Implementation of K-Modes Algorithm for Clustering of Stress Causes in University Students

Ni Putu Mira Novita Dewi[a1], Ida Bagus Gede Dwidsamara[a2]

[a]Informatics Department, Udayana University
Bali, Indonesia
[1]miranovitad@gmail.com
[2]dwidasmara@unud.ac.id

Abstract

Stress is an inevitable part of life in a college environment. The variety of factors that cause stress in students, it is necessary to cluster the factors that cause stress in students to see the description of the characteristics of each cluster of students. The clustering process is carried out to identify the causes of stress in student groups and their relationship to these internal and external factors. Cluster analysis can be used as a reference to decide on efforts to handle and prevent increased stress in students.

The clustering process is carried out using the Python programming language. The algorithm used is the k-modes clustering algorithm. This algorithm is suitable for clustering categorical data. The optimal number of clusters obtained from the implementation of the elbow method is three clusters. Cluster 1 is a cluster with a mild stress level, the main cause of stress is academic issues. Cluster 1 is the only group where the majority of the cause of stress is not financial. Cluster 2 is a cluster with a high stress level which causes various stressors. However, cluster 2 is the only cluster where the cause of stress is on careers and on involvement in hostels, clubs, and society. Cluster 3 is a cluster with a medium stress level. This cluster is the only cluster dominated by male gender. The main cause of stress in this cluster is academic and financial.

Keywords: *Stress, k-modes, elbow method, clustering, university student*

1. Introduction

Changes in the values of life due to globalization cause a variety of problems faced by society. Personal inability to solve social problems and meet environmental demands can cause stress in a person. Stress is the body's reaction that occurs when someone faces a threat, pressure, or a change. Stress definitely happens to anyone, including children, adolescents, adults, or the elderly. If the amount of stress experienced by a person is too much, it can be bad for their physical and mental condition [1]. Mental disorders are still a problem that needs more attention. A mental disorder that is too severe may cause a person to have suicidal thoughts. According to a survey conducted by WHO, more than 800,000 people in various parts of the world each year die because of suicide. In fact, suicide is the 2nd largest cause of death that occurs in someone aged 15 to 29 years.

Stress is an inevitable part of life in a college environment. According to research conducted at the University of Gondar in Ethiopia, the prevalence of emotional mental disorders in college students was 40.9%, whereas according to a study conducted at the German University the prevalence of emotional mental disorders in college students was 22.7% [2]. The stress experienced by students is caused by two factors which are divided into internal factors and external factors. Internal factors consist of gender, socioeconomic status, student personality characteristics, coping strategies, ethnicity and culture, and intelligence. While external factors consist of job or academic demands and student relations with their social environment [3].

The variety of factors that cause stress in students, it is necessary to cluster the factors that cause stress in students to see the description of the characteristics of each cluster of students. The clustering process is carried out to identify the causes of stress in student groups and their relationship to these internal and external factors. Cluster analysis can be used as a reference to decide on efforts to handle and prevent increased stress in students.

Clustering is the process of grouping data into classes or clusters, so that objects in a cluster have a high similarity to each other, but are very different from objects in other clusters [4]. Research on clustering has been carried out, including grouping villages based on indicators of diarrhea disease using k-means clustering [5], identifying clusters of the working age population in South Sumatra Province using k-modes which aim to facilitate the government in making policies [6], data clustering drugs at Pekanbaru General Hospital with the k-means algorithm [7], and the application of the k-means cluster for the effect of emotional intelligence and stress on student achievement [4].

The data used in this study were student response data regarding the ideal student life factors that are important in student life. The data was obtained from the Kaggle website in the form of files in the comma separated values format. Meanwhile, the clustering algorithm used in this study is the k-modes algorithm. The k-modes algorithm is a process of grouping data into a group, so that the data in a group has very large similarities, but is not very similar to data in other clusters [8]. The k-modes algorithm is a modification of the k-means algorithm which tries to provide a solution for grouping categorical data. The k-modes algorithm is designed to solve problems in the k-means algorithm, namely its use is only limited to the use of numeric data (interval / ratio) because the k-means algorithm is grouped by calculating the average of one data with other data, Meanwhile, for the case of categorical data, the average value cannot be calculated and cannot be applied using the k-means method. The final result of this research is student groups based on the causes of stress both from internal and external factors.

2. Research Methods

The process in this research includes problem analysis, data collection from the Kaggle website, data preprocessing, clustering using k-modes, and the last is evaluating the results of the cluster formed. After getting the results from clustering, each cluster that is formed will be analyzed to determine the characteristics of each cluster. The following is a flowchart of the clustering process in this study which can be seen in **Figure 1**.

