Classification of Stingless Bee Honey

Muhammad Zariff Wafiy, Nor Liyana Mohd Shuib
Department of Information Systems, Faculty of Computer Science and Information Technology,
Universiti Malaya, Kuala Lumpur, Malaysia
17204013@siswa.um.edu.my

***Abstract*—Stingless bee honeys are one of the most prominent biological consumables due to its medicinal properties which includes anti-inflammatory, antibacterial, and antimicrobial qualities. Due to the abundance of species of stingless bee across the world where each species produces different quality of honey, it is difficult for potential consumers to choose which type of honey is the most suitable for them to purchase and utilize based on their needs or current health condition. Honeys from different species differ based on physicochemical properties and antioxidant activities. The objective of this project is to help consumers as well as researchers to have an insight on which species is a particular sample from. A machine learning approach is implemented onto a multiclass classification problem to classify samples of stingless bee honey based on different attributes including chemical composition and antioxidant assays. The data used in this project is obtained from the Centre for Natural Products and Drug Discovery, Universiti Malaya. Random Forest is chosen as the main model after evaluating classifiers' performances based on accuracy though, the accuracy of Random Forest Classifier is sub par at best. Hence, the classification of stingless bee honey should be taken as a rough indicator on which species the particular sample belongs to.**

***Keywords*—Biological consumables; Stingless Bee Honey; Machine Learning; Multiclass Classification; Random Forest**

1. INTRODUCTION

The stingless bee is a type of bee which can be found on practically every continent across the globe (Abd Jalil, 2017). This bee produces honey, which has been extensively utilized across time and space. According to Fletcher (2017), most tropical and subtropical areas are home to stingless bees (Meliponini), which number over 500 species and are dispersed across the Neotropical, Afrotropical, and Indo-Australian regions. . Stingless bees are crucial to the environment, the economy, and cultures; they serve as the primary pollinators for a variety of tropical wild and domesticated plants. For many years, people have exploited their products, such as honey, pollen, and cerumen, as a means of revenue. In rural areas of America, indigenous people, particularly the Mayan, are culturally attached to stingless bees. According to a research officer in the Malaysian Agricultural Research and Development Institute (MARDI), due to their miniscule size, stingless bees are unique in that they can pollinate small flowers, something that the comparative large honey bee cannot achieve.

Honey produced from stingless bees *Meliponini* are far different from the ones produced by regular honey bees *A.mellifera* mainly due to its distinctive feature of being preserved naturally in the cerumen. This contributes to its medicinal effects, particularly in the healing of wounds (Abd Jalil, 2017). In fact, it is said to have superior nutritional and therapeutic benefits compared to *A.mellifera* honey (Shamsudin et

al., 2019). Although stingless bee honey has long been regarded as a premium functional food, its purported therapeutic benefits have not been linked to any particular bioactive ingredients (Fletcher, 2020).

For the past few decades, *A.mellifera* has been the subject of the majority of studies on honey due to the species' global adaptability, however, the literature extols the merits of the many qualities of stingless bee honey, particularly in regard to its moisture content, distinctive flavor, and more overt scent (Alves et al. 2005). According to Fletcher (2017), despite the presence of numerous studies on the physicochemical and nutritional makeup of stingless bee honey have been made, there are only a few bioactive components that have been identified. While all of these investigations concluded that the composition of stingless bee honey is different from that of regular bee honey, no thorough identification of the main components and possible therapeutically active compounds has been reported. There are also the emergence of physicochemical analyses pertaining to the quality of stingless bee honey, however, these analyses do not give insight about the geographical origin of the honey which is due to the consistency of stingless bee honey's metabolite composition which could not be reproduced (Razali, 2018).

The development in the field of Artificial Intelligence allows the usage of classification models through the approach of Machine Learning for differentiating various labels based on certain attributes and variables. This study aims to utilize the advanced method of developing machine learning classifiers to classify different species of stingless bee provided different honey samples based on physicochemical properties as well as antioxidant activities by implementing the approach of multiclass classification. In addition, This study would also identify the correlation between chemical compounds and antioxidant activities, as well as to build an interactive dashboard for model deployment.

2. LITERATURE REVIEW

This research aims to provide an alternative method of machine learning to classify samples of stingless bee honey based on its attributes. This approach is different from the regular method of using chemical tests to identify which samples belong to which species of stingless bee.

The concept of classifying stingless bee honeys is common in the field of study of chemistry, but not so much in the computer science domain.
A research conducted suggests that classification of raw stingless bee honeys by bee species origins can be done by using the NMR- and LC-MS-Based metabolomics approach (Razali, 2018). Proton nuclear magnetic resonance (1H-NMR) spectroscopy, an ultra-high performance liquid chromatography-quadrupole time of flight mass spectrometry (UHPLC-QTOF MS), and chemometrics methods were used to analyze and categorize raw stingless bee honeys according to the origins of the bee species.

In accordance with the origins of the bee species, Heterotrigona itama, Geniotrigona thoracica, and Tetrigona apicalis, the honey samples could be divided into three distinct groups. As the discriminant metabolites or possible chemical markers, d-Fructofuranose (*H. itama* honey), -d-Glucose, d-Xylose, -d-Glucose, and l-Lactic acid, Acetic acid, and l-Alanine (*T. apicalis* honey) were all detected. d-Fructofuranose was identified by 1H-NMR data. It may be argued that the classification method by bee species origins using the UHPLC-QTOF MS-based metabolomics methodology can quickly identify

the quality of honey in terms of originality and purity (Razali, 2018).

