Transactions of the Nigerian Association of Mathematical Physics
Volume 7, **(***March, 2018***), pp63 – 70**

PLANT LEAVES CLASSIFICATION USING K-NEAREST NEIGHBOR ALGORITHM

[1]Usman M. A., [2]Olayiwola M.O., [1]Folorunsho S.O., [1]Solanke O. O., [1]Hammed F.A., [1]Okusaga S.T. and [1]Oyebo, A.B.

[1]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State.
[2]Department of Mathematical Sciences, Osun State University, Osogbo.

Abstract

This paper investigates plant leaf classification using k-Nearest Neighbor Algorithm. Plant classification is important for agriculturist, botanist and medicine, and is especially significant to the biology diversity research. As plants are vitally important for environmental protection, it is more important to identify and classify them accurately. This study is aimed at building a machine learning model to classify plant species using extracted features (shape, margin and texture) of plants leaves. This project applied the k-Nearest Neighbor (k-NN) model to automatically classify and predict plant species. k-Nearest Neighbor model was used as a base learner and also as a bagged ensemble with varying values of k = 3, 5, 7, 9, 11, 13, 17, 19, 21. The classification performance of both k-NN and its ensemble model were compared based on accuracy and log-loss values. It was observed from the results obtained that the k-NN with k=3 outperformed all other models with the accuracy value of 88.89% and the log-loss value of 1.17. It is also observed that as the value of k is increasing the accuracy and Log Loss of the model is decreasing.

***Keywords*:** Plant classification, k-NN, Machine learning, log-loss, plant leaf, Ensemble, Bagging

1. Introduction

Leaves are very important for trees they provide food for the whole tree. Leaves use a very special process called *photosynthesis* to convert energy from sunlight into sugars and starches that a tree uses as food. Leaves have an important chemical inside of them called *chlorophyll.* Which is what makes them green, and is also what allows them to do photosynthesis [1,2,3]. Plants exist everywhere we live, as well as places without us. Many of them carry significant information for the development of human society. The urgent situation is that many plants are at the risk of extinction. So it is very necessary to setup a database for plant protection [4,5]. We believe that the first step is to teach a computer how to classify plants. For all forms of life, plants form the basic food staples, and this is just one reason why plants are important. They are the major source of oxygen and food on earth since no animal is able to supply the components necessary without plants.[6,7,8] The ability to know, or identify plants allows them to assess many important rangeland or pasture variables that are critical to proper management, range condition, proper stocking rates, forage production, wildlife habitat quality, and rangeland trend, either upward or downward. Natural resource managers, especially those interested in grazing and wildlife management must be able to evaluate the presence or absence of many plant species in order to assess these variables [9,10,11,12].

Therefore it necessary for preserving these natural resources and hence, it is great requirement to correctly and quickly identify plant species and classify them accordingly.[13,14] The classification process of plant species identifies the different kinds of plants. Thus builds the system to classify the plant species. Many researchers have tried to classify the plant species based on the various features of plants. Many of them have used leaf biometric features for plant classification as they are more readily available in almost all season. [15,16] In the classification system, plants are classified on the basis of color, shape venation, margin, color and texture features [17]. For identification of the plant species, individual plants are classified to a species based on its leaf physiological characteristics. Plant taxonomy theory says that leaves and flowers are more important components of plant biometry that are essentially used for classifying plants. Leaves are the components readily

Correspondence Author: Usman M.A., Email: usmanma@yahoo.com, Tel: +2348033454676

available in almost all season. Therefore are used in plant classification. Leaves are important part of plant and are distinct in shape and venation. A leaf is containing three basic parts, leaf lamina, stem and a small leaf like component at the base of the stem called stipules. According to [18], if the plant classification is based on only two dimensional images, it is very difficult to study the shapes of flowers, seedling and morph of plants because of their complex three dimensional structures.[19] This is very much important to improve this process by identifying plant species by using computerized techniques.

2. LITERATURE REVIEW

Several plant leafs machine learning Algorithm classification have been developed based on various features and classifiers. Many of these works were based on artificial neural networks (ANN) as Machine Learning Models due to their adaptability and scalability. Table 1 below shows some recent works on plant leaves for automatic classification of plant species.

