Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

AGARWOOD OIL QUALITY CLASSIFIER USING MACHINE LEARNING

M. A. Abas[1,*], N. S. A. Zubir[1], N. Ismail[1], I. M. Yassin[1], N. A. M. Ali[2], M. H. F.Rahiman[1], N. T. Saiful[3] and M. N. Taib[1]

[1]Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

[2]Natural Product Program, Forest Research Institute of Malaysia, Kepong, Selangor, Malaysia

[3]Faculty of Industrial and Science Technology, University Malaysia Pahang, 26300 Gambang, Pahang, Malaysia

Published online: 05 October 2017

ABSTRACT

Agarwood oil is known as one of the most expensive and precious oils being traded. It is widely used in traditional ceremonies and religious prayers. Its quality plays an important role on the market price that it can be traded. This paper proposes on a proper classification method of the agarwood oil quality using machine learning model k-nearest neighbour (k-NN). The chemical compounds of the agarwood oil from high and low quality are used to train and build the k-NN classifier model. Correlation-based feature selection was used to reduce the dimension of the data before it is being fed into the model. The results show a very high accuracy (100%) model trained and can be used to classify the agarwood oil quality accurately.

Keywords: agarwood oil; k-nearest neighbours; quality; machine learning.

Author Correspondence, e-mail: mohdaqib93@yahoo.com

doi: http://dx.doi.org/10.4314/jfas.v9i4s.4

1. INTRODUCTION

Agarwood oil is a precious and expensive oil that is extracted from agarwood. Nowadays, agarwood oil is widely traded internationally and the market demand for it is very huge especially from countries in the Middle East, Japan and Malaysia [1]. The agarwood oil is used as incenses, perfumes, traditional ceremonies and religious prayers [1-3]. The quality of the agarwood oil plays an important role in classifying the market price section that it will be sold as. A high quality agarwood oil can be sold for USD 126 to USD 633 per tola [4].

Machine learning is the science of programming computers, so they can learn from data. It is a field at the intersection of artificial intelligence, statistics and computer science. It shines when it is used to solve problems that are too complex for traditional approaches or when there is no known algorithm [5-6]. Machine learning algorithms that learn from input and output data are known as supervised learning algorithm. K-Nearest neighbours (k-NN) is one of the supervised learning algorithms that is widely used due to its simplicity and efficiency [7-10].

K-NN classification algorithm is arguably one of the most fundamental and simple classification methods. It is also known as lazy learner as it will delay abstracting from data until the model is asked to make a prediction, where later it will directly compare the queries with instances in the dataset. The k-nearest neighbours classification method represents the dataset in a feature space by taking each of descriptive features to be the axes of coordinate system. Then, each instance will be place within feature space based on values of its descriptive features. Distance metric is used in k-NN to measure the distance between two instances in the feature space. The prediction given by the model will be the target feature of instances (depends on the value of k) that is located nearest to the query in feature space.

Some of the strength of K-NN algorithm is that it is very easy to understand and often it will give a reasonable performance. The process of building the nearest neighbours is usually very fast but as the size of your dataset increase the prediction will get slower. The number of hyperparameter used to tune the k-NN model is also small compared to models like support vector machine and random forest classifier. Like any other machine learning algorithm, k-nearest neighbours also have its weakness such as the problem with curse of dimensionality

[5, 9].

Curse of dimensionality is the trade-off between the number of input features used in the dataset and density of instances in the feature space [9]. In curse of dimensionality problems, the higher number of input features, the lower the predictive classification power of the model. This happens because the predictive classification power depends highly on a reasonable sampling density of training instances used in the feature space.

Correlation-based feature selection is one of the techniques of feature selection. Feature selection is used to reduce the number of input features to the subset of input features that are most useful [5-6, 9]. Pearson correlation coefficient is a technique used to measure the linear correlation between two features. In feature selection, correlation coefficient is used to find the relationship that exists between each feature. If there is a strong relationship between features, it often means that those features are redundant and having both of them would not provide any extra information for the classifier model [9].

2. LITERATURE REVIEW

2.1. Agarwood Oil Quality

There has been numerous amount of literatures that covers on agarwood and its quality in past few years. These studies that has been carried out contribute a lot as a foundational knowledge on classifying agarwood oil using its chemical compounds. In [4], it presents a review on agarwood and its quality determination. It states that agarwood and its essential oil have different quality or grades and they are being traded in market according to the quality. It mentions that the chemical profiles of agarwood and its oil are key in differentiating between high and low quality as relying on physical appearances such as its color and odor have some drawbacks.

In [11], the author use correlation analysis between high grade and low grade Aquilaria Malaccensis (AM). There are 7 selected samples of AM identified using Gas Chromatograph-Mass Spectra (GC-MS) and the samples are being separated into two groups which are high grade and low grade. The result of the correlation test shows that between similar grade there are high correlation between the compound. However, comparing between two different grades of high and low, there are no significant correlation that exist between those two grades.