Furthermore, a study titled *Entomological authentication of stingless bee honey by 1H NMR-based metabolomics approach* implements the utilization of 1H NMR data. Using 1H NMR data generated from both water dilution and chloroform extract spectra, orthogonal projection to latent structure-discriminant analysis (OPLS-DA) enabled successful entomological classification (Zuccato et al., 2017).

Another research conducted suggests a comparative characterization of physicochemical and antioxidants properties of processed *Heterotrigona itama* from different origins and classification by chemometrics analysis (Shamsudin, 2019). In this research, stingless bee honey produced by *Heterotrigona itama* from different botanical origins was characterized and discriminated. Three types of stingless bee honey collected from acacia, gelam, and starfruit nectars were analyzed and compared with *Apis mellifera* honey.

The results of this research showed that stingless bee honey samples from the three different botanical origins were significantly different in terms of their moisture content, pH, free acidity, total soluble solids, color characteristics, sugar content, amino acid content and antioxidant properties. Stingless bee honey was significantly different from *Apis mellifera* honey in terms of physicochemical and antioxidant properties (Shamsudin, 2019).

In addition, there is also research regarding the classification of stingless bee honey based on species, dehumidification process and geographical origins using physicochemical and ATR-FTIR chemometric approach (Ismail, 2021). Using chemometric methods, this study examined the physicochemical characteristics and chemical profiles of Malaysian stingless bee honey from various species, the dehumidification procedure, and geographical origins. A total of 122 samples from various species were gathered across Malaysia's several states. Each of the investigated samples' physicochemical attributes fell within the acceptable limits set by the Malaysian standard.

The chemometric analyses demonstrated the validity of physicochemical characteristics and ATR-FTIR to distinguish stingless bee honey based on the dehumidification procedure and geographic origins but not based on species. The results suggest that, in addition to the current physicochemical study, the categorization and quality control of Malaysian stingless bee honey may also benefit from the application of an ATR-FTIR-based chemometric (Ismail, 2021).

One study in the field of computer science suggests the classification of entomological origin of honey based on its physicochemical and antioxidant properties (Kek, 2017). In this study, raw Malaysian honey's physicochemical and antioxidant characteristics were employed as indicators to pinpoint its entomological source of the bee species Apis dorsata, Apis mellifera, Apis cerana, or Heterotrigona itama. The following physical and chemical characteristics of the sample were measured and analyzed: moisture content, water activity, specific gravity, viscosity, pH, free acidity, electrical conductivity, color (L*, a*, and b*), color intensity, and antioxidant properties, such as the DPPH free radical scavenging activity power (1/IC50), ascorbic acid equivalent antioxidant content (AEAC), ferric ion reducing antioxidant power (FRAP), and total phenolic content (TPC).

Using hierarchical cluster and principal component analysis, honeys were divided into

two main categories based on their physicochemical and antioxidant properties: those from honey bees (Apis spp.) and Trigona stingless bees (*Heterotrigona itama*). Heterotrigona itama, a species of stingless bee, produces honey in Kelulut that can be distinguished from honey from Apis spp. by its high moisture content of 33.24 g/100 g, free acidity of 136.8 meq/kg, color intensity of 990.3 mAU, AEAC of 26.64 mg/100 g, and FRAP of 41.95 mg AAE/100 g. Entomological categorization of honey aids in honey identification and lowers honey fraud (Kek, 2017).

Another study conducted by Kek which was published in 2016, regarding the concept of classification of stingless bee honeys is titled *Classification of Honey from Its Bee Origin via Chemical Profiles and Mineral Content*. This study includes the honey from *A.mellifera* as well as *Meliponini* to be classified for their origins. Raw honeys are classified based on its mineral contents, heavy metals, as well as chemical profiles which include predominant sugars, proximate composition, hydroxymethylfurfural content, together with diastase activity. Two methods implemented in this research which includes principal component analysis and hierarchical cluster analysis show a relatively high chance of differentiating honey by its species (Kek, 2016). This study finds that potassium and sodium were the major elements contained within raw regular honey samples. Contradictingly, more protein and less sugar of fructose as well as glucose can be identified in the samples of stingless bee honey (Kek, 2016).

Furthermore, another study is done to investigate the provenance establishment of stingless bee honey using multi-element analysis in combination with chemometrics techniques (Shadan, 2017). This study aims to provide the accurate identification of stingless bee honey geographical origin in combating fraudulent activities for the purpose of honey consumer protection. The methods implemented include an inductively coupled-plasma optical emission spectrometer as well as few technical methods such as principal linear discriminant analysis (LDA) together with component analysis (PCA). According to Shadan (2017), the result of this study shows an 87% correct classification rate through cross-validation for PCA, whereas a 96.2% correct classification rate is obtained with the utilization of LDA. This provides sufficient support for assigning the provenance of stingless bee honeys.

