process Classifier, GaussianNB, Decision Tree Classifier, Decision Tree Regressor, linear models such as Linear Regression, RidgeCV, Lasso, ElasticNet, Logistic RegressionCV, SGD Classifier, Perceptron and Passive Aggressive Classifier. An ensemble of the classifiers is used, and they show a better accuracy than used separately.

All the work is done in Python using scikit-learn library. Scikit-learn is a machine learning library for Python programming language. It contains numerous classification, regression and clustering algorithms..

A. Dataset

The dataset used consists of 43 black sigatoka images in RGBA format and 220 bacterial wilt disease images in RGBA format and 359 healthy leaf images in RGBA format [7].We consider black sigatoka images and healthy leaves images for ensembling. RGBA format is the color channel representation of red, green, blue and alpha channel.

B. Analysis

Classifier selection

Nine supervised learning classifiers were tested. Those are Decision trees,Adaboost,SVM, KNN, Naïve Bayes, MLP-NN, Random forest, GBM and Logistic classifiers.

- Decision Tree: Decision trees are commonly applied in data mining. Decision trees condense the connection between attributes and the class of an object in a branching tree structure. End of the tree branches are named leaves. Dividing lasts to a user-defined end point. Estimates are made by sorting new instances until a leaf is reached [9].
- AdaBoost: On frequently revised editions of the data AdaBoost fits a series of weak learners. Then to generate the final prediction, all those predictions from weak learners are merged through a weighted majority vote [17].
- Support Vector Machine: The large margin classifier and a vector space based which finds the boundary far from training point [18]. In a hyperplane, classes are represented as a vectors. SVM margins by utilizing a function which increases the two different class distance[9] .
- K-Nearest Neighbour: Voting obtained from nearest neighbours, KNN predicts the class. Training is not required. It picks values from the nearest example by calculating the distance from all [9].
- Naïve Bayes: From the prior probabilities of the training set, testing set probabilities are calculated. Naïve Bayes is a probabilistic classifier. When relationship is being built, each variable is treated independently from others.
- Multilayer Perceptron: A classical type of neural network. Used in classification prediction problems. One or more layers of neurons are there. The input layer is used for feeding data, one or more layers as hidden layers for abstraction, output layer is used for predicting the outcomes [19] .
- Random Forest: Decision trees can be ensembled to form random forests[16]. From every decision tree to predict its class label random forest obtains the vote for testing example. Different ensemble models can have different voting schemes. We run the RandomForest algorithm by keeping the default values of attributes defined by scikit-learn python library.

- Logistic Regression: Output values are modelled as binary value. In Logistic regression the input values combined to predict output values, in a linear way using weights.
- Gradient Boosting Classifier: Many weak learning algorithms are combined together in gradient boosting classifiers for strong predictive modelling. It is very effective in classifying complex datasets [20].
- Ensembling: Combination of nine classifiers were taken for ensembling and had brought down to ensembling of few classifiers from the nine used, whichever scored good.

IV. RESULTS AND DISCUSSIONS

A. Tables and figures

Table-I shows the results of performance of separate classifiers on the dataset considering healthy and black sigatoka images. Table-II shows the results of ensemble methods run on the healthy and black sigatoka images. The overall Ensemble ROC-AUC score is 0.929. The figure Fig.1 shows the correlation matrix for the ensemble methods. The ensembles of different classifiers and their ensembling score are illustrated in the Table-III. The ensembling of SVM,Logistic, MLP-NN, AdaBoost gives a good ROC_CURVE and hence considered for classification.

Table- I: Performance measures of classifiers

Classifier/Regressor used	Parameter value	Mean Squared Error	Variance	Accuracy	Score
LinearRegression	Default python sklearn values	0.06	-0.14		-0.13904972809214744
RidgeCV	alpha=numpy.logspace(-6, 6, 13)	0.05	0.16		0.16313190596721125
Lasso	alpha=0.1	0.06	-0.09		--0.08940924062079181
ElasticNet	random_state=0	0.06	-0.09		-0.0894092406207918 1
Logistic RegressionCV	cv=5, random_state=0,multi_class='multinomial'	0.04	0.24	1.0	0.9586776859504132
SGD Classifier	loss="hinge", penalty="l2", max_iter=5	0.07	-0.36	0.9113475177 304965	0.9256198347107438
Perceptron	tol=1e-3, random_state=0	0.17	-2.18	0.9042553191 489362	0.8264462809917356
Passive Aggressive Classifier	max_iter=1000, random_state=0, tol=1e-3	0.14	-1.58	0.9042553191 489362	0.859504132231405
K-Neighbors Classifier	n_neighbors=3	0.06	-0.06	0.9361702127 659575	0.9421487603305785
K-Neighbors Regressor	n_neighbors=2	0.04	0.24		0.24185463659147843
Gaussian Process Classifier	Default sklearn python values	0.05	0.09	1.0	0.9504132231404959
Gaussian Process Regressor	Default sklearn python values	0.06	-0.14		-0.14046470000134614