In [12], the author presents a classification technique of Z-score to discriminate between high and low quality of agarwood oil. The author uses 6 samples that has been analyzed with GC-MS to examine their chemical profiles. The Z-score technique is used to determine the significant compound that is inside those samples as the extraction gave at least 43 volatile compound presents on the agarwood oil. It is found that there are 6 significant chemical compound that is used to successfully discriminate between high and low quality agarwood oils which are β-agarofuran, α-agarofuran, 10-epi-γ-eudesmol, α-eudesmol, dihydrocollumellarin and γ-eudesmol.

2.2. Machine Learning in Agriculture Field

Machine leaning technique also has been implemented in agriculture domain before. Some of the literatures are [13-16]. In [13], it carried out a research on predicting the yield of crops by using machine learning models. The correlation between past environmental patterns and crop production rate is used to train the models. The data was taken for duration of 18 years from 1992 to 2010. Then, the models were compared with each other to measure the effectiveness on an unknown climate variable.

In [14], it proposed a technique to identify plant disease by using pattern recognition and leaf texture analysis. The system that is used required single leaf as an input to be segmented, analyzed and find the texture model. Then, it uses multiclass SVM technique to classify the extracted pattern. The machine learning model produce an accurate result on classifying the grape leaves into health and diseased. The results shown reach up to 96.66% by increasing the training-testing ratio to 100:30.

In [15], also uses machine learning technique to make an early detection of grapes diseases. However, the author combines machine learning technique with IoT to help the farmers and experts. It uses Hidden Markov Model and then provides and alert system to the farmer via SMS. The algorithm used in the machine learning model takes the input from temperature, moisture sensors and humidity. The results shown using Hidden Markov Model is more accurate and precise than using statistical model.

In [16], the author uses k-Nearest Neighbor (k-NN) approach to grade the quality of agarwood oil. The k-Nearest Neighbor is chosen because of its simplicity and efficiency. The author separates the dataset with 80% being allocated as training data and 20% being allocated as the testing data. The Euclidean distance metric is used in this k-NN implementation. The result shows that using k = 1 and k = 3 gives the best accuracy for training and testing, which is 87.5% and 83.33% respectively. The accuracy of training and testing decreases as the number of k increase and at k = 5 the accuracy of training data drops to 77.08% and accuracy of testing data drops to 75.00%.

3. METHODOLOGY

The whole process of building, training and evaluating the k-NN classifier model are done using anaconda software, which contains Python programming language and all its necessary data science modules.

The data used in this experiment is collected from agarwood oil samples, species Aquilaria Malaccensis which are obtained from Forest Research Institute Malaysia (FRIM), Kepong, Selangor, Malaysia and Faculty of Industrial Sciences and Technology, Universiti Malaysia Pahang (UMP). The collection process of the agarwood oil data is first by extracting the agarwood oils using GC-MS. GC-MS analysis is done to extract the chemical compounds in the agarwood oil of two qualities which are high and low quality. Fig. 1 shows the overall methodology used to in this experiment.

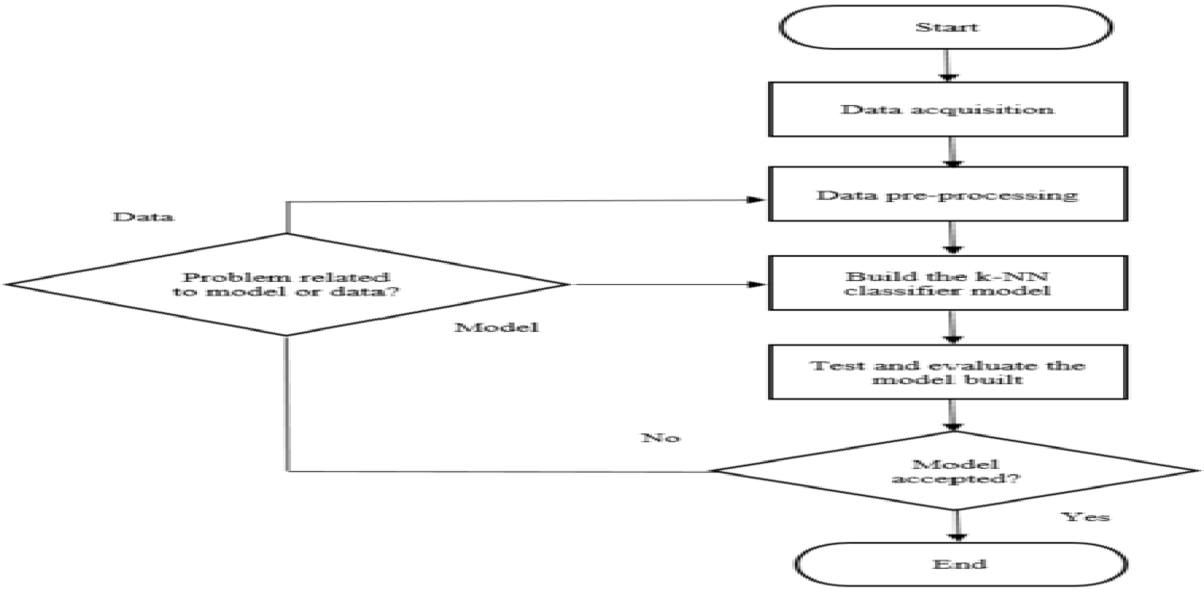

Start
Data acquisition
Data pre-processing
Build the k-NN classifier model
Test and evaluate the model built
Model accepted?
Yes
End
No
Problem related to model or data?
Data
Model

Fig.1. Flowchart of overall experiment

3.1. Data Acquisition

Data is acquired in this stage. The dataset used to train the k-NN classifier model consists of 117 samples of agarwood oil with 7 input features (the chemical compounds of sample) and 1 output features (quality of the sample).