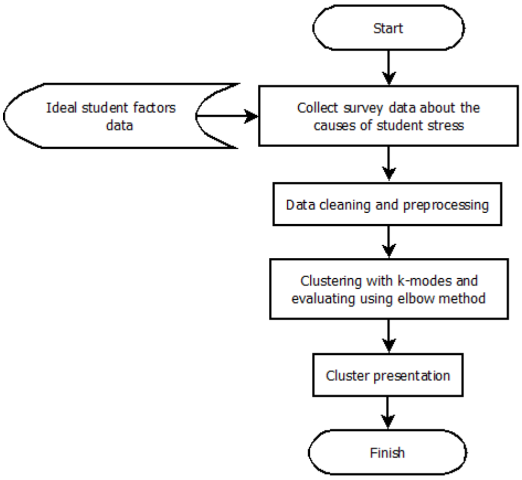

Start
Ideal student factors data
Collect survey data about the causes of student stress
Data cleaning and preprocessing
Clustering with k-modes and evaluating using elbow method
Cluster presentation
Finish

Figure 1. Research Flow Diagram

The flow diagram above shows that this study begins with the process of collecting data related to the causes of stress in students. The data is taken from the ideal student factors data set. After the data is sorted, the next process is to import and clean the data. Data cleaning is done by checking whether the imported data is NULL or not. Furthermore, the preprocessing process is carried out by carrying out label encoding so that the data can be processed by the method or algorithm that will be used. After cleaning and preprocessing data, the next step is clustering. The clustering method used is by using k-modes. To evaluate the number of clusters used the elbow method. This method is used to determine the optimal number of clusters. After that, a cluster presentation is carried out. At this stage, conclusions are made about the characteristics of each cluster.

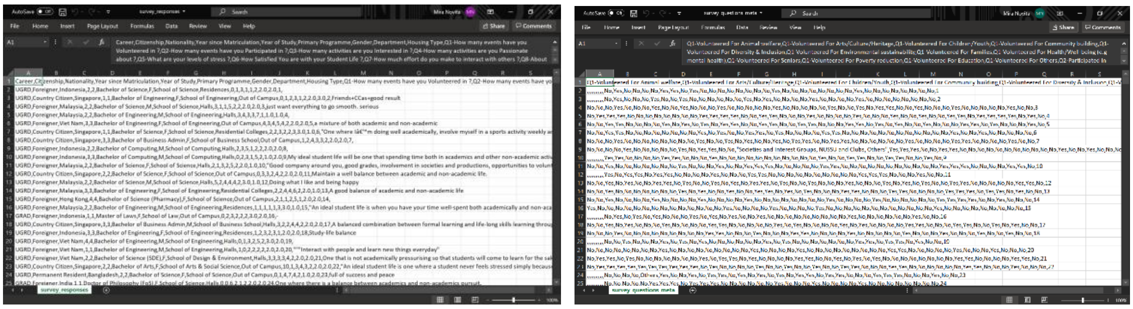
2.1. Problem Analysis and Literature Review

Stress is something that can happen to anyone. However, if the amount of stress faced by a person is so much it can have an impact on that person's mental health. At present, mental health issues need more attention because if they are not handled seriously it can trigger a person to have suicidal thoughts. Stress is an inevitable part of student life. Student stress is caused by various factors, both internal and external. So, in this study the authors will cluster the causes of stress in student groups and their relationship with external and internal factors to decide on efforts to handle and prevent stress in students.

This research begins by collecting journals related to the proposed method. The author collects journals that discuss the k-modes algorithm which is a modified algorithm of k-means. In addition to collecting journals related to the method of clustering, the author also collects journals related to the cases that will be raised in this study, namely regarding student stress and the factors that usually influence it.