Hence, PCA and LDA are proven to be a viable method in the identification of stingless bee honey. A latest study done by Raypah et al. (2022) with the usage of near-infrared spectroscopy with chemometrics for identification and quantification of adulteration in high-quality stingless bee honey also applies the methods of PCA and LDA. This research implements the utilization of several chemometric tools including PCA, PCA-LDA, HCA, and PLSR where a total of 97.95% of explained variance is achieved (Raypah et al., 2022). This results in a complete distinction between pure and adulterated honeys.

Other than that, machine learning is also used to provide an alternative method in the discrimination of stingless bee honey. This is demonstrated in a research conducted by local Malaysian researchers, titled *Discrimination of Malaysian stingless bee honey from different entomological origins based on physicochemical properties and volatile compound profiles using chemometrics and machine learning.* It highlights that VOC markers are viable to be utilized to differentiate honey from different stingless bees (Sharin et al., 2021). In addition,

physicochemical and VOC profiles of stingless bee honey aid in honey authentication (Sharin et al., 2021).

Another study done internationally, specifically in the country of Venezuela, characterizes local stingless bee honey by multivariate analysis of physicochemical properties (Vit et al., 1998). Collected stingless bee honey samples are analyzed for several compositional factors by which, the most influential factors include reducing sugars, sucrose, and diastase activity for a correct classification (Vit et al., 1998). In addition, the entomological origin of honey samples are identified through the usage of three multivariate analysis methods (Vit et al., 1998).

3. PROBLEM FORMULATION

3.1. Problem Statements

The main problem identified in this project was the physicochemical and antioxidant characteristics of stingless bee honey were different from those of *Apis mellifera* honey as well as between various species of *Meliponini*. (Shamsudin et al., 2019).

In addition, another concern related to the correlation between chemical compounds and antioxidant activities are also unclear for this specific project.

3.2. Objectives

There are several objectives that are proposed to be achieved by the end of this project. Below are the list of primary objective as well as secondary objectives: -

Primary Objective:

1. To develop a classification model to classify different species of stingless bee through samples of stingless bee honey.

Secondary Objectives:

1. To identify the correlation between different chemical compositions with different antioxidant activities.
2. To deploy the classification as an interactive application that allows users to predict species of stingless bees based on honey attributes.

4. METHODOLOGY

This section of the study presents the overall methodology or data science pipeline that is applied to achieve all of the objectives listed. This includes data collection, Exploratory Data Analysis, data preprocessing, descriptive analytics, data modeling, model evaluation, and model deployment.

4.1 Data Collection

Dataset containing samples of stingless bee honeys are collected from a research team from Centre for Natural Resources and Drug Discovery(CENAR) in collaboration with The Faculty of Pharmacy, authorized by the Fundamental Research Grant Scheme(FRGS).

4.1.1. Sample Location

The samples of stingless bee honey are extracted from various sources of stingless bee keeping farms across five states in Malaysia including Kedah, Selangor, Johor, Pahang, and Terengganu. The states are associated with a code of capitalized alphabets for each sample. In addition, the code also refers to the exact location of the beekeeping farm. For example, A *itama* refers to a sample extracted from Kg. Kekabu, Felda Sg. Tiang, Pendang Kedah while C *itama* refers to a sample extracted from Kg. Tualak, Kuala Nerang, Kedah. All the samples

are extracted gradually from August 2019 to October 2022.

4.1.2. Sample Iteration

Each stingless bee honey sample is associated with an iteration that refers to the order of extraction in a particular beekeeping farm. For example, *itama i*, *itama ii*, *itama iii*, are referencing the first, second, and the third samples respectively extracted from a stingless beekeeping farm.

4.1.3. Properties Description

In the dataset, the features indicate attributes related to testing the chemical composition as well as antioxidant activities of stingless bee honey samples. Two chemical compositions namely phenolic compounds and flavonoid compounds are tested in each sample to understand the quantity of phenolic compounds as well as flavonoid compounds present in samples through the column of TPC and TFC respectively.

In addition, antioxidant assays are also used to test antioxidant activities present in honey samples under the colorimetric analysis. This is done through the assays of ABTS as well as FRAP that reflect the column name respectively. ABTS refers to a chemical assay which is done through a mechanism known as free radical scavenging whereas, FRAP refers to an assay which is done through a mechanism known as reducing agent.

4.1.4. Dataset Description

The dataset contains a total of 158 stingless bee honey samples, with 10 features altogether. Table 1 below is a brief description of the dataset's attributes.

Attribute	**Datatype**	**Description**
Sample	String	Species name of sample combined with extract iterations and state codes
PH	Float	The potential of Hydrogen of samples
TPC (µg GAE/mg honey)	Float	The Total Phenolic Content
TFC (µg QE/mg honey)	Float	The Total Flavonoid Content
ABTS (% inhibition)	Float	Antioxidant assay of 2,2'-azino-bis (3-ethylbenzothiazoline6-sulfonic acid to test antioxidant activities in samples
FRAP (µg FeSO4/mg honey)	Float	Antioxidant assay of Ferric Reducing Antioxidant Power to test antioxidant activities in samples
Water content (% weight/weight)	String	The moisture level present in samples

Fructose content (% weight/weigh t)	Float	The content of Fructose sugar in samples
Glucose content (% weight/weigh t)	Float	The content of Glucose sugar in samples
Remark	String	A side note to be referenced as for 'Out Of Range' values

Table 1 : Dataset description

4.2. Data Preprocessing

The data is preprocessed before further utilization to have cleaned data for this project. The data cleaning steps includes renaming columns, splitting columns, replacing and removing values, imputing missing values, changing data types, removing unused columns, standardizing values, and randomizing the dataset. For this project, the attributes of the samples are only taken to compare different species of stingless bee honey. The various species compared are including H.Itama, T.Binghami, T.Apicalis, H.Erythrogastra, G.Thoracica, H.Fimbriata, and T.Melanoleuca.