3.2. Data Pre-Processing

Data pre-processing done for this experiment consists of correlation-based feature selection, min-max feature scaling and stratified k-fold cross validation technique for data splitting method. The number of fold used for the cross-validation technique will be set to 10 in this experiment. The dataset must undergo data pre-processing technique to ensure the data used can provide most information and increase the predictive power of the classifier model that will be build.

3.2.1. Correlation-Based Feature Selection

The Pearson correlation coefficient which is used in correlation-based feature selection is given by

$$\rho_{x,y} = \frac{\mathrm{cov}\,(x,y)}{\sigma_x \sigma_y} \qquad (1)$$

Based on the formula, Pearson correlation coefficient ($\rho_{x,y)}$ is the covariance (cov) between two features x and y which is standardized by their standard deviations (σ). The covariance can be computed as

$$\mathrm{cov}(X, Y) = \sum_{i=1}^{N} \frac{1}{N} (x_i - \mu_x)(y_i - \mu_y)$$

(2)

The formula for standard deviation is given as

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_n - \mu)^2} \tag{3}$$

The mean (μ) is given as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_n \tag{4}$$

The end product of the dataset after applying feature selection technique would only to include input features that are predictive and interacting, while discarding any redundant and irrelevant features.

3.2.2. Min-Max Scaling (Feature Scaling)

Feature scaling is used to rescale the values of features in the dataset. Some machine learning models are sensitive to the values of data used. Hence, feature scaling is used so that the data representation is more suitable for the machine learning algorithm.

Min-max scaling technique (also called normalization) is used to rescale the data for the values to fall in a specific range (usually between 0 to 1 or -1 to 1), while at the same time still maintains the relative difference that exist between each feature. Formula for min-max scaling is given as [6, 9]:

$$x_{new} = a + \frac{x_i - \min(x)}{\max(x) - \min(x)}(b - a) \tag{5}$$

where x_i is the value of x at i^{th} instance, a and b are the points to rescale. For this experiment, the range of min-max scaling is between 0 to 1, thus a = 0 and b = 1. Min-max scaling is usually used if the value in features need to be in a specific range for calculation and computation.

3.2.3. Stratified k-Fold Cross-Validation

Cross-validation technique is one of the data splitting technique used to get a better evaluation than the normal hold-out test set [5-6, 9]. In k-fold cross validation, the available data is being partitioned into k-equal sized folds. Then evaluation of the model is performed k times. The ‘stratified’ means that each set in the fold contains approximately the same percentage of samples of each class that exist in the dataset as the complete set. Usually, stratified k-fold cross validation is preferred to normal k-fold cross validation because it results in more reliable estimation of the generalization performance, especially when used to split dataset that is not balance [6].

The process of evaluation for cross validation is generally considered as more stable and thorough than using the normal hold-out test set. It also solves the possibilities of having ‘lucky split’ of data during randomly splitting in hold-out test set, where ‘easy’ data is randomly put into test set, and thus making the prediction score unrealistically higher than it supposed to. Even though cross validation method is considered better than hold-out test set, it is very computational hungry as it will train k model instead of only one model in hold-out test set [6, 9].

3.3. Build the k-NN Classifier Model

For this experiment, the k-NN classifier model is built using Euclidean distance metric calculation and the number of neighbours is varied between 1 to 20 to detect the appropriate number of neighbours to be used for the model to get the best scores and results. The Euclidean distance is defined as [9]:

$$\mathrm{Euclidean}\ (a, b) = \sqrt{\sum_{i=1}^{m} (a[i] - b[i])^2}$$

(6)

The Euclidean distance involves more computation of squaring and square root and this would be an important aspect to look when the datasets for the model is very large. Euclidean distance is often used as default distance metrics in k-NN [9-10].

3.4. Test and Evaluate the Model

Model built was tested using the classification accuracy score and basic confusion matrix-based performance measures. Measurements such as precision, recall and F_1 measure would also be used to get more granular and precise results on the model built.

Classification accuracy indicates how well the model actually performs in classifying the output. The formula for classification accuracy score is given as [9]:

$$\mathrm{accuracy} = \frac{\mathrm{TP+TN}}{(\mathrm{TP+TN+FP+FN})} \tag{7}$$

The formula for precision (8), recall (9) and F_1 measure (10) is given as follows

$$\mathrm{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP+FP})} \tag{8}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{(\mathrm{TP+FN})} \tag{9}$$

$$F_1\text{measure} = 2 \times \frac{(\text{precision}\times\text{recall})}{(\text{precision}+\text{recall})} \tag{10}$$

4. RESULTS AND DISCUSSION

The results for the experiment carried out is presented in this section.