2.2. Data Collection

The dataset used in this study is the level of student stress and its causes obtained from the ideal student factor survey data. This data is downloaded from the Kaggle website as a comma separated values file. Not all data will be used in this study. Therefore, it is necessary to preprocess the data to choose which data can and is good for the clustering process. In the ideal student factor raw data, there are two interrelated .csv files. In the clustering process, not all of the attributes in the raw data are needed, only attributes related to the research topic that will be used in the process. Therefore, a preprocessing process is needed by reducing some of the attributes that are not needed in the clustering process.



```
Career,Citizenship,Nationality,Year since Matriculation,Year of Study,Primary Programme,Gender,Department,Housing Type,Q1-How many events have you Volunteered in ?,Q2-How many events have you Participated in ?,Q3-How many activities are you Interested in ?,Q4-How many activities are you Passionate about ?,Q5-What are your levels of stress ?,Q6-How Satisfied You are with your Student Life ?,Q7-How much effort do you make to interact with others ?,Q8-About
1  Career,Citizenship,Nationality,Year since Matriculation,Year of Study,Primary Programme,Gender,Department,Housing Type,Q1-How many events have you Volunteered in ?,Q2-How many events have yo
2  UGRD,Foreigner,Indonesia,2,2,Bachelor of Science,F,School of Science,Residences,0,1,3,1,1,2,2.0,2.0,1,
3  UGRD,Country Citizen,Singapore,1,1,Bachelor of Engineering,F,School of Engineering,Out of Campus,0,1,2,3,1,2,2.0,3.0,2,Friends+CCAs+good result
4  UGRD,Foreigner,Malaysia,2,2,Bachelor of Science,M,School of Science,Halls,3,1,1,5,2,2,2.0,2.0,3,just want everything to go smooth. serious
5  UGRD,Foreigner,Malaysia,2,2,Bachelor of Engineering,M,School of Engineering,Halls,3,4,3,3,7,1,1.0,1.0,4,
6  UGRD,Foreigner,Viet Nam,3,3,Bachelor of Engineering,F,School of Engineering,Out of Campus,4,3,4,5,4,2,2.0,2.0,5,a mixture of both academic and non-academic
7  UGRD,Country Citizen,Singapore,1,1,Bachelor of Science,F,School of Science,Residential Colleges,2,2,3,2,2,3,3.0,1.0,6,"One where Iâ€™m doing well academically, involve myself in a sports activity weekly a
8  UGRD,Country Citizen,Singapore,3,3,Bachelor of Business Admin,F,School of Business School,Out of Campus,1,2,4,3,3,2,2.0,2.0,7,
9  UGRD,Foreigner,Indonesia,2,2,Bachelor of Computing,M,School of Computing,Halls,2,3,5,1,2,2,2.0,2.0,8,
10 UGRD,Foreigner,Indonesia,3,3,Bachelor of Computing,M,School of Computing,Halls,0,2,3,1,5,2,1.0,2.0,9,My ideal student life will be one that spending time both in academics and other non-academic activ
11 UGRD,Foreigner,Malaysia,2,2,Bachelor of Science,F,School of Science,Halls,2,1,3,2,5,2,2.0,1.0,10,"Good company around you, good grades, involvement in societies and productions, opportunities to volunt
12 UGRD,Country Citizen,Singapore,2,2,Bachelor of Science,F,School of Science,Out of Campus,0,3,3,2,4,2,2.0,2.0,11,Maintain a well balance between academic and non-academic life.
13 UGRD,Foreigner,Malaysia,2,2,Bachelor of Science,M,School of Science,Halls,5,2,4,4,4,2,3.0,1.0,12,Doing what I like and being happy
14 UGRD,Foreigner,Malaysia,3,3,Bachelor of Engineering,F,School of Engineering,Residential Colleges,2,2,4,4,6,2,2.0,1.0,13,A good balance of academic and non-academic life
15 UGRD,Foreigner,Hong Kong,4,4,Bachelor of Science (Pharmacy),F,School of Science,Out of Campus,2,1,1,2,5,1,2.0,2.0,14,
16 UGRD,Foreigner,Malaysia,2,2,Bachelor of Engineering,M,School of Engineering,Residences,1,1,1,1,1,3,3.0,1.0,15,"An ideal student life is when you have your time well-spent both academically and non-aca
17 GRAD,Foreigner,Indonesia,1,1,Master of Laws,F,School of Law,Out of Campus,0,2,3,2,2,2,3.0,2.0,16,-
18 UGRD,Country Citizen,Singapore,3,3,Bachelor of Business Admin,M,School of Business School,Halls,3,2,2,4,4,2,2.0,2.0,17,A balanced combination between formal learning and life-long skills learning throu
19 UGRD,Foreigner,Indonesia,3,3,Bachelor of Engineering,F,School of Engineering,Residences,1,2,3,2,3,1,2.0,2.0,18,Study-life balance
20 UGRD,Foreigner,Viet Nam,4,4,Bachelor of Engineering,M,School of Engineering,Halls,0,1,3,2,5,2,3.0,2.0,19,
21 UGRD,Foreigner,Viet Nam,1,1,Bachelor of Engineering,M,School of Engineering,Halls,1,0,2,2,2,2,2.0,2.0,20,"""Interact with people and learn new things everyday"
22 UGRD,Foreigner,Viet Nam,2,2,Bachelor of Science (SDE),F,School of Design & Environment,Halls,3,3,3,3,4,2,2.0,2.0,21,One that is not academically pressurising so that students will come to learn for the sak
23 UGRD,Country Citizen,Singapore,2,2,Bachelor of Arts,F,School of Arts & Social Science,Out of Campus,10,1,3,4,3,2,2.0,2.0,22,"An ideal student life is one where a student never feels stressed simply becaus