Table 1.0: List of References

Authors	Techniques	Result
[19]	KNN algorithms for plant classification based on leaf images	k-NN gives better performance
[2]	Plant texture classification using Gabor co-occurrences	
[6]	shape and vein, color, and texture features to classify leaves using probabilistic neural network	accuracy of 93.75%
[11]	classifying plant leaves using the 2-dimensional shape feature, using distributed hierarchical graph neuron (DHGN) for pattern recognition and k-Nearest Neighbors (k-NN) for pattern classification	distributed hierarchical graph neuron (DHGN) with accuracy of 83.67%
[16]	for classification of plants which was based on the characterization of texture properties	The proposed systems ability to classify and recognize a plant from a small part of the leaf is its advantageous thing
[17]	Applied a feature fusion technique using the Gabor filter in the frequency domain and fusing the obtained features with edge based feature extraction. The extracted features were trained using 10 fold cross validation and tested with CART and RBF classifiers to measure its accuracy	accuracy 85.93 % with low relative error for a nine class problem

3. METHODOLOGY

The methodology adopted follow the knowledge discovery pattern and it's presented in figure 1. k-Nearest Neighbor as a base learner and its bagging ensemble using k-NN as a base were trained and tested on the same dataset. Stratified cross validation was then used on each dataset. This process was repeated for k ϵ {3,5,7,9,11,13,15,17,19,211} and all result with k were used. Classifier were implemented using Python open source technologies on Anaconda Machine Learning suite (Jupyter).

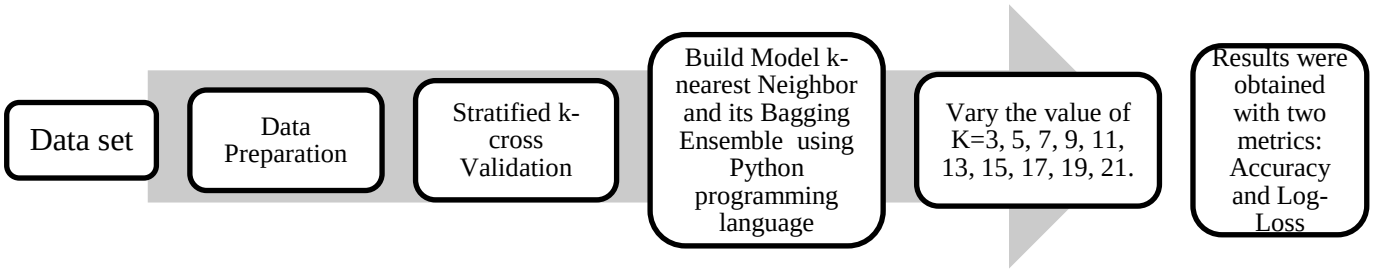


Figure 1: The Design Methodology

Table 1.1: DATA SET

Dataset	No of Samples	No of attributes	Type of attributes
All plant variety	Train: 991 Test: 595	194 193	Margin: 64 Shape: 64 Texture: 64 Id: 1 Species 1

3.2 K NEAREST NEIGHBOR

The k-NN is metric-based algorithm that solves classification problem by looking for the shortest distance between the test data and training sets in the feature space. The distance is generally computed in Pythagorean sense (by finding the square root of the sum of differences).

Suppose the training set, using the features below is defined as;

$$x = \begin{bmatrix} x_{11} & x_{12} & x_{13} & K & K_{1N} \\ x_{21} & x_{22} & x_{23} & K & K_{2N} \\ x_{31} & x_{32} & x_{33} & K & K_{3N} \\ M & M & M & M & M \\ x_{m1} & x_{m2} & x_{m3} & K & K_{MN} \end{bmatrix} \quad (1.0)$$

Where M = 1,584, the number of observations in these dataset. The number of features here is N=192. The k-NN algorithm computes Euclidean distance between test data and all entries in the training sets and then finds the nearest point (shortest distance) from the training dataset to the datasets as: TEST x is define in equation 3.2

$$D(x_{TEst}, x_m) = \sqrt{\sum_{m=1}^{M}(x_{TEST} - x_n)^2} \quad (1.1)$$

Where m=1, 2, 3, 4, 5…., M. The k-NN considers only the k nearest neighbors denoted as {X1,Kxk}as the member(s) of the set is define in equation 3.3.

$$kNNSpace = \{x_j \,|d(x, x_i) \leq d(x, x_j)\} \quad (1.2)$$

The k-NN rules involves classifying a test sample by assigning it to the most frequently represented among the k nearest samples.