4.1. Correlation-Based Feature Selection

Fig. 2 shows the heatmap for the Pearson correlation coefficient computed for all the input features in the dataset. From here, it can be seen that input features 5, 6 and 7 are strongly correlated. In classifying the agarwood oil using the chemical compound, it is assumed that having all these 3 features in the dataset would not provide any additional information for the k-NN classifier model and at the same time, having all 3 features would increase the problems of curse of dimensionality. Hence, only feature 5 is chosen and features 6 and 7 are discarded from being used.

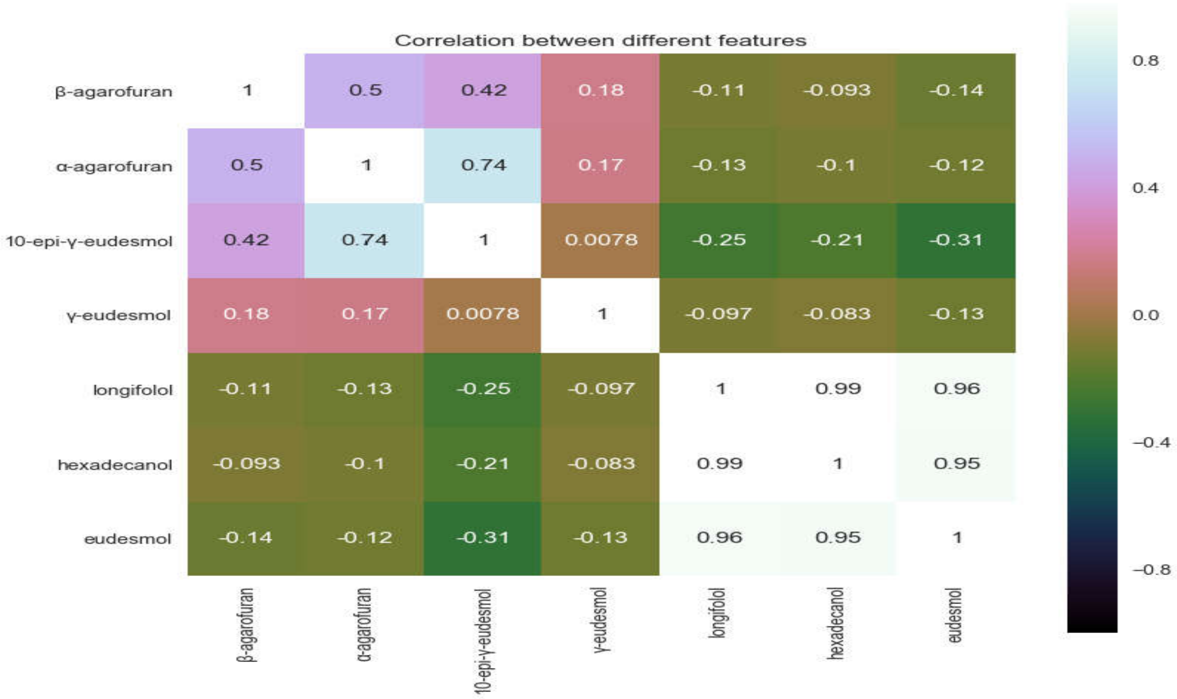


Fig.2. Heatmap of correlation-based feature selection

4.2. Min-Max Scaling (Feature Scaling)

Fig. 3(a) shows the boxplot for distribution of the original dataset, Fig. 3(b) shows the boxplot for distribution of the dataset after undergone min-max scaling technique and Fig. 3(c) shows the dataset information before and after applying the min-max scaling technique. From these figures, it can be seen that for original dataset, the values of each input features in the dataset varies widely with feature 3 having the widest range of values up to 21.45 while feature 1 and feature 2 only have values up to 6.97 and 3.21 respectively. It can also be seen that feature 5 contains some outliers in it. From Fig. 3(b), we can see the distribution of values after applying min-max scaling have the same range which is between 0 to 1. This would help in training a better k-NN classifier model as it will reduce the sparsity in the feature space.