24 UGRD,Permanent Resident,Bangladesh,2,2,Bachelor of Science,F,School of Science,Out of Campus,0,1,4,7,4,2,1.0,2.0,23,full of success and peace
25 GRAD,Foreigner,India,1,1,Doctor of Philosophy (FoS),F,School of Science,Halls,0,0,6,2,1,2,2.0,2.0,24,One where there is a balance between academics and non-academics pursuit.
```

```
Q1-Volunteered For Animal welfare,Q1-Volunteered For Arts/Culture/Heritage,Q1-Volunteered For Children/Youth,Q1-Volunteered For Community building,Q1-Volunteered For Diversity & Inclusion,Q1 Volunteered For Environmental sustainability,Q1 Volunteered For Families,Q1 Volunteered For Health/Well-being (e.g mental health),Q1-Volunteered For Seniors,Q1-Volunteered For Poverty reduction,Q1-Volunteered For Education,Q1-Volunteered For Others,Q2-Participated In
1  Q1-Volunteered For Animal welfare,Q1-Volunteered For Arts/Culture/Heritage,Q1-Volunteered For Children/Youth,Q1-Volunteered For Community building,Q1-Volunteered For Diversity & Inclusion,Q1-V
2  ,,,,,,,,No,Yes,No,No,No,No,Yes,Yes,No,Yes,No,No,No,No,No,No,No,No,Yes,No,No,No,No,No,No,No,No,Yes,No,No,No,No,No,No,No,No,1
3  ,,,,,,,,No,Yes,No,No,No,Yes,No,No,Yes,No,No,No,No,No,No,Yes,No,Yes,No,No,No,No,Yes,No,No,No,No,Yes,No,No,No,No,No,No,No,No,2
4  No,No,No,Yes,No,No,No,Yes,No,No,Yes,No,No,No,Yes,No,No,No,No,No,Yes,No,No,No,No,No,Yes,No,No,Yes,No,Yes,Yes,No,No,Yes,No,No,No,Yes,No,No,No,No,No,Yes,No,No,3
5  No,Yes,Yes,Yes,No,No,No,No,No,No,No,Yes,Yes,Yes,Yes,No,No,Yes,Yes,No,No,No,Yes,No,No,No,No,Yes,No,No,No,Yes,No,No,Yes,No,No,Yes,Yes,No,Yes,Yes,Yes,Yes,No,Yes,No,4
6  No,No,Yes,Yes,No,No,No,No,Yes,No,Yes,No,Yes,No,Yes,Yes,No,No,Yes,Yes,No,No,Yes,Yes,No,No,No,Yes,Yes,No,Yes,No,Yes,No,No,Yes,No,No,No,Yes,Yes,No,Yes,No,No,No,Yes,No,5
7  No,No,Yes,Yes,No,No,No,No,No,No,No,No,Yes,No,Yes,No,No,Yes,Yes,No,Yes,No,No,No,No,Yes,Yes,No,No,No,No,No,No,No,No,No,No,No,No,No,Yes,No,No,Yes,No,No,No,No,No,6
8  No,No,No,Yes,No,No,No,No,No,No,No,No,No,Yes,No,No,Yes,No,Yes,No,No,Yes,No,Yes,Yes,No,No,Yes,Yes,No,No,No,No,No,No,Yes,No,No,No,No,Yes,Yes,No,No,No,No,Yes,No,No,7
9  No,No,No,No,Yes,No,No,No,No,No,No,Yes,No,Yes,Yes,No,No,"Societies and Interest Groups, NUSSU and Clubs, Others",Yes,Yes,Yes,No,No,Yes,Yes,No,No,No,No,No,No,No,No,No,No,No,Yes,No,No,Yes,No,No,No
10 ,,,,,,,,Yes,Yes,No,No,No,No,Yes,Yes,No,No,No,Yes,No,No,No,No,No,No,No,No,No,No,No,No,Yes,No,No,Yes,Yes,No,No,Yes,Yes,No,No,Yes,No,9
11 No,No,Yes,No,No,No,No,No,No,No,Yes,No,No,No,Yes,No,No,Yes,Yes,Yes,No,No,No,No,No,Yes,Yes,No,No,No,No,No,No,No,No,No,No,No,Yes,Yes,Yes,No,No,No,No,Yes,Yes,No,10
12 ,,,,,,,,Yes,No,Yes,Yes,No,Yes,Yes,No,No,No,No,Yes,No,No,No,Yes,Yes,No,No,No,No,No,No,No,No,No,No,Yes,Yes,Yes,No,No,No,Yes,No,No,11
13 No,No,Yes,No,Yes,No,No,Yes,Yes,No,Yes,No,No,Yes,No,Yes,No,Yes,Yes,No,Yes,No,No,No,Yes,No,No,No,Yes,No,No,No,Yes,No,Yes,No,Yes,No,No,Yes,Yes,No,No,No,No,No,Yes,Yes,No,12
14 No,No,Yes,No,No,No,No,No,No,No,Yes,No,Yes,No,No,Yes,No,No,Yes,Yes,No,Yes,No,Yes,No,No,No,Yes,Yes,No,No,No,Yes,Yes,No,No,No,No,Yes,Yes,Yes,No,Yes,Yes,No,Yes,No,No,13
15 No,No,Yes,No,No,No,No,Yes,No,No,No,No,Yes,No,No,No,Yes,No,No,No,No,No,No,No,No,Yes,No,No,No,No,Yes,No,No,No,No,No,No,Yes,Yes,Yes,Yes,No,No,Yes,No,No,No,14
16 Yes,No,No,No,No,No,No,No,No,No,No,No,No,No,No,Yes,No,No,No,No,Yes,No,No,No,Yes,No,No,No,No,No,No,No,No,No,No,No,No,No,Yes,No,No,No,No,No,No,No,No,No,No,15
17 ,,,,,,,,No,Yes,No,Yes,No,Yes,No,No,No,Yes,No,Yes,No,Yes,No,No,Yes,No,No,No,No,No,No,No,No,No,Yes,No,No,No,No,No,No,No,Yes,No,No,16
18 No,No,Yes,No,Yes,No,No,No,Yes,No,No,No,No,Yes,Yes,No,No,No,No,No,Yes,No,Yes,No,No,No,No,Yes,No,Yes,No,No,No,No,Yes,Yes,No,No,No,Yes,Yes,No,No,Yes,No,Yes,No,No,17
19 No,No,No,Yes,No,No,No,No,No,No,No,No,No,Yes,Yes,No,No,No,Yes,No,No,No,Yes,Yes,No,No,No,No,Yes,No,No,No,No,Yes,No,No,No,Yes,No,No,No,Yes,No,No,Yes,No,No,18
20 ,,,,,,,,No,No,Yes,No,No,No,Yes,No,Yes,No,Yes,No,No,No,No,No,No,Yes,No,Yes,No,No,No,No,No,No,No,Yes,Yes,No,Yes,Yes,No,Yes,No,No,19
21 No,No,No,No,No,No,No,No,Yes,No,No,No,No,No,No,No,No,No,Yes,No,No,No,No,No,Yes,No,No,No,No,No,No,No,No,No,Yes,Yes,No,No,No,No,No,Yes,No,No,No,Yes,No,No,20
22 No,Yes,Yes,No,Yes,No,No,No,No,No,No,No,Yes,No,Yes,Yes,No,Yes,No,No,No,No,Yes,Yes,No,No,Yes,No,Yes,No,Yes,No,No,No,No,No,No,No,Yes,Yes,No,No,No,No,Yes,Yes,No,21
23 No,Yes,Yes,Yes,Yes,Yes,Yes,Yes,Yes,Yes,Yes,No,Yes,No,No,No,No,Yes,No,No,Yes,No,No,Yes,No,No,Yes,Yes,No,No,No,Yes,Yes,No,No,No,No,No,No,Yes,Yes,No,No,No,Yes,No,No,No,22
24 ,,,,,,,,No,No,No,No,Others,Yes,No,No,Yes,No,Yes,Yes,No,Yes,No,Yes,No,Yes,No,Yes,Yes,No,Yes,Yes,No,No,No,Yes,Yes,Yes,No,No,No,Yes,No,No,23
25 ,,,,,,,,No,No,No,No,No,Yes,Yes,Yes,No,Yes,Yes,Yes,No,No,No,No,No,No,Yes,No,No,No,Yes,No,No,No,No,Yes,No,No,No,No,No,No,No,24
```