GaussianNB	Default sklearn python values	0.05	0.09	0.8085106382978723	0.7768595041322314
Decision Tree Classifier	Default sklearn python values			1.0	0.9421487603305785
Decision Tree Regressor	Default sklearn python values			1.0	0.09022556390977421
Logistic RegressionCV	cv=5, random_state=0,multi_class='multinomial'	0.04	0.24	1.0	0.9586776859504132
SGD Classifier	loss="hinge", penalty="l2", max_iter=5	0.07	-0.36	0.9113475177304965	0.9256198347107438
Perceptron	tol=1e-3, random_state=0	0.17	-2.18	0.9042553191489362	0.8264462809917356
Passive Aggressive Classifier	max_iter=1000, random_state=0, tol=1e-3	0.14	-1.58	0.9042553191489362	0.859504132231405
MLP Classifier	solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5, 2), random_state=1			0.9326241134751773	0.9504132231404959

Table- II: Ensemble methods and their performance

Classifier	Score
SVM	0.964
KNN	0.883
Naïve Bayes	0.782
MLP – NN	0.893
Random forest	0.922
GBM	0.905
Logistic	0.931
Decision Tree	0.630
Adaboost	0.876

Table-III: Ensemble ROC-AUC Score for different combinations of classifiers

Classifier	**Ensemble ROC-AUC Score**
SVM, KNN, Naïve Bayes	0.912
SVM, Logistic, Random forest	0.951
KNN, MLP-NN, AdaBoost	0.886
SVM, Logistic, MLP-NN, AdaBoost	0.965

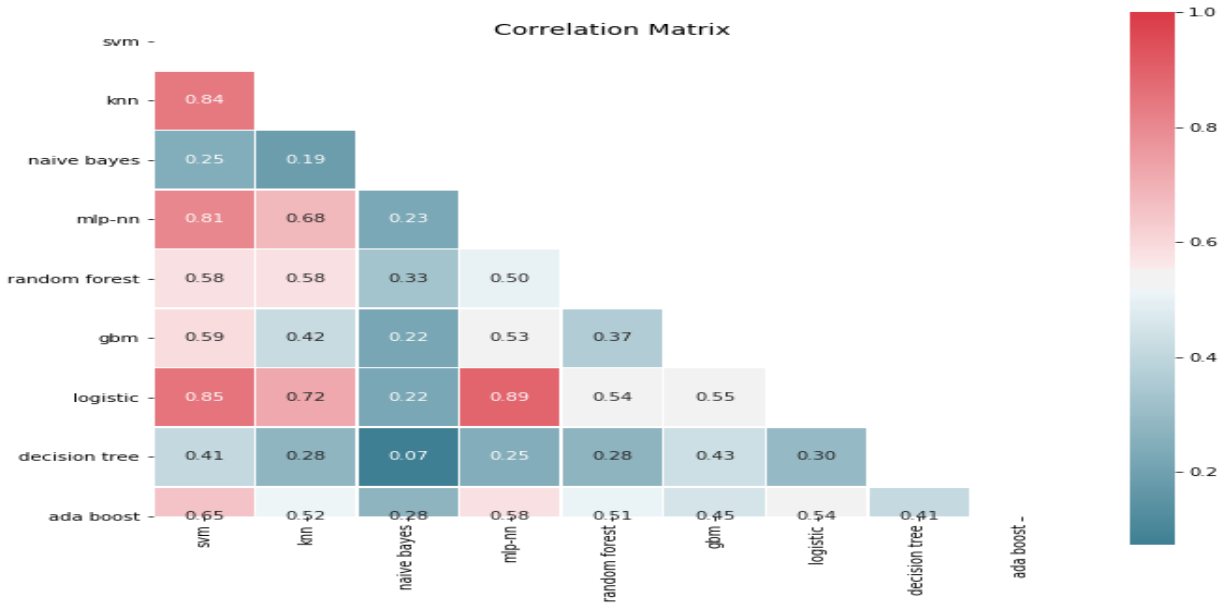

Fig.1.Correlation matrix for ensemble

V. CONCLUSION

The ensembling of SVM,Logistic , MLP-NN and AdaBoost gives a good score and hence is considered for the classification of black sigatoka disease in banana plants using data mining algorithms. These algorithms convert the image pixel data into arrays and classification is done based on those arrays rather that that of the direct image classification using only pixels as in case of other machine learning and deep learning algorithms . Thus it can be seen ensembing achieves good accuracy than that of individual classifier models. In future hybridization of these algorithms maybe done. And the algorithms may be run on different data sets . Deep learning techniques can be used in future for the prediction of diseases.