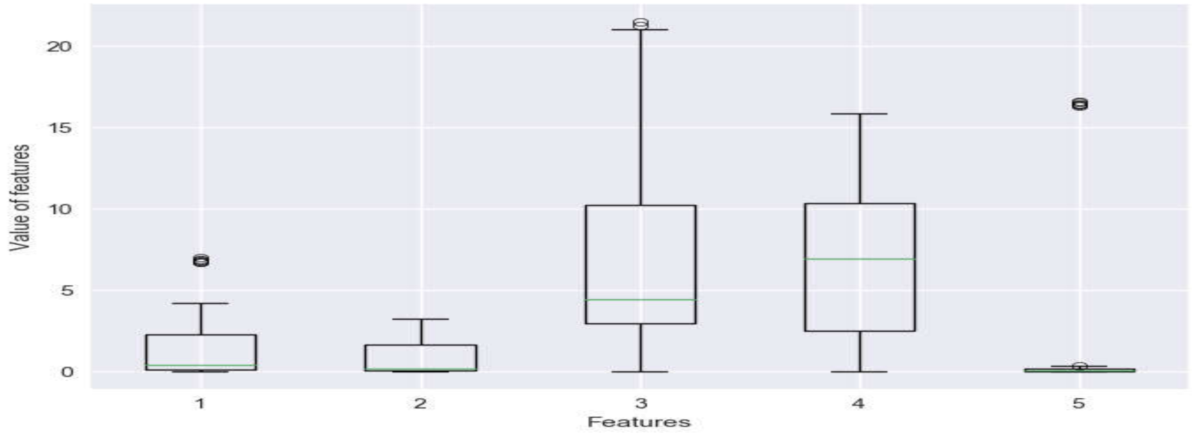


Fig.3. (a) Boxplot of original data distribution

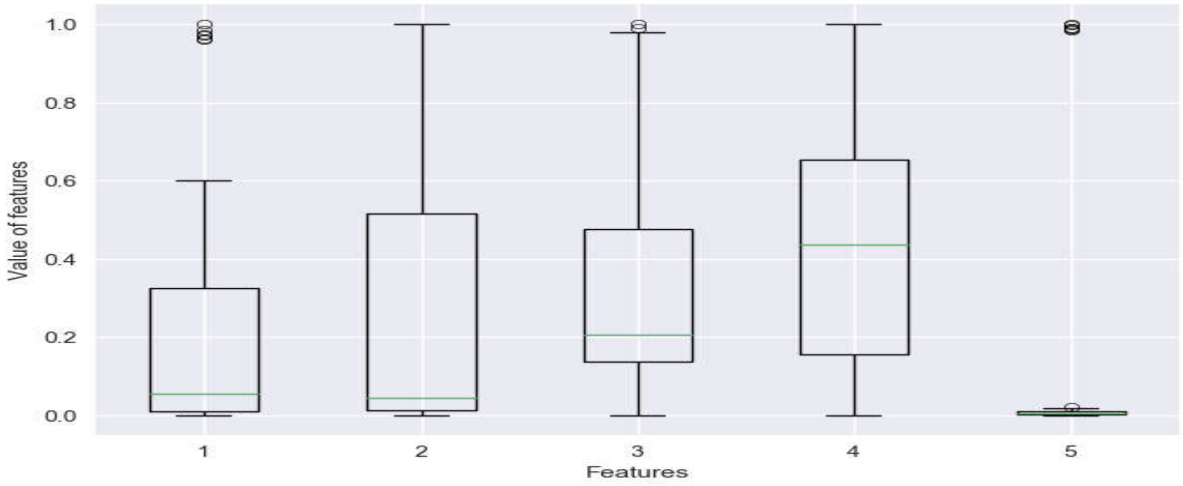
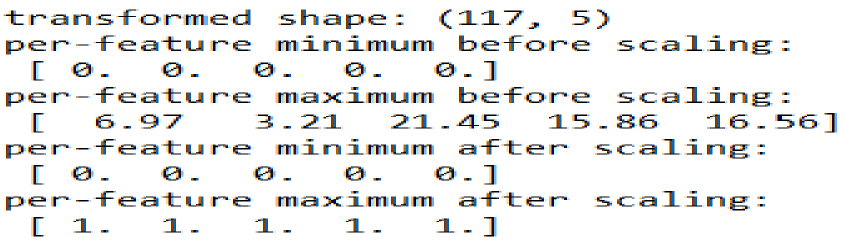


Fig.3. (b) Boxplot of data distribution after min-max scaling



```
transformed shape: (117, 5)
per-feature minimum before scaling:
 [ 0.   0.   0.   0.   0.]
per-feature maximum before scaling:
 [   6.97    3.21   21.45   15.86   16.56]
per-feature minimum after scaling:
 [ 0.   0.   0.   0.   0.]
per-feature maximum after scaling:
 [ 1.   1.   1.   1.   1.]
```

Fig.3. (c) Dataset information before and after applying min-max scaling

4.3. Evaluation of the k-NN Classifier Model

Fig. 4 shows the graph plot for training accuracy and testing accuracy of the model. Overall from the graph plot, it shows the classifier model used with stratified k-fold cross validation splitting method are very stable to predict the agarwood oil quality with classifying capabilities of 100% for number of neighbours used between 1 to 9 and more than 89% throughout the whole range of number of neighbours used in this experiment.