4.2.1. Renaming Columns

Some columns in the initial dataset contain the complex names of chemical compounds as well as antioxidant assays. To avoid confusion and have a better understanding, the columns will have to be renamed to much simpler names for reference.

4.2.2. Splitting values into New Columns

Each sample name is associated with codes that references the beekeeping farm location where the sample is extracted. The alphabetical codes were splitted from the sample name and are stored in a new column 'State'.

4.2.3. Replacing and Removing Values

With the new column created, the state codes were replaced with the actual state names. For instance, the state code A, B, and C are replaced with the value of 'Kedah', while the state code D and E are replaced with the value of 'Selangor'. This is done to increase the readability of the data.

4.2.4. Imputing Missing Values

The column of Water content (% weight/weight), renamed Water contains the value of 'OOR' which is referencing the value being out of range. To address this issue, all 'OOR' values will be replaced with 0 to reference a null value. The figure 0 will then be replaced with a suitable value that is discussed with the research team which is 32.

4.2.5. Changing Data Types

Despite having the missing values imputed, the datatype of the Water column is still counted as a string. Hence, it is necessary to change the datatype to a numerical figure to be used as inputs for the model development.

4.2.6. Removing Unused Columns

There are several columns that are considered as unnecessary and will not be used for building machine learning models. The columns include 'Sample_ID', 'State', and 'Remark'.

4.2.7. Standardizing Values and Randomizing Dataset

Finalizing the data preprocessing steps will be to standardize all the values contained within the dataset especially within the column that contains sample names. All sample names are capitalized to have only a single name representing one species.

Furthermore, the data is sorted in the order of sample extraction for the column of sample names which could cause bias in data splitting and data modeling. Hence, the dataset is randomized.

4.3. Descriptive Analytics

To achieve the secondary objective of this project which is to observe and determine the correlation between chemical compounds and antioxidant activities in honey samples, a descriptive analytics approach is implemented. A few graphs are plotted to see how chemical compounds correlate with antioxidant activities. Below are the graphs to show the relationship between TPC as the chemical compound and ABTS as well as FRAP as antioxidant assays that show the antioxidant activity.

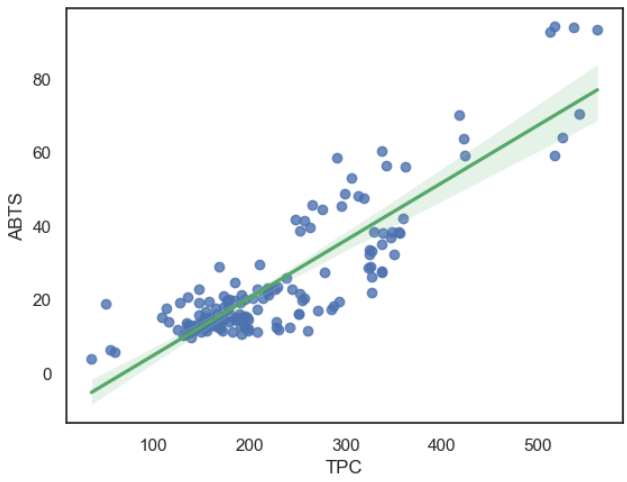

Figure 1: Scatter plot of TPC vs ABTS

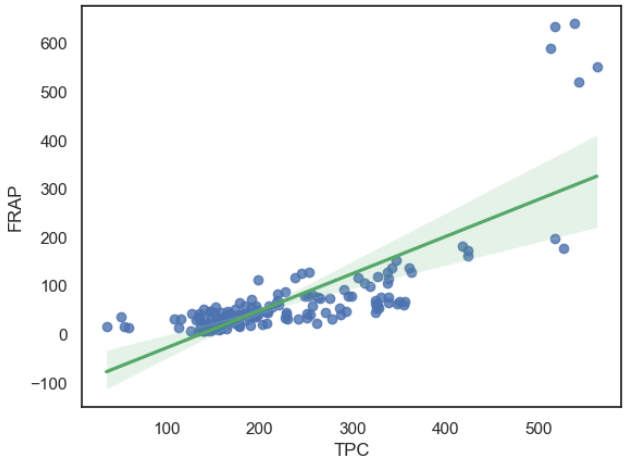

Figure 2: Scatter plot of TPC vs FRAP

In Figure 1, it is observed that the correlation of TPC and ABTS is relatively linear. The same could be said for TPC and FRAP (Figure 2). Nevertheless, below are the graphs to show the relationship between TFC as the chemical compound and ABTS as well as FRAP as the antioxidant assays that shows the antioxidant activity.