4. EXPERIMENTAL DESIGN

Jupiter Notebook python language ANACONDANAVIGATOR which (includes 250+ popular data science packages and the *conda* package and virtual environment manager for Windows, Linux, and MacOS were used for the implementation of the k-NN model. The KNN model intends to classifier plant leafs based on which stores all cases, classify new cases based on similarity measure and chosen the best nearest neighbor (k) for the classification of the plant leafs.A flow diagram Graphical Description of a System's data & how the process transform the data.

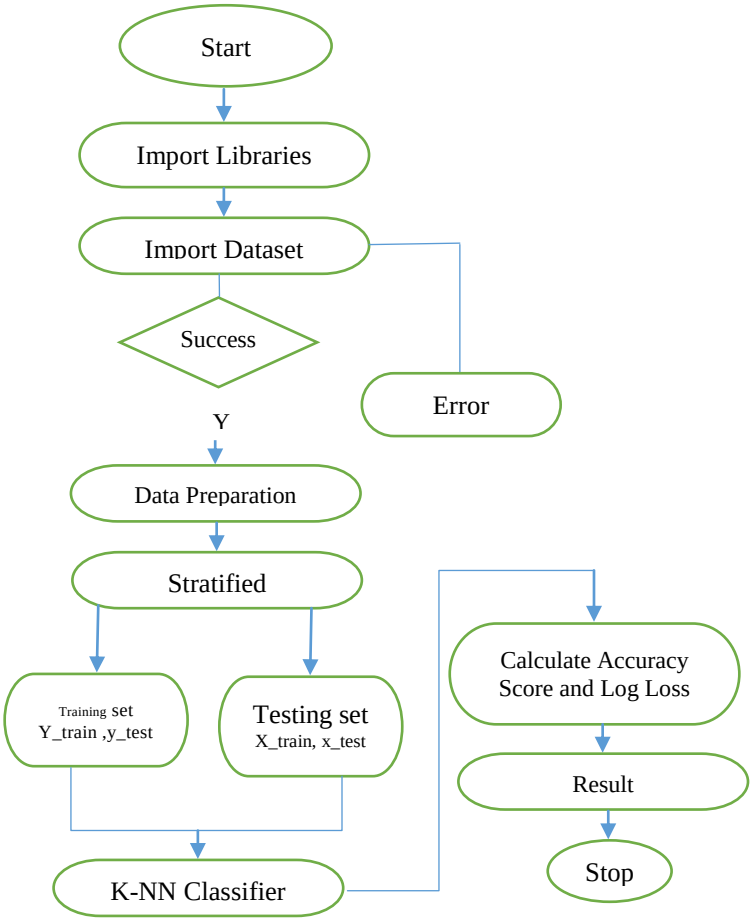


Figure 2: flow diagram of k-Nearest Neighbor Model

4.2 EVALUATION

To evaluate classification, Stratified K-Fold Cross Validation were used. Training and testing sets contains approximately the same percentage of samples of each target class as the complete set. . Finally, un-sampled data are allocated to the validation set. In this study, 60% of each data set was used for training, 20% for testing, and 20% for validation. These data splitting ratio were selected as it has been proven to be effective for many models.

4.3 CALCULATED MEASURES FOR THE EVALUATION

In order to effectively measure the classification accuracy of the proposed K-nearest neighbor model and the k-NN ensemble for classification of plant leaves, Metrics use for the performance of k-NN base learner and k-NN ensemble is measured and compared. For the evaluation phase, Accuracy Score and Log Loss metrics were used.

4.4 LOG LOSS

The classifier assigns a very small probability to the correct class then the corresponding contribution to the Log Loss will be very large indeed. Naturally this is going to have a significant impact on the overall Log Loss for the classifier.

4.5 ACCURACY SCORE

The evaluation phase is to compare the accuracy of the K-nearest neighbors (K-NN) using accuracy score, log loss and classification report for the proof of better performance of the model. The result visualization phase is to enable users see at a glance the classification accuracy of K-nearest neighbors (K-NN) and its evaluation summary.