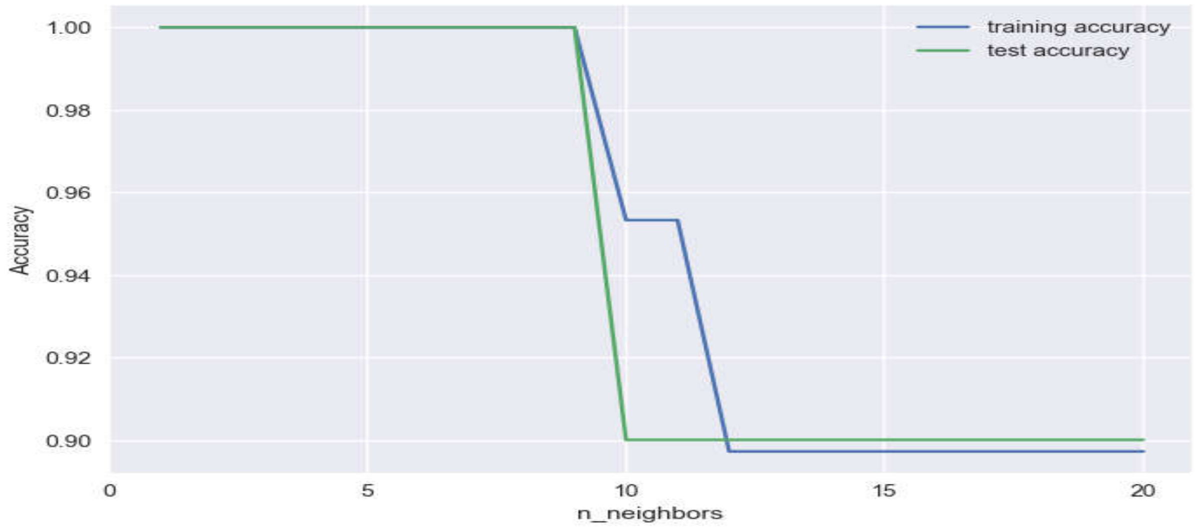
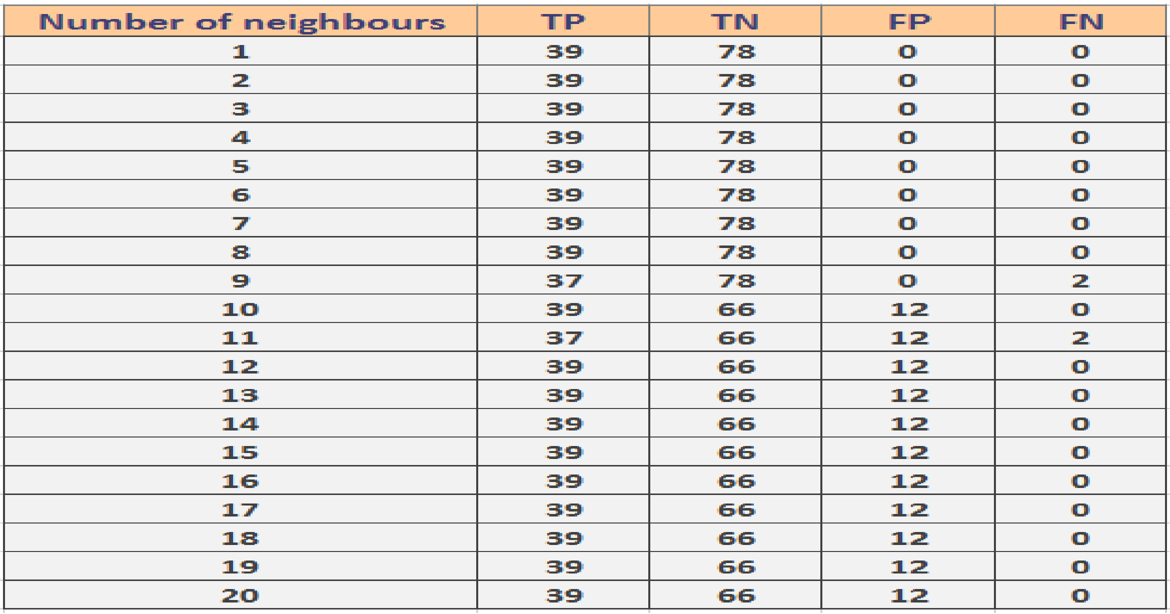


Fig.4. Graph plot for training accuracy and testing accuracy score

Table 1 shows the results for confusion matrix done on the testing set for k = 1 to k = 20. Notice that the testing set includes all of 117 instances used when using stratified k-fold cross-validation split. This is because for each iteration of fold, different subset of the dataset is used as training and testing. After all 10 iterations are completed, each of the instances will be completely used as testing set. Low quality agarwood oil is being set as the positive class (39 instances) and high quality agarwood oil are being set as negative class (78 instances). The classification made are perfect from number of neighbours used between 1 to 8 then the results drop as the number of neighbours used increases. Notice that from Fig. 4, even though at k = 9, the accuracy is perfect. But when looking at Table 1, there are 2 instances of low quality are being wrongly predicted as high quality. This is because the accuracy (in Fig. 4) is being calculated as the average accuracy result of each iteration performed, while the confusion matrix result is based on all the 117 instances used as testing set in 10 iterations.

Table 1. Confusion matrix score for testing set



Number of neighbours	TP	TN	FP	FN
1	39	78	0	0
2	39	78	0	0
3	39	78	0	0
4	39	78	0	0
5	39	78	0	0
6	39	78	0	0
7	39	78	0	0
8	39	78	0	0
9	37	78	0	2
10	39	66	12	0
11	37	66	12	2
12	39	66	12	0
13	39	66	12	0
14	39	66	12	0
15	39	66	12	0
16	39	66	12	0
17	39	66	12	0
18	39	66	12	0
19	39	66	12	0
20	39	66	12	0

Fig. 5 shows the graph plot for precision, recall and F$_1$ measure used and the positive class used in the calculation is set for low quality. Table 2 provides a more detailed information by having both high quality and low quality set for positive class. Based on the figures and table, the precision, recall and F$_1$ measure score for both high quality and low quality when k = 1 to k = 8 shows a perfect score. This means the classifier model are able to classify accurately when using those range of numbers of neighbours. As the number of neighbours used increase more than 8, the scores decreases. However, for precision when high quality is used as positive class has perfect score except for on k = 9 and k = 11 which indicates there are mistakes of classifying low quality as high quality. Recall score reading when low quality is used as positive class also has perfect score except for k = 9 and k = 11 because of the same reason as above.