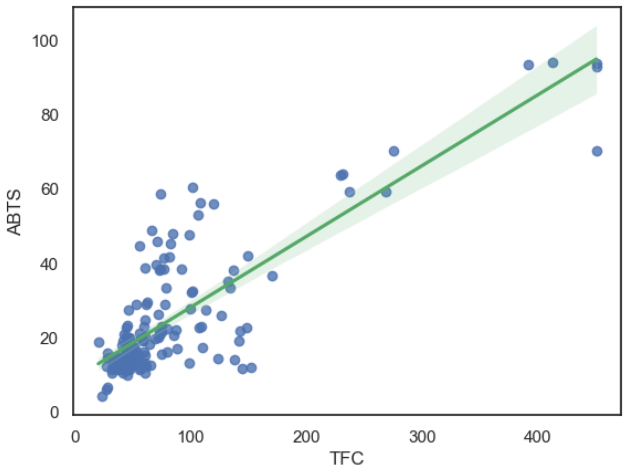

Figure 3: Scatter plot of TFC vs ABTS

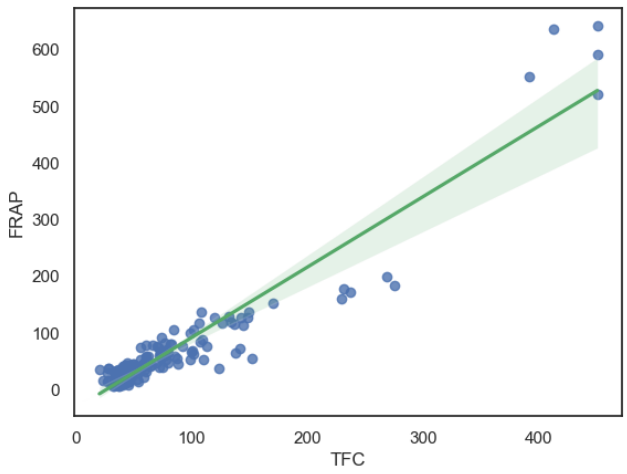

Figure 4: Scatter plot of TFC vs FRAP

Both of the graphs (Figure 3 & 4) visualize a linear correlation between TFC and the antioxidant assays of ABTS and FRAP respectively.

Hence, based on all four graphs, it can be concluded that the chemical compounds have a positive correlation with antioxidant activities as the chemical compounds present in stingless bee honey samples increases, the antioxidant activities in the samples increases as well.

4.4. Data Modeling

In this project, there are several classification models that are used for the purpose of multiclass classification of stingless bee honeys namely Logistic Regression, Decision Trees, K-Nearest Neighbours, Support Vector Machine, and also Random Forest. The models are chosen based on popularity, common uses and performance.

Prior to developing machine learning models, the dataset should be splitted into two sets which are a train set and a test set with the ratio of 80:20 respectively.

For each individual classifier, the attributes of honey are designated as inputs, whereas the sample name is designated as the label. The train set is fitted into each classifier, followed by a prediction made onto the test set. The results of the predictions of classifiers are based on accuracy where the higher the accuracy produced by a classifier, the better the classifier's performance and vice versa.

4.5. Model Evaluation

Classifiers performance are needed to be evaluated to have a single model with the highest accuracy as the main model for deployment. The evaluation metrics utilized in this study are confusion matrix, precision, recall, as well as f1-score.

4.6. Model Deployment

Given the development of machine learning models centers around the usage of Python as the main programming language, the web application developed is by using Gradio which is built on top of Python. Gradio is an open-source Python library for building interactive machine learning web applications. It allows users to easily create and deploy their own machine learning models as web applications with a simple API. It supports a wide variety of machine learning frameworks and libraries making it the best candidate to deploy the developed random forest classifier.

5. RESULTS

The results of this study is shown in this section which comprises the classification model results as well as the final web application that is developed to deploy the model.

5.1. Performance of machine learning models

Each model is built and trained to perform classification and are evaluated based on accuracy. Random Forest has the highest accuracy among the classifiers hence, it is chosen as the proposed model for deployment.

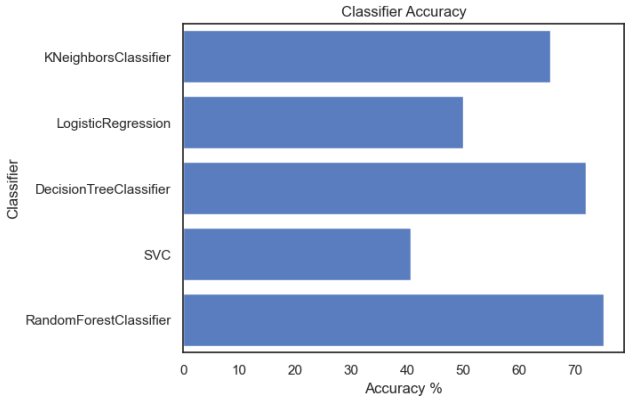

Figure 5: Horizontal bar graph of accuracy of classifiers

Random Forest has the accuracy of 75% followed by Decision Trees and K-Nearest Neighbours with the accuracy of 71.88% and 62.63% respectively (Figure 5). Under the surface, Random Forest is basically a Classification and Regression Trees algorithm (CART), except it is composed of many trees instead of just a single tree (Dobilas, 2021). The relatively high accuracy of the Random Forest classifier is due to the advantages of the model itself which includes improved performance and also improved robustness. In Random Forest, the number of trees can affect the performance of the classifier. Higher number of trees gives better performance but makes the complexity higher, while a lesser number of trees provides a quick and simple classification process with decent performance.