5. RESULT PRESENTATION AND DISCUSSION

The result according to the k-NN based model and the ensemble of k-NN model and type metrics (Log Loss and Accuracy Score).Figure 3, presents the result of k-NN as a based learner model for k=3, 5, 7, 9, 11, 13, 15, 17, 19, 21. On the accuracy metrics. For accuracy score, the higher the value the better the result of the model. As observed for the result in figure 4, k-NN with value of k=3 i.e. 3 neighbors has the highest value. This showed us that the model had the best performance when 3 neighbor's distances are measured out once.

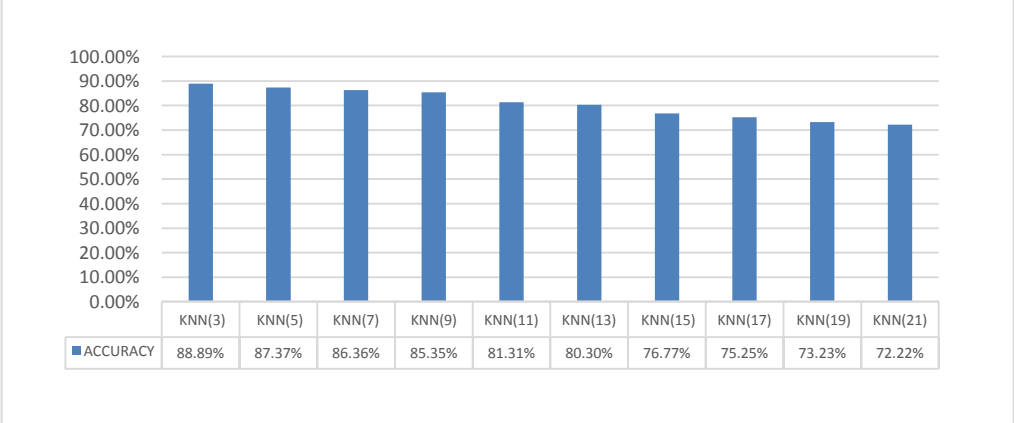


Figure 3: RESULT OF ACCURACY SCORE ON k-NN Based Model

RESULT ON LOG LOSS METRICS

Log loss metric value measures the lower probability of the model making the right prediction. The lower the value of log loss, the better the model. Figure 4 presents the result of k-NN as a based learner model for k=3, 5, 7, 9, 11, 13, 15, 17, 19, 21. On the Log Loss metrics, the k-NN model had the best result when k=7 i.e. 7 nearest neighbor.

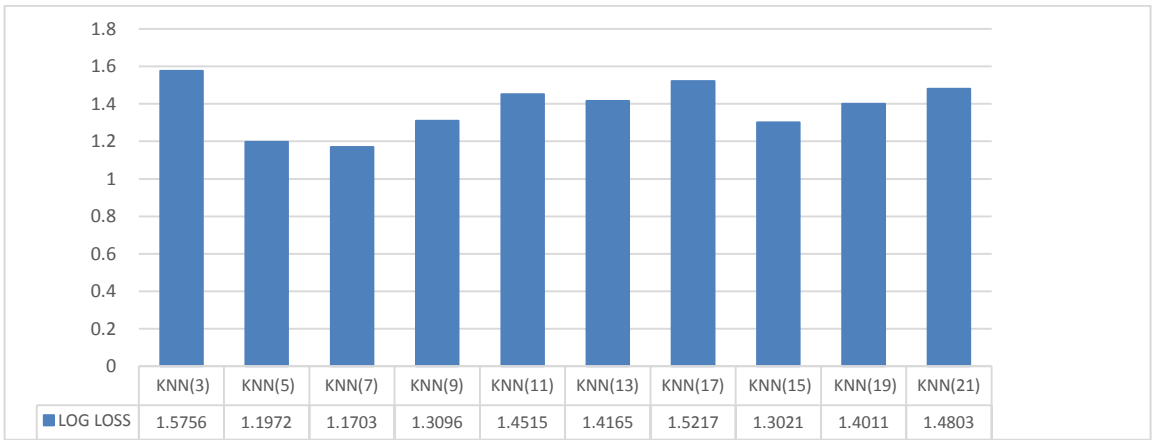


Figure 4: RESULT GENERATED BY LOG LOSS ON k-NN Model

RESULT ON ACCURACY SCORE METRICS

Figure 5: presents the result of k-NN as a based learner model for k=3, 5, 7, 9, 11, 13, 15, 17, 19, 21. On the accuracy metrics. For accuracy score, the higher the value the better the result of the model. As observed for the result in figure 4.6, k-NN with value of k=5 i.e. 17 neighbors has the highest value. This showed us that the model had the best performance when 17 neighbor's distances are measured out once.