Figure 2. Raw Data

The data attributes used include nationality, department, year of study, gender, level of stress, as well as several surveys regarding stressors such as stress about adjustment, academic, financial, family, friendships, romantic relationships, health related issues, career, involvement. in clubs, and about others.

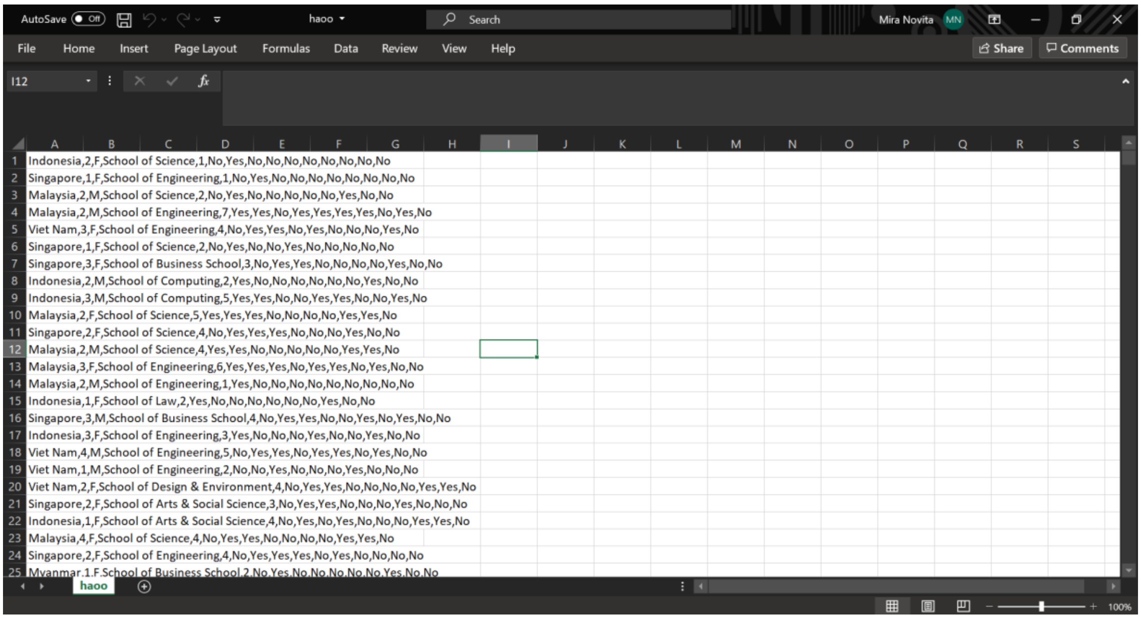
2.3. Preprocessing Data

Preprocessing is a step taken to process the initial data so that it gives good results when processed with the method that has been chosen. Data preprocessing or data preparation can also be interpreted as steps taken to clean data and convert data so that it has the same standard [9]. In the preprocessing stage, attribute reduction is carried out to sort out which data will be used in the clustering process, data cleaning, and encoding labels. Because the raw data has a lot of attributes, it is necessary to reduce attributes to select which attributes are needed in the clustering process. Before heading to the attribute reduction process, it is necessary to integrate data first to combine the two raw data files. Data integration is carried out on Microsoft Excel.

After the data is integrated, the next process is to reduce attributes. The attributes chosen are attributes related to the research topic, where in this study, attributes related to internal and external factors that cause stress in students are needed. Here it can be seen that the data attributes used are nationality, year of study, gender, department, stress level, and several causes of stress which will be described in detail in **Table 1**.