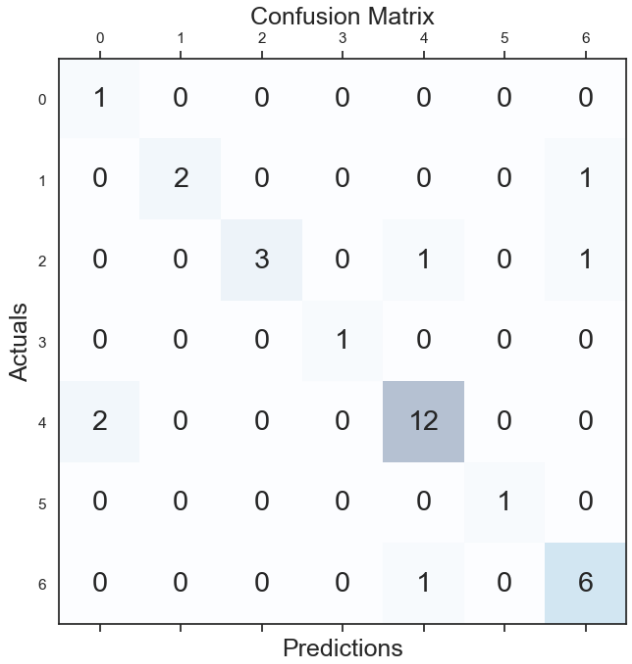

Figure 6: Confusion matrix of Random Forest Classifier

According to Heydarian (2022), a powerful method for analyzing a multi-class classifier is the 2-dimensional confusion matrix to show the distribution of false predictions in one view. Confusion matrix allows the identification of correct predictions which consist of True Positive(TP) and True Negative(TN) and false predictions which consist of False Positive(FP) and False Negative(FN).

Interpreting a multiclass confusion matrix can be quite challenging. In Figure 6, the numbers in both y and x-axis represent the species name that is sorted alphabetically respectively. As an instance, the figure 0 corresponds to the first species in alphabetical order which is *Apicalis*. In the case of a true positive (TP) and true negative (TN), the value that is aligned with the same classes is the amount of correct predictions. Nevertheless, there is a limitation of using a regular confusion matrix to evaluate multiclass classifiers, where this constraint causes uncertainty on the number of false negative (FN)

or false positive (FP) connected with each label (Heydarian, 2022).

	precision	recall	f1-score	support
Apicalis	0.33	1.00	0.50	1
Binghami	1.00	0.67	0.80	3
Erythrogastra	1.00	0.60	0.75	5
Fimbriata	1.00	1.00	1.00	1
Itama	0.86	0.86	0.86	14
Melanoleuca	1.00	1.00	1.00	1
Thoracica	0.75	0.86	0.80	7
accuracy			0.81	32
macro avg	0.85	0.85	0.82	32
weighted avg	0.86	0.81	0.82	32

Figure 7: Precision, Recall, and F1-score of Random Forest Classifier

Through the confusion matrix in Figure 6, certain evaluation metrics can be derived which includes precision, recall, and f1-score for each class and macro and weighted average of all classes to measure the overall performance of a classifier (Heydarian, 2022). Through Figure 7, for each class, precision refers to the correct positive predictions relative to the total positive predictions whereas recall refers to the correct positive predictions relative to total actual predictions (Zach, 2021). The f1-score can be calculated by the formula of:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

5.2. Web application

Gradio provides a simple web application framework to deploy machine learning models. The interactive web application developed is called Pollenize - A Stingless Bee Honey Classifier. The web application consists of two pages which are the login page and the main interface.

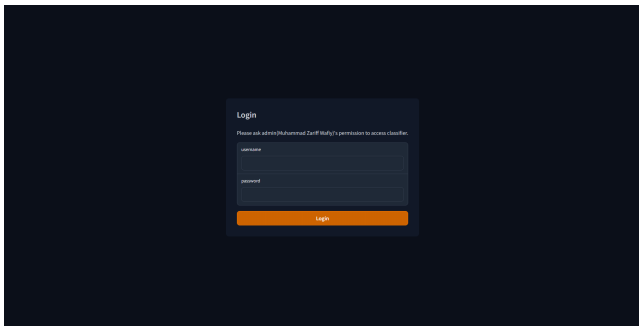

Figure 8: Login page of Pollenize

Figure 8 displays the login page that requests text inputs for admin username and password which can only be obtained by the administrator of the Pollenize. Text entered in the password section is censored to provide a higher security level for the web application.