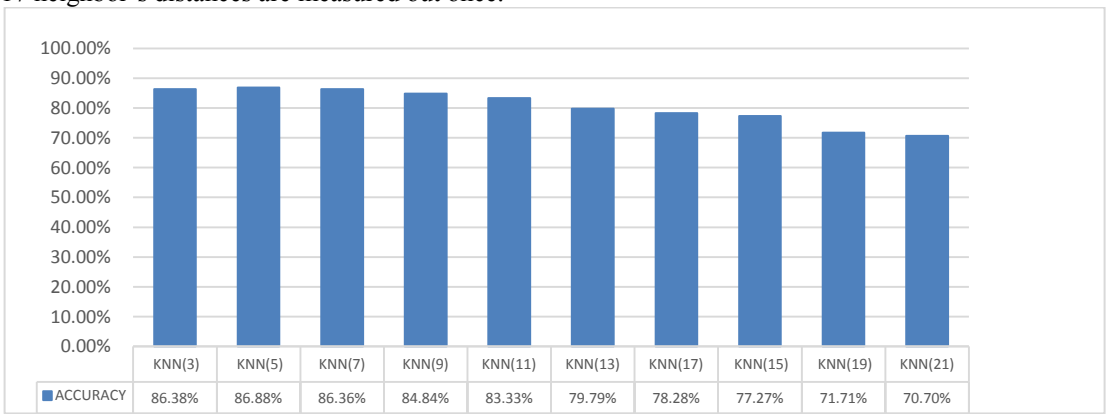


Figure 5:RESULT GENERATED BY THE ACCURACY ON k-NN ENSEMBLE MODEL

RESULT OF K-NN ENSEMBLE MODELS ON LOG LOSS METRICS

Log loss metric value measures the probability of the model making the right prediction. The lower the value of log loss, the better the model. Figure 6: presents the result of k-NN as a based learner model for k=3, 5, 7, 9, 11, 13, 15, 17, 19, 21. On the Log Loss metrics, the k-NN model had the best result when k=3 i.e. 7 nearest neighbor.

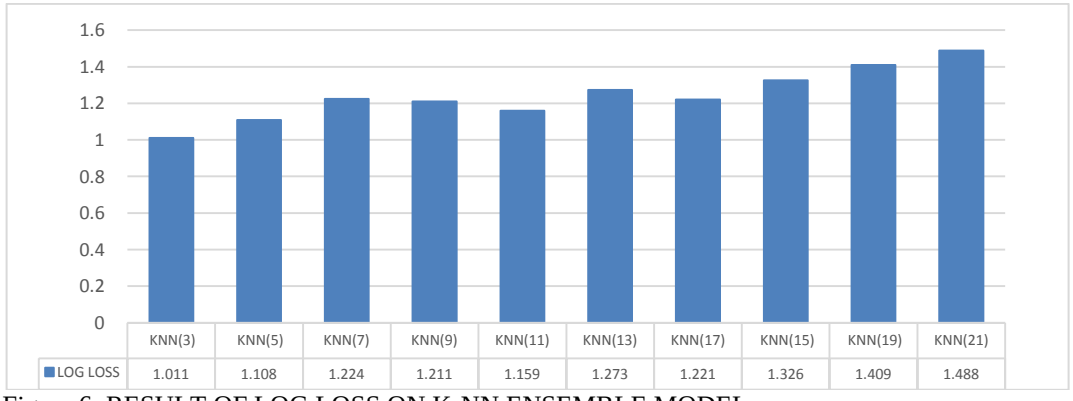
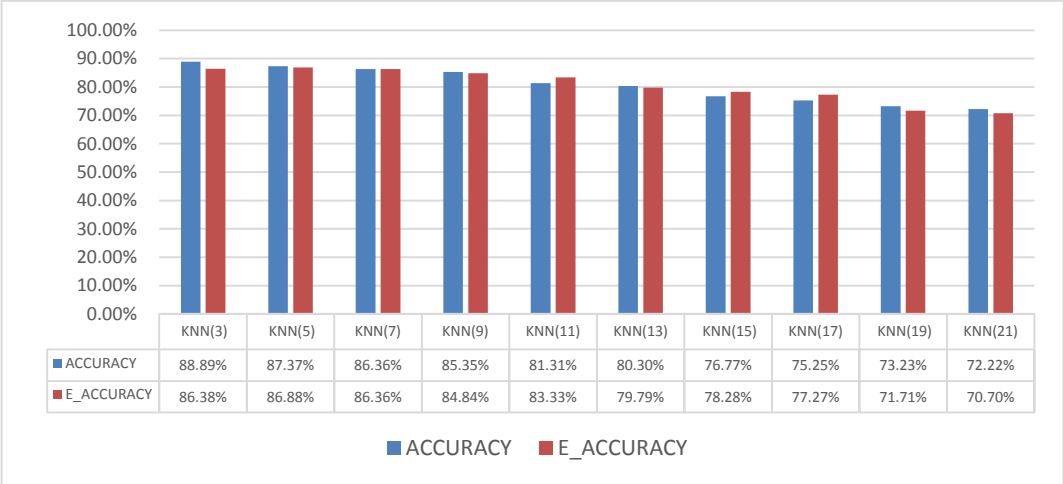