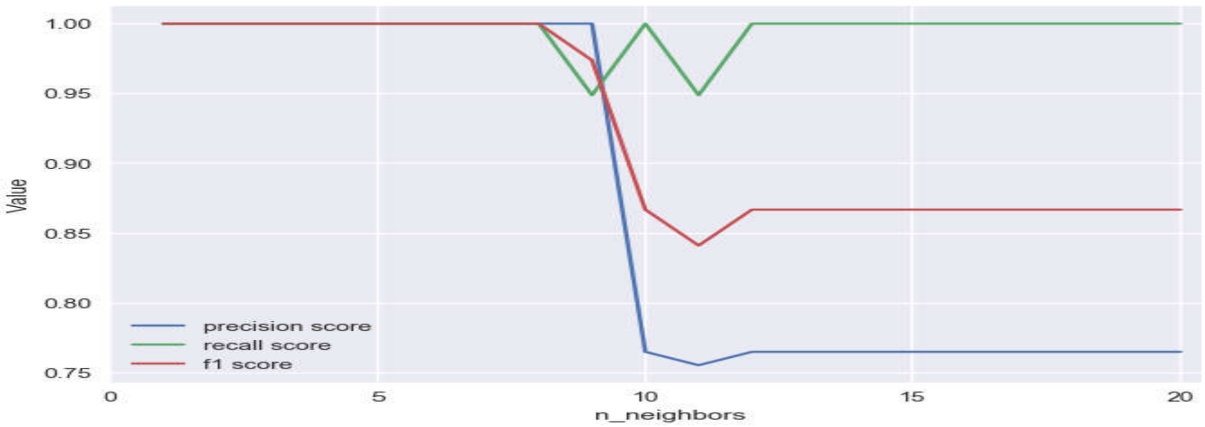
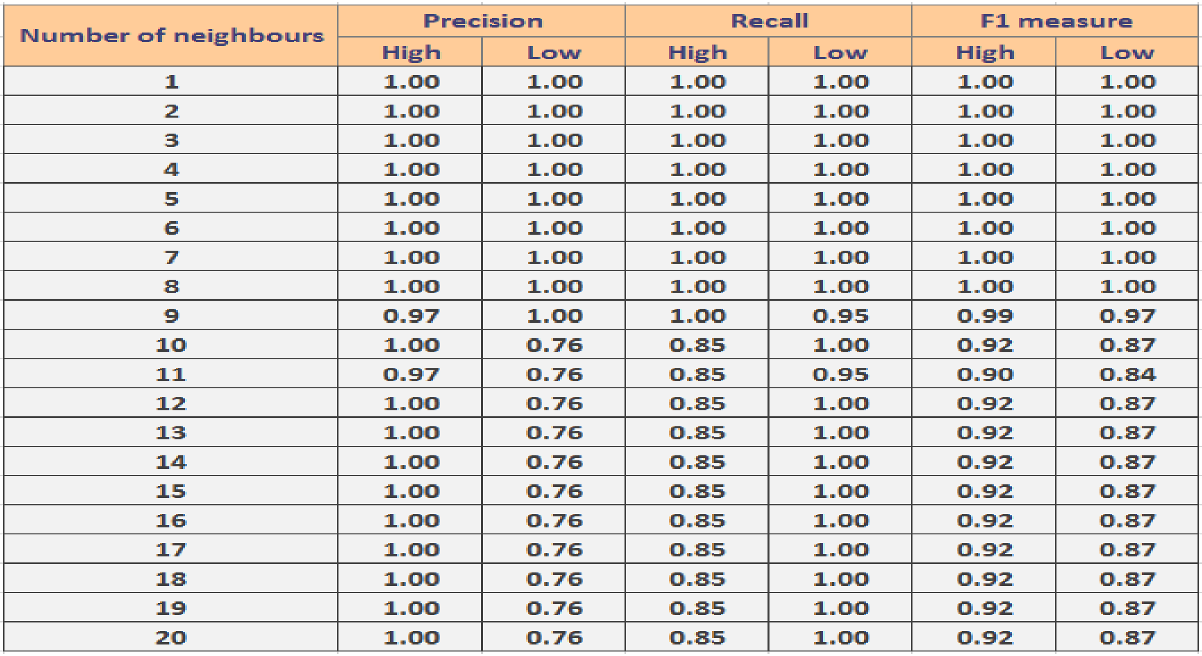