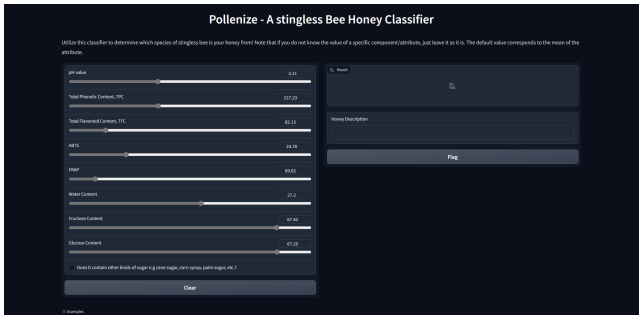

Figure 9: Default main interface of Pollenize

As shown in Figure 9, after entering the correct admin username and password, the user will be directed to the main interface of Pollenize. This interface consists of several sections which include sliders for attribute inputs, a text output to state the result of prediction, and a textbox to provide few descriptions regarding the predicted honey. The value of inputs are defaulted to be the mean value for each attribute which is reflected by the position of the sliders.

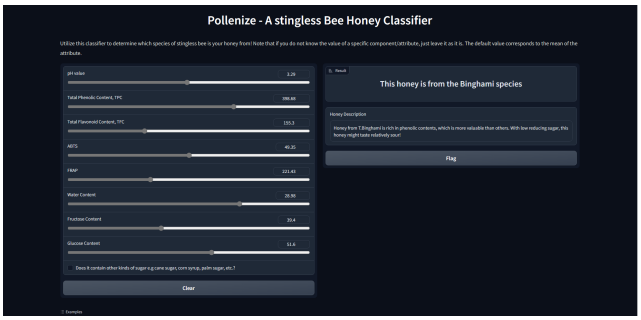

Figure 10: Adjusted inputs for different parameters

Based on Figure X shown, the sliders are adjusted to provide a different set of inputs towards each attribute of the honey. The inputs are live which makes this web application interactive. For instance, a single adjustment of any sliders will automatically invoke an output based on the current set of inputs.

6. DISCUSSIONS

Overall, each and every single one of the objectives proposed in this study have been achieved and the problem statements have been addressed. The first objective of finding the correlation between chemical attributes and antioxidant activities of stingless bee honey samples has been achieved by plotting several graphs to show the said relation. It is observed that chemical attributes do indeed contribute to the total antioxidant activities of stingless bee honey.

The next objective of this project which is to develop a machine learning classification model to classify different species of stingless bee honey based on attributes is achieved by having all the classifiers evaluated, which the Random Forest classifier performs the best by the metrics of accuracy.

The last objective of developing an interactive web application is also met with the end product of Pollenize - A Stingless Bee Honey Classifier. The web application receives a set of attribute values as the user inputs and is fitted into the proposed Random Forest classifier to predict the origin of the honey.

7. CONCLUSION

In conclusion, the classification of stingless bee honey is approached with machine learning techniques of multiclass classification which separates different honey samples into different types of stingless bee species. The dataset obtained from the FRGP research contains the physicochemical compounds, chemical composition, as well as antioxidant activities of honey samples as variables for classification. The data is preprocessed with a set of data cleaning steps before the development of models to have standardized and cleaned data. Several machine learning classifiers are chosen to train the dataset and are evaluated with accuracy. Random Forest classifier holds the top spot for the highest accuracy of 75% compared to the other classification models hence, it is chosen as the main model for deployment through the means of Gradio web application framework.

This study is by no means perfect nor advanced in the field of Computer Science. There are several identifiable limitations of this study which include limited samples and the absence of adulterated honey. The data utilized in this study only contains 158 samples of stingless bee honey samples which is a relatively small figure. This has absolutely impacted the results of the study by having a sub par accuracy of classifiers performance. There is also the problem of the absence of adulterated honey samples in the dataset. With the presence of adulterated honeys, the horizon of this study could be widened to include the question of honey purity and authenticity.

For future reference, this study could be improved on several aspects including performing multiclass classification based on beekeeping locations. This is also to observe which attributes will be affected more if the locations of beekeeping are different due to various weather conditions and geographic terrain. In addition, the attributes used as features in this study are only a portion of a larger collection of attributes used to test the properties of stingless bee honey samples. Hence, more honey attributes could be added in the dataset to increase the number of features for model development. Furthermore, the objective of future studies could be expanded to include the question of the purity and authenticity of stingless bee honeys. For this to be valid, adulterated honey samples must be present and should be included in the dataset to have observations of impure honeys. The result will have a bigger impact towards the community of honey consumers as they can identify the purity and authenticity of stingless bee honey that they purchased or are planning to.

ACKNOWLEDGEMENTS

This endeavor would not have been possible without my Data Science Project supervisor, Dr. Nor Liyana Mohd Shuib for her unlimited efforts in guiding me towards completing this study. I also could not have undertaken this journey without the research team under the FRGS-FP035-2021 project, led by Assoc. Prof. Dr. Najihah Mohd Hashim along with Dr. Zuwairi Saiman, En.Nazil Afiq, and En. Kasful Askra Sakika who assisted me in providing data insights and feedback.