REFERENCES

1. C. C. Aggarwal, Data mining: the textbook. Switzerland: Springer Internaltional Publishing Switzerland, 2015.
2. Majumdar, J., Naraseeyappa, S. & Ankalaki, S. Analysis of agriculture data using data mining techniques: application of big data. J Big Data 4, 20 (2017) https://doi:10.1186/s40537-017-0077-4
3. Halobi website https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/
4. National Institute of Open Schooling website http://oer.nios.ac.in/wiki/index.php/Plant_Diseases
5. Yonow T, Ramirez-Villegas J, Abadie C, Darnell RE, Ota N, Kriticos DJ (2019) Black Sigatoka in bananas: Ecoclimatic suitability and disease pressure assessments. PLoS ONE 14(8): e0220601. https://doi.org/10.1371/journal.pone.0220601
6. Alabi Israel, published on Feb,2017 at the website https://medium.com/@alabiisrael/how-to-effectively-eliminate-black-sigatoka-disease-on-your-plantain-farm-5f47219b106a
7. Godliver Owomugisha, Implementation-BBW and BBS Diseases (2014), GitHub repository, https://github.com/godliver/source-code-BBW-BBS
8. Ayub, U., & Moqurrab, S. A. (2018). Predicting crop diseases using data mining approaches: Classification. 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG). https://doi:10.1109/icpesg.2018.8384523
9. Hill, M.G. & Connolly, P.G. & Reutemann, Peter & Fletcher, Dale. (2014). The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. Computers and Electronics in Agriculture. 108. 250 - 257. https://doi:10.1016/j.compag.2014.08.011
10. Owomugisha, Godliver & Quinn, John & Mwebaze, Ernest & Lwasa, James. (2014). Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease. 1st International conference on the use of mobile ICT in Africa, At 9-10th December 2014, Stellenbosch, South Africa
11. P. R. Rothe and R. V. Kshirsagar," Cotton Leaf Disease Identification using Pattern Recognition Techniques", International Conference on Pervasive Computing (ICPC),2015.
12. Aakanksha Rastogi, Ritika Arora and Shanu Sharma," Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic" 2nd International Conference on Signal Processing and Integrated Networks (SPIN)2015.
13. S. S. Sannakki and V. S. Rajpurohit,"Classification of Pomegranate Diseases Based on Back Propagation Neural Network," International Research Journal of Engineering and Technology (IRJET), Vol2 Issue: 02 | May-2015.
14. Ratih Kartika Dewi and R. V. Hari Ginardi," Feature Extraction for Identification of Sugarcane Rust Disease", International Conference on Information, Communication Technology and System, 2014.
15. Yuan Tian,Chunjiang Zhao, Shenglian Lu and Xinyu Guo," SVM-based Multiple Classifier System for Recognition of Wheat Leaf Diseases," Proceedings of 2010 Conference on Dependable Computing (CDC'2010),November 20-22, 2010.
16. U. Ayub and S. A. Moqurrab, "Predicting crop diseases using data mining approaches: Classification," 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG), Mirpur Azad Kashmir, 2018, pp. 1-6. https://doi: 10.1109/ICPESG.2018.8384523
17. scikit website https://scikit-learn.org/stable/modules/ensemble.html
18. Online book chapter from https://nlp.stanford.edu/IR-book/pdf/15svm.pdf
19. Machine learning website https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/
20. Website https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/#:~:targetText=Gradient%20boosting%20classifiers%20are%20a,used%20when%20doing%20gradient%20boosting
21. Deenan, Surya Prabha & SatheeshKumar J, (2014). Study on Banana Leaf Disease Identification Using Image Processing Methods. International Journal of Research in Computer Science and Information Technology. 2. 89-94.
22. Almadhoun, Hamza Rafiq and Abu Naser, Samy S., Banana Knowledge Based System Diagnosis and Treatment (July 25, 2018). International Journal of Academic Pedagogical Research (IJAPR), 2(7), 1-11, July 2018. Available at SSRN: https://ssrn.com/abstract=3219792

23. Tejada, J. A., & Gara, G. P. P. (2017). LeafcheckIT. Proceedings of the 3rd International Conference on Communication and Information Processing - ICCIP '17. https://doi:10.1145/3162957.3163035
24. Vipinadas.M.J, A.Thamizharasi (2016). Detection and Grading of diseases in Banana leaves using Machine Learning. International Journal Of Scientific & Engineering Research, Volume 7, Issue 7, July-2016.
25. Douglas H. Marín, Ronald A. Romero, Mauricio Guzmán, and Turner B. Sutton (2003). Black Sigatoka: An Increasing Threat to Banana Cultivation, Plant Disease / Vol. 87 No. 3

AUTHORS PROFILE

Srivalli Devi S, is a research scholar, currently doing her Ph.D at the PG & Research Dept. of Computer Science, at Chikkanna Govt. Arts College, Tirupur, Tamilnadu,India. Her areas of interest include data mining, machine learning, cryptography and robotics. Her research is focused on solving agricultural problems with the help of data science.

Dr.A.Geetha, is currently heading the PG & Research Dept. of Computer Science, at Chikkanna Govt. Arts College, Tirupur -2. She has got more than 22 years of Collegiate Teaching experience. A highly motivated teacher, with the thirst for learning new concepts. She regularly updates herself by attending workshops and hands on training sessions, to keep herself abreast with latest tools and techniques. Her research area includes Data Mining and Big Data Analytics. She has got more than 50 publications in National and International Conferences and in reputed journals.