Fig.5. Graph plot for precision, recall and F$_1$ measure score

Table 2. Precision, recall and F_1 measure score



Number of neighbours	Precision		Recall		F1 measure	
	High	Low	High	Low	High	Low
1	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00
9	0.97	1.00	1.00	0.95	0.99	0.97
10	1.00	0.76	0.85	1.00	0.92	0.87
11	0.97	0.76	0.85	0.95	0.90	0.84
12	1.00	0.76	0.85	1.00	0.92	0.87
13	1.00	0.76	0.85	1.00	0.92	0.87
14	1.00	0.76	0.85	1.00	0.92	0.87
15	1.00	0.76	0.85	1.00	0.92	0.87
16	1.00	0.76	0.85	1.00	0.92	0.87
17	1.00	0.76	0.85	1.00	0.92	0.87
18	1.00	0.76	0.85	1.00	0.92	0.87
19	1.00	0.76	0.85	1.00	0.92	0.87
20	1.00	0.76	0.85	1.00	0.92	0.87

5. CONCLUSION

The agarwood oil quality classifier model using k-NN is successfully built. Based on the score of the accuracy and results of performance measure, it can be concluded that the k-NN classifier model using number of neighbours of 1 to 8 provide the best results for classification of agarwood oil quality with 100% classification accuracy score. The computation efficiency and accuracy score obtained from this paper is higher when compared to the previous paper of classifying agarwood oil using k-NN method in [16].

6. ACKNOWLEDGEMENTS

The authors would like to express my sincere appreciation and thanks to FRIM and UMP for the data collection that is being used to build the model for this agarwood oil classification system. The authors would also like to thank Faculty of Electrical Engineering and Institute of Research Management and Innovation (IRMI), Universiti Teknologi MARA (UiTM) especially grant reference code 600-IRMI/MyRA 5/3/LESTARI (0040/2016) for financial support in this project. The authors would like to graciously acknowledge the Ministry of Higher Education and Universiti Teknologi MARA for supporting this research work through Grant No: 600-RMI/DANA 5/3/ARAS(4/2015).

7. REFERENCES

[1] NorAzah M A, Majid J A, Mailina J, Said A A, Husni S S, Hasnida H N, Arip M A, Chang Y S. Profiles of selected supreme Agarwood oils from Malaysia. In Herbal Globalisation: A New Paradigm for Malaysian Herbal Industry, 2008, pp. 393-398

[2] Hidayat W, Shakaff A Y, Ahmad M N, Adom A H. Classification of agarwood oil using an electronic nose. Sensors, 2010, 10(5):4675-4685

[3] Ishihara M, Tsuneya T, Uneyama K. Fragrant sesquiterpenes from agarwood. Phytochemistry, 1993, 33(5):1147-1155

[4] Lias S, Ali N A, Jamil M, Zainal M H, Ab Ghani S H. Classification of pure and mixture Agarwood oils by Electronic Nose and Discriminant Factorial Analysis (DFA). In IEEE International Conference on Smart Sensors and Application, 2015, pp. 7-10

[5] Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow. California: O'Reilly Media Inc., 2017

[6] Müller A. C., Guido S. Introduction to machine learning with Python. California: O'Reilly Media, 2016

[7] Tan S. An effective refinement strategy for KNN text classifier. Expert Systems with Applications, 2006, 30(2):290-298

[8] Zhang N, Yang J, Qian J J. Component-based global k-NN classifier for small sample size problems. Pattern Recognition Letters, 2012, 33(13):1689-1694

[9] Kelleher J. D., Mac Namee B., D'Arcy A. Fundamentals of machine learning for predictive data analytics. Massachusetts: The MIT Press, 2015

[10] Brink H., Richards J. W., Fetherolf M. Real-world machine learning. Connecticut: Manning Publications Co., 2016

[11] Ismail N, Rahiman M H, Jailani R, Taib M N, Ali N A, Tajuddin S N. Investigation of common compounds in high grade and low grade Aquilaria Malaccensis using correlation analysis. In IEEE Control and System Graduate Research Colloquium, 2012, pp. 277-281).

[12] Ismail N, Rahiman M H, Taib M N, Ali N A, Jamil M, Tajuddin S N. Classification of the quality of agarwood oils from Malaysia using Z-score technique. In IEEE 3rd International Conference on System Engineering and Technology, 2013, pp. 78-82

[13] Rahman M M, Haq N, Rahman R M. Machine learning facilitated rice prediction in Bangladesh. In IEEE Annual Global Online Conference on Information and Computer Technology, 2014, pp. 1-4

[14] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1):1-47

[15] Patil S S, Thorat S A. Early detection of grapes diseases using machine learning and IoT. In 2nd IEEE International Conference on Cognitive Computing and Information Processing, 2016, pp. 1-5

[16] Ismail N, Rahiman M H, Taib M N, Ali N A, Jamil M, Tajuddin S N. The grading of agarwood oil quality using k-Nearest Neighbor (k-NN). In IEEE Conference on Systems, Process and Control, 2013, pp. 1-5

How to cite this article:
Abas MA, Zubir NSA, Ismail N, Yassin IM, Ali NAM, Rahiman MHF, Saiful NT, Taib MN. agarwood oil quality classifier using machine learning. J. Fundam. Appl. Sci., 2017, *9(4S), 62-76.*