REFERENCES

Abd Jalil, M. A., Kasmuri, A. R., & Hadi H. (2017). Stingless Bee Honey, the natural wound healer: A Review. Skin Pharmacology and Physiology, 30(2), 66-75. Retrieved November 5, 2022, from https://doi.org/10.1159/000458416

Braghini, F., Biluca, F. C., Schulz, M., Gonzaga, L. V., Costa, A. C., & Fett, R. (2021). Stingless Bee Honey: A precious but unregulated product - reality and expectations. *Food Reviews International, 38*(sup1), 683-712. https://doi.org/10.1080/87559129.2021.1884875

Dobilas, S. (2022, February 5). *Random Forest Models: Why are they better than single decision trees?* Medium. Retrieved January 24, 2023, from https://towardsdatascience.com/random-forest-models-why-are-they-better-than-single-decision-trees-70494c29ccd1

Fletcher, M.T., Hungerford, N.L., Webber, D. *et al.* Stingless bee honey, a novel source of trehalulose: a biologically active disaccharide with health benefits. *Sci Rep* 10, 12128 (2020). https://doi.org/10.1038/s41598-020-68940-0

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access, 10,* 19083-19095. https://doi.org/10.1109/access.2022.3151048

Ismail, N. F., Maulidiani, M., Omar, S., Zulkifli, M. F., Mohd Radzi, M. N., Ismail, N., Jusoh, A. Z., Roowi, S., Yew, W. M., Rudiyanto, R., & Ismail, W. I. (2021). Classification of stingless bee honey based on species, dehumidification process and geographical origins using

physicochemical and ATR-FTIR chemometric approach. *Journal of Food Composition and Analysis, 104,* 104126. https://doi.org/10.1016/j.jfca.2021.104126

Kek, S. P., Chin, N. L., Tan, S. W., Yusof, Y. A., & Chua, L. S. (2016). Classification of honey from its bee origin via chemical profiles and mineral content. *Food Analytical Methods, 10*(1), 19-30. https://doi.org/10.1007/s12161-016-0544-0

Kek, S. P., Chin, N. L., Yusof, Y. A., Tan, S. W., & Chua, L. S. (2017). Classification of entomological origin of honey based on its physicochemical and antioxidant properties. *International Journal of Food Properties, 20*(sup3). https://doi.org/10.1080/10942912.2017.1359185

Pimentel, T.C., Rosset, M., Sousa, J.M., Oliveira, L. I., Mafaldo, I.M., Pintado, M. M., Souza, E.L., & Magnani, M. (2021). Stingless Bee Honey: An overview of health benefits and main market challenges. *Journal of Food Biochemistry*, 46(3). https://doi.org/10.1111/jfbc.13883

Raypah, M. E., Zhi, L. J., Loon, L. Z., & Omar, A. F. (2022). Near-infrared spectroscopy with chemometrics for identification and quantification of adulteration in high-quality stingless Bee Honey. *Chemometrics and Intelligent Laboratory Systems, 224,* 104540. https://doi.org/10.1016/j.chemolab.2022.104540

Razali, M., Zainal, Z., Maulidiani, M., Shaari, K., Zamri, Z., Mohd Idrus, M., Khatib, A., Abas, F., Ling, Y., Rui, L., & Ismail, I. (2018). Classification of raw stingless bee honeys by bee species origins using the NMR-and LC-MS-based metabolomics approach. *Molecules,* 23(9), 2160. https://doi.org/10.3390/molecules23092160

Shadan, A. F., Mahat, N. A., Wan Ibrahim, W. A., Ariffin, Z., & Ismail, D. (2017). Provenance establishment of stingless bee honey using multi-element analysis in combination with chemometrics techniques. *Journal of Forensic Sciences, 63*(1), 80-85. https://doi.org/10.1111/1556-4029.13512

Shamsudin, Selamat, Sanny, A.R, Jambari, & Khatib. (2019). A Comparative Characterization of Physicochemical and Antioxidants Properties of Processed Heterotrigona itama Honey from Different Origins and Classification by Chemometrics Analysis. Molecules, 24(21), 3898. https://doi.org/10.3390/molecules24213898

Sharin, S. N., Sani, M. S., Jaafar, M. A., Yuswan, M. H., Kassim, N. K., Manaf, Y. N., Wasoh, H., Zaki, N. N., & Hashim, A. M. (2021). Discrimination of Malaysia stingless bee honey from different entomological origins based on physicochemical properties and volatile compound profiles using chemometrics and machine learning. *Food Chemistry*, *346*, 128654. https://doi.org/10.1016/j.foodchem.2020.128654

Souza, E.C.A., Menezes, C. & Flach, A. Stingless bee honey (Hymenoptera, Apidae, Meliponini): a review of quality control, chemical profile, and biological

potential. *Apidologie* 52, 113–132 (2021). https://doi.org/10.1007/s13592-020-00802-0

Vit, O., Oddo, L. P., Marano, M. L., & Salas de Mejias, E. (1998). Venezuelan stingless bee honeys characterized by multivariate analysis of physicochemical properties. *Apidologie, 29*(5), 377-389. https://doi.org/10.1051/apido:19980501

Zach. (2021, October 7). *How to calculate F1 score in Python (including example).* Statology. Retrieved January 27, 2023, from https://www.statology.org/f1-score-in-python/

Zuccato, V., Finotello, C., Menegazzo, I., Peccolo, G., & Schievano, E. (2017). Entomological authentication of stingless bee honey by 1H NMR-based Metabolomics Approach. *Food Control, 82,* 145-153. https://doi.org/10.1016/j.foodcont.2017.06.024