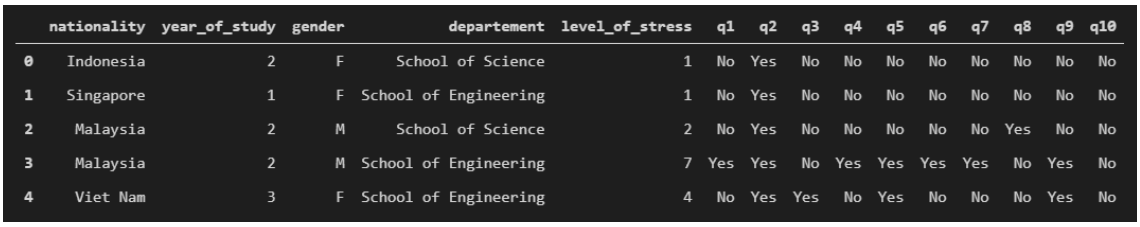
Table 1. Attributes for the Clustering Process

Attributes	Explanation
nationality	Student nationality
year_of_study	Length of study (1-5 years)
gender	Student gender
departement	Department where students study
level_of_stress	Stress level (levels 1-9)
q1	Stressed about Adjustment issues
q2	Stressed about Academic issues
q3	Stressed about Financial issues
q4	Stressed about Family issues
q5	Stressed about Friendships
q6	Stressed about Romantic relationships
q7	Stressed about Health related issues
q8	Stressed about Career related issues
q9	Stressed about My involvement in hostel, clubs, societies, interest groups.
q10	Stressed about Others



```
1  Indonesia,2,F,School of Science,1,No,Yes,No,No,No,No,No,No,No,No
2  Singapore,1,F,School of Engineering,1,No,Yes,No,No,No,No,No,No,No,No
3  Malaysia,2,M,School of Science,2,No,Yes,No,No,No,No,No,Yes,No,No
4  Malaysia,2,M,School of Engineering,7,Yes,Yes,No,Yes,Yes,Yes,Yes,No,Yes,No
5  Viet Nam,3,F,School of Engineering,4,No,Yes,Yes,No,Yes,No,No,No,Yes,No
6  Singapore,1,F,School of Science,2,No,Yes,No,No,Yes,No,No,No,No,No
7  Singapore,3,F,School of Business School,3,No,Yes,Yes,No,No,No,No,Yes,No,No
8  Indonesia,2,M,School of Computing,2,Yes,No,No,No,No,No,No,Yes,No,No
9  Indonesia,3,M,School of Computing,5,Yes,Yes,No,No,Yes,Yes,No,No,Yes,No
10 Malaysia,2,F,School of Science,5,Yes,Yes,Yes,No,No,No,No,Yes,Yes,No
11 Singapore,2,F,School of Science,4,No,Yes,Yes,Yes,No,No,No,Yes,No,No
12 Malaysia,2,M,School of Science,4,Yes,Yes,No,No,No,No,No,Yes,Yes,No
13 Malaysia,3,F,School of Engineering,6,Yes,Yes,Yes,No,Yes,Yes,No,Yes,No,No
14 Malaysia,2,M,School of Engineering,1,Yes,No,No,No,No,No,No,No,No,No
15 Indonesia,1,F,School of Law,2,Yes,No,No,No,No,No,No,Yes,No,No
16 Singapore,3,M,School of Business School,4,No,Yes,Yes,No,No,Yes,No,Yes,No,No
17 Indonesia,3,F,School of Engineering,3,Yes,No,No,No,Yes,No,No,Yes,No,No
18 Viet Nam,4,M,School of Engineering,5,No,Yes,Yes,No,Yes,Yes,No,Yes,No,No
19 Viet Nam,1,M,School of Engineering,2,No,No,Yes,No,No,No,Yes,No,No,No
20 Viet Nam,2,F,School of Design & Environment,4,No,Yes,Yes,No,No,No,No,Yes,Yes,No
21 Singapore,2,F,School of Arts & Social Science,3,No,Yes,Yes,No,No,No,Yes,No,No,No
22 Indonesia,1,F,School of Arts & Social Science,4,No,Yes,No,Yes,No,No,No,Yes,Yes,No
23 Malaysia,4,F,School of Science,4,No,Yes,Yes,No,No,No,No,Yes,Yes,No
24 Singapore,2,F,School of Engineering,4,No,Yes,Yes,Yes,No,Yes,No,No,No,No
25 Myanmar,1,F,School of Business School,2,No,Yes,No,No,No,No,No,Yes,No,No
```

Figure 3. Data used for clustering in .csv Format

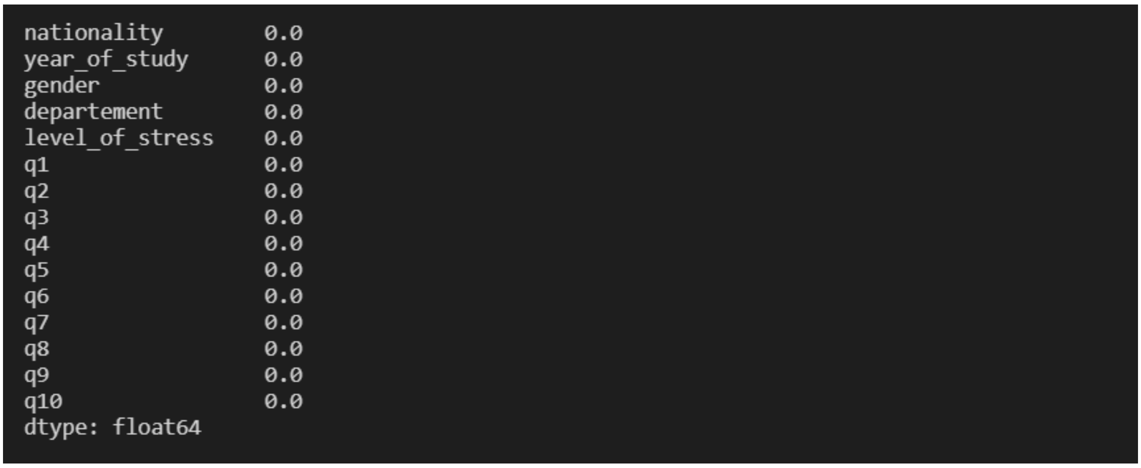


```
  nationality  year_of_study  gender            departement  level_of_stress   q1   q2   q3   q4   q5   q6   q7   q8   q9  q10
0   Indonesia              2       F      School of Science                1   No  Yes   No   No   No   No   No   No   No   No
1   Singapore              1       F  School of Engineering                1   No  Yes   No   No   No   No   No   No   No   No
2    Malaysia              2       M      School of Science                2   No  Yes   No   No   No   No   No  Yes   No   No
3    Malaysia              2       M  School of Engineering                7  Yes  Yes   No  Yes  Yes  Yes  Yes   No  Yes   No
4    Viet Nam              3       F  School of Engineering                4   No  Yes  Yes   No  Yes   No   No   No  Yes   No
```

