CHAPTER 1
INTRODUCTION

1.1 Rationale

The development of machine learning is very important in various aspects such as IT, education, social, health, and so on [1]. In the health domain, it is widely recognized that food has a significant impact on providing energy to the human body [2]. To be more precise, nutrition derived from the food we consume enables us to perform daily activities such as working, running, and other physical endeavors. Nutrition is one of the fundamental resources that our body requires to function properly [3]. Today, many research using food as an object and only discuss about the quality of food, the best food to consume in certain restaurant/website [4], and the nutrients present in different foods, there has been minimal research focused on clustering food ingredients based on their macronutrients and research micronutrients value. According to experts [5] macronutrients are the nutrients we need in larger quantities that provide us with energy, whereas micronutrients are vitamins and minerals needed by the body in very small amounts. Although they are required in small amounts, their impact on a body's health are critical, and deficiency in any of them can cause severe and even life-threatening conditions [6]. In this studies, we define macronutrients as nutrients that provide most of the energy to the human body, like protein, fat, and carbohydrate, and micronutrients as vitamins and minerals that are essential for our health, like other nutrients such as water, energy, and zinc.

In previous studies, food ingredients were used to classify information on how they are essential for the human body. For example, a paper was published that presented how food could provide vital information to improve diet and nutrition service ontology [7]. To achieve this, the authors used a K-Means methodology, combined with the recommender system. However, using the K-Means methodology for clustering has a significant drawback, as it requires a randomly determined K value. The results obtained by using K-Means depend on the initial values set, making it less flexible. In another studies they use the K-Function method for the case of classifying the area nearest to schools and fast food restaurants [8]. It is a good idea to provide the neighbourhood with information about nearby schools and restaurants but applying the K-Function method is more difficult to identify areas if many restaurants are in the same location as more than one school. So far we have seen the use of clusters as a more popular method and food as a pretty good object, but we need to explore more especially about the nutrients that are present in each food.

Nutrition is the primary source of energy for the human body [9]. Our daily activities, such as work, exercise, and household chores, require energy, which is derived from the food we consume. The composition of the food we eat must be balanced [10] to avoid causing diseases and inefficiency in the body. The nutritional needs of each individual vary, depending on their daily activities and body shape [11]. People who often do more activities than others need more nutrition from food and vice versa if the activity tends to be little, then the nutrients needed are also at sufficient levels. From other aspects, for example, for people who have a history of certain diseases it is also a component that needs to be considered for their nutritional needs, because certain diseases will avoid certain foods/nutrients so as not to put the sufferer in an unwanted condition. [12]. So, when viewed from various aspects, it is necessary to provide information about the nutrients in food for use by people who have excessive activity, have a history of certain diseases or people who exercise regarding the nutritional needs they need. This nutritional information is taken from food ingredients or processed food/food products. By paying attention to this, we aim to cluster food ingredients based on their nutritional value by dividing the dataset into macronutrients and micronutrients to identify similarity clusters and evaluate them.

To cluster food ingredients based on nutritional value, several stages are required [13]. These stages include data preparation, modeling, and data visualization, all of which are critical to the success of clustering food ingredients. The modeling stage employs two methods to analyze and understand the obtained results from different perspectives. The analysis and modeling results are expected to show unique similarities based on food nutrition. Analytical data plays a crucial role in representing information on food ingredients and their respective nutrients. The study results can serve as the basis for further research on nutritional recommendations for energy in the body by adjusting aspects of activity, body shape, and lifestyle. Data analytics is a useful method for extracting valuable information from food datasets, and several steps are necessary to accomplish this, including data collection, data cleaning, data modeling, and data visualization. To ensure the best result, we will not rely on a single clustering method but focus on multiple clustering models to enhance the flexibility and score evaluation of this research.

This research aims to cluster food ingredients based on nutrition using dynamic clustering, which utilizes multiple methods to cluster datasets automatically. Unlike the static clustering approach that works on a predefined number of clusters, dynamic clustering can determine the number of clusters without predefined number of clusters [14].

1.2 Problem Formulation

Currently, many researchers use food as an object and only talk about how good the food is at a particular restaurant/site, or what food is best recommended for elderly, and how much nutrition is wasted in a particular place. Nevertheless, no one is talking about how food or food ingredients can be grouped based on their nutrition, both macro and micro nutrients. In other research, we can learn about the popularity of a recipe or ingredient from many foods in a restaurant or from a website that provides menu and recipe lists, but we cannot find the nutritional content of the food ingredients, so we miss out on one of the most important components of food, which is nutrition. Why is nutrition so important? because with nutrition, we know about food and what is good for our bodies. For many people, having good nutrition can affect their lives or activities. The nutrients they get from food can be converted into energy or strength. Furthermore, nutrients can help people understand what foods they can and cannot eat based on their eating patterns or healthy methods to maintain their food or maintain their body's health.

Nutrition itself contains very detailed information, so it takes more time to understand it. For example, a list of foods usually has many nutrient contents such as fats, proteins, etc. This makes it difficult for people to read and understand it. In addition, comparing the nutritional information of one ingredient with another will take longer, so nutritional information between food ingredients and others will be more often ignored. Therefore, it is proposed as a solution to make food ingredients into groups based on the nutrients they possess.

The results of related studies say that clusters of food products will be recommended to users, and similar food products will also be recommended using a knowledge-based system [1] With the hope of using a hybrid clustering-based food recommendation method that uses chronic disease, diet, and nutrition-based clustering ontologies, better information can be provided. Nevertheless, the research only has recommendations based on food and chronic diseases, not on nutrient content. Therefore, we decided to involve nutrition in the clustering of food ingredients and when we tried to group food ingredients based on nutrition, we found a similar but not identical study on food waste. Food waste is all food that is not consumed or discarded even though it is fit for human consumption [2]. Food waste research has already used a nutrient-based foundation, but only on carbohydrates, meat, and fats. In general, this does not provide enough information for those who want to know about small nutrients. Therefore, the researcher chose to use macronutrient and micronutrient datasets for this thesis and used agglomerative clustering methods that are suitable for numerical datasets, have flexibility in determining the number of clusters, and support the number of datasets that are

not too large. In addition to the agglomerative method, the researcher also used affinity propagation because it applies the concept of dynamic clustering.

Based on those, the research question of this thesis is:

a. How to clusters the food ingredients based on nutrition using agglomerative clustering and affinity propagation?

In addition, there are two sub-research questions of this thesis that should be answered, namely:

1. Which one has higher performance of silhouette coefficient : the agglomerative or affinity propagation?
2. When should we use these two methods for macronutrient and micronutrient?

1.3 Objective and Hypothesis

The goal of this thesis is to determine which algorithm is suitable for each nutrient: macronutrients and micronutrients.

The hypothesis of this thesis is that the relationship between macronutrient and micronutrient datasets and agglomerative clustering methods depends on the linkage type and the number of clusters, whereas affinity propagation is influenced by the preference value.

1.4 Assumption

For research in this thesis, we use several assumptions that support the writing of this thesis, including implementing several Python libraries to meet the needs of analysts and implementation in data analytics, then we assume the food ingredients dataset here includes the food ingredients themselves and processed foods as well as food. products and use the clustering method to group similar food ingredients.

1.5 Scope and Delimitation

1. The system will group the nutrient from food ingredients using two method clustering.
2. The system will focus on seeing how the result of each algorithm clusters the food ingredients based on macro and micro nutrient.
3. The system uses steps such as data preparation and modeling before clustering and data analysis for cluster evaluation.