Figure 6: RESULT OF LOG LOSS ON K-NN ENSEMBLE MODEL

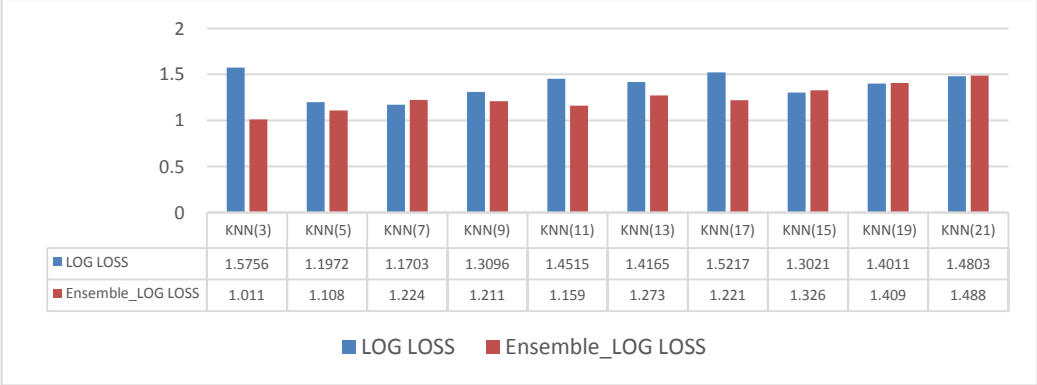
COMPARISON OF ACCURACY SCORE BETWEEN K-NN MODEL AND K-NN ENSSEMBLE MODEL

The result presented at this section is also observed that as the value of k is increasing the accuracy Log Loss of the model is decreasing.



	KNN(3)	KNN(5)	KNN(7)	KNN(9)	KNN(11)	KNN(13)	KNN(15)	KNN(17)	KNN(19)	KNN(21)
ACCURACY	88.89%	87.37%	86.36%	85.35%	81.31%	80.30%	76.77%	75.25%	73.23%	72.22%
E_ACCURACY	86.38%	86.88%	86.36%	84.84%	83.33%	79.79%	78.28%	77.27%	71.71%	70.70%

Figure 7:Comparison of Accuracy Score of K-NN Model and K-NN Model Ensemble



	KNN(3)	KNN(5)	KNN(7)	KNN(9)	KNN(11)	KNN(13)	KNN(17)	KNN(15)	KNN(19)	KNN(21)
LOG LOSS	1.5756	1.1972	1.1703	1.3096	1.4515	1.4165	1.5217	1.3021	1.4011	1.4803
Ensemble_LOG LOSS	1.011	1.108	1.224	1.211	1.159	1.273	1.221	1.326	1.409	1.488

Figure 8: Comparison of Log Loss Result of K-NN Model and K-NN Ensemble Model

6. CONCLUSION

The project aimed at building a machine learning model to classify plant species using plant leaves. The data set was binary plant leaves extracted features. k-NN machine learning model with k=3,5,7,9,11,3,15,17,19,21 was used. Also, a bagged ensemble of k-NN classifier was also built for the experiment. It was observed that k-NN based classifier with k=3 neighbors out perform all other models and its performance is better than all the neighbors of k-NN Ensemble. It had the highest accuracy value and also the lowest value for log loss. The result generated corroborated with the result of (Wilson, 1972) and (Folorunso & Adeyemo, 2017) saying that the asymptotic property of k-Nearest Neighbor is three. It is also observed that as the value of k is increasing the accuracy score is decreasing and Log Loss of the model is decreasing.