Figure 4. Display of .csv Data When Imported and Ready for the clustering process

p-ISSN: 2301-5373
e-ISSN: 2654-5101

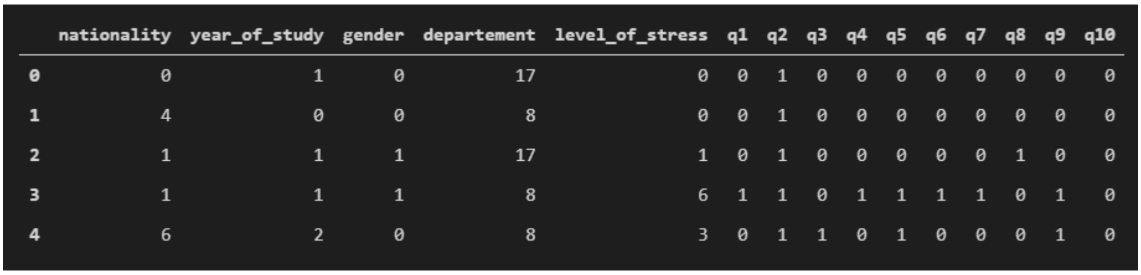
Data cleaning is the process of removing noise and inconsistent or irrelevant data [10]. Data cleaning is also related to the process of analyzing the quality of data because if the data to be used is of poor quality, it will affect the final results of the data mining process (in this study the clustering process). In this study, the authors have checked the data and made sure the data is clean and ready for the clustering process. From the results of checking, there is no data that is NULL.



```
nationality        0.0
year_of_study      0.0
gender             0.0
departement        0.0
level_of_stress    0.0
q1                 0.0
q2                 0.0
q3                 0.0
q4                 0.0
q5                 0.0
q6                 0.0
q7                 0.0
q8                 0.0
q9                 0.0
q10                0.0
dtype: float64
```

Figure 5. NULL Check Results

At this stage, an encoding label is also carried out for the data to be used. Label encoding is a process carried out to change the form of data into a form that can be understood by machines so that the data can be processed and processed by the method or algorithm.



	nationality	year_of_study	gender	departement	level_of_stress	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
0	0	1	0	17	0	0	1	0	0	0	0	0	0	0	0
1	4	0	0	8	0	0	1	0	0	0	0	0	0	0	0
2	1	1	1	17	1	0	1	0	0	0	0	0	1	0	0
3	1	1	1	8	6	1	1	0	1	1	1	1	0	1	0
4	6	2	0	8	3	0	1	1	0	1	0	0	0	1	0

Figure 6. Label Encoding Results

2.4. Proposed Method

K-modes clustering was first introduced by Huang in 1998 as a clustering method modified from the k-means method. K-means is a simple unsupervised algorithm whose results are quite good for general clustering problems. However k-means usually works on attributes or data with numeric values, not categorical values. Therefore, k-means is modified to be used for categorical data. The result of this modification is called k-modes. Modifications made to the k-means method are:

a. The distance between two data points X and Y is the number of features in X and Y whose values are different (simple dissimilarity measure), formally formulated as follows:

$$d_1(X,Y) = \sum_{j=1}^{m} \delta\,(x_j, y_j) \tag{1}$$

where

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

Explanation:

$d(x,y)$ = data distance from x to y
x_j = feature value j from x

y_j = feature value j from y
m = number of features

b. Change the means to modes.
c. Use frequency to search mode. The mode is the data value that appears the most. The formation of the centroid is by looking for the mode of each feature.

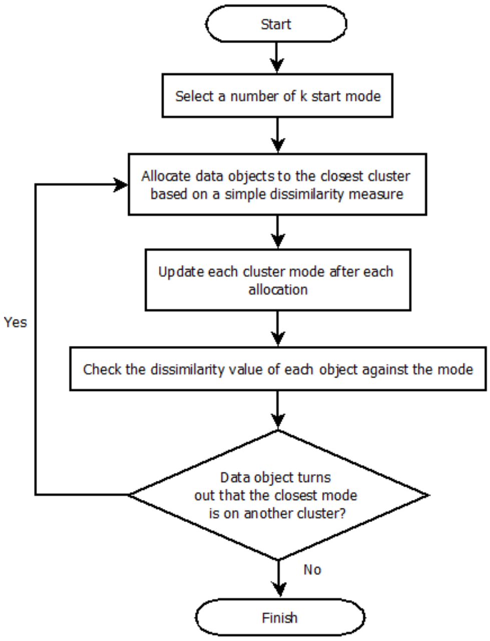

Start
Select a number of k start mode
Allocate data objects to the closest cluster based on a simple dissimilarity measure
Update each cluster mode after each allocation
Check the dissimilarity value of each object against the mode
Data object turns out that the closest mode is on another cluster?
Yes
No
Finish