7. REFERENCE

[1]. Charles, M., James, C., & James, O. (2016). PLANT LEAF CLASSIFICATION USING PROBABILISTIC. *INTEGRATION*. *https://www.researchgate.net/publication/266632357, II(34)*, 1-2.

[2]. Cope, J.S., Remagnino, P., Barman, S., & Wilkin, P. (2010). Plant Texture classification using gabor co-occurences. *In Advances in Visual computing, Springer Berlin Heidelberg*, pp. 669-677.

[3]. Folorunso, S. O., & Adeyemo, A. A. (2017). An Emperical Experimental Survey of Application of Wilson's Edited Nearest Neighbor As a Sampling And Data Reduction Scheme to Alleviate Class Imbalance Problem. *Jounal of teh Nigeria Association of Mathematical Physics*, 239-248.

[4]. Homesciencetools. (2018, April 10). Retrieved from learning-center.homesciencetools.com: https://learning-center.homesciencetools.com/article/learn-about-leaves/

[5]. Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 234-244.

[6]. Kadir, A., Nugroho, L.E., Susanto, A., & Santosa, P. (2011). Neural Network Application in Forliage Plant Identification. *International Journal of Computer Applications (0975-8887),* pp. 15-22.

[7]. Kevin, M., & Bernhard , S. (2018, 06 12). *https://en.wikipedia.org/wiki/Machine_learning*. Retrieved from wikipedia: https://en.wikipedia.org/wiki/Machine_learning

[8]. Kevin, Z. (2018, 04 22). *Kevin Zakka's Blog*. Retrieved from Academic Journal: https://kevinzakka.github. io/2016/07/13/k-nearest-neighbor/

[9]. Manisha , A., Dr. Ramesh, R. M., Dr. Pravin , Y., & Dr.Ashok, T. G. (2016). Plant Classification Based on Leaf Features. *IBMRD's Journal of Management & Research*, 1-5.

[10]. Matthew, M. (2017). *kdnuggets*. Retrieved from seven-more-steps-machine-learning-python: https://www. kdnuggets. com/ 2017/ 03/seven-more-steps-machine-learning-python.html/2

[11]. Mohamed, E. R., & Abdelmalek, A. (2015). The First International Conference on Big Data, Small Data, Linked Data and Open Data. *AllData*, 31-34.

[12]. Nitin, B., & Vandana. (2010). Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security (IJCSIS)*, 302-305.

[13]. Noel , B. A. (2015). *KDnuggets*. (AYLIEN Text Analysis blog) Retrieved April 13, 2018, from kdnuggets: https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

[14]. Paplinski , A. P. (2018, 06 12). *neural_networks chapter 7*. Retrieved from byclb: https://www.byclb.com/TR/ Tutorials /neural_networks/ch7_1.htm

[15]. Parekh, A., Jyotismita , C., & Ranjan. (2012). Designing an Automated System for Plant Leaf Racognition. *nternational Journal of Advances inEngineering & Technology, volume 2*(Issue 1), pp. 149-158.

[16]. Rashad, M.Z., el-Desouky, B.S., & Manal, K. (2011). Plant images Classification Based on Textural Features. *International Journal of Computer Science & Informn Technology (IJCSIT),* 3(4), pp. 15-20.

[17]. Sumathi, C.S. & Senthil K.A. (2012). Edge and Texture Fusion for Plant Leaf Classification. *International Journal of Computer Science and Telecommunications*, 3(6), pp. 6-9.

[18]. Wilson, D. L. (1972). Asymptotic Properties of nearest neighbor Rules Edited Data. *IEEE Transaction on Systems, Man, and Cybernetics*, 408-421.

[19]. Zhang, S. & Chau, K.W. (2009). Dimension duction using semisupervised locally linear embedding for plant leaf classification. *Emerging Intelligent Computing Technology and Aplications, Springer Berlin Heidelberg,* pp. 958-955.