Figure 7. K-modes Flowchart

The following are the steps for k-modes clustering by Huang:

a. Select a starting mode number k
b. Allocate data objects to the closest cluster based on a simple dissimilarity measure. Update each cluster mode after each allocation.
c. After all data objects have been allocated to a cluster, check the dissimilarity value of each object against the mode. If a data object turns out to be the closest mode to another cluster, move the object to the appropriate cluster and update the second cluster mode.
d. Repeat step 3 until none of the data objects change clusters.

2.5. Optimization of the Number of Clusters with the Elbow Method

The elbow method is a method to determine the right number of clusters through the percentage of the comparison between the number of clusters that will form an elbow at a point [11]. The elbow method plots the value of the cost function produced by different values of k [12]. The different percentage results from each cluster value can be shown using a graphic as the source of information. The principle of this elbow method is to select the cluster value and then add the cluster value to be used as a data model in determining the best cluster. When the cluster value has the greatest decline and forms an angle, the number of cluster values is said to be the most appropriate value.

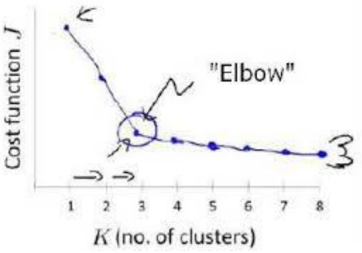


Figure 8. Graph of K Value

Elbow method is the process of finding the optimal cluster that works through the calculation of the cost function. For K-means, the cost is defined as the sum of squares error within the cluster and gives information on how scattered the points from a cluster are. Therefore, the lower the cost, the nearer the points in the cluster. Nevertheless, to compute the cost for K-modes the Euclidean distance has to be replaced for the Hamming distance. By plotting the cost function against the number of clusters an elbow should be found [13].

$$cost = \sum_{i=1}^{n} \sum_{i=1}^{k} d_{xc} \tag{2}$$

Explanation:

K = cluster c
X_1 = distance object to i
C_k = cluster center i

3. Result and Discussion

The clustering process is carried out using the Python programming language. The algorithm used is the k-modes clustering algorithm which Huang has previously researched. This algorithm is suitable for clustering categorical data. After doing the research stage of data collection, integration, data cleaning, label encoding, clustering with k-modes, and optimization of the number of clusters with the elbow method. The optimal number of clusters obtained from the implementation of the elbow method is three clusters due to a drastic decrease in the number. The elbow chart can be seen in the image below:

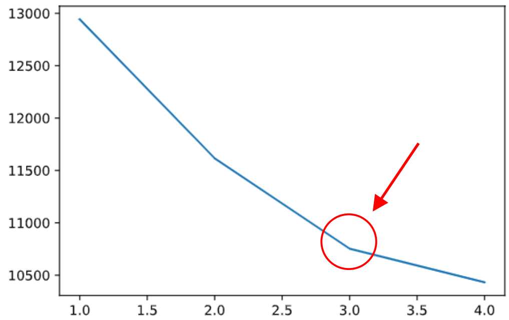
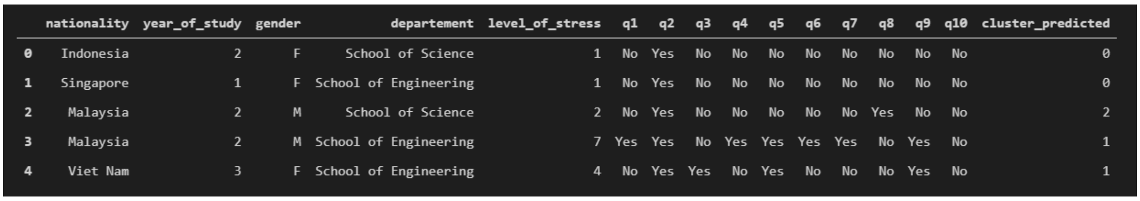


Figure 9. Elbow Method Graph

The total amount of imported data is 2663 data from students who are currently studying in several countries in Southeast Asia. After clustering with the k-modes algorithm, the number of clusters is 3, the following results are obtained: cluster 1 contains 1666 students, cluster 2 contains 393 students, and cluster 3 contains 604 students.



```
  nationality year_of_study gender           departement level_of_stress  q1   q2   q3   q4   q5   q6   q7   q8   q9  q10 cluster_predicted
0   Indonesia             2      F     School of Science               1  No  Yes   No   No   No   No   No   No   No   No                 0
1   Singapore             1      F School of Engineering               1  No  Yes   No   No   No   No   No   No   No   No                 0
2    Malaysia             2      M     School of Science               2  No  Yes   No   No   No   No   No  Yes   No   No                 2
3    Malaysia             2      M School of Engineering               7 Yes  Yes   No  Yes  Yes  Yes  Yes   No  Yes   No                 1
4    Viet Nam             3      F School of Engineering               4  No  Yes  Yes   No  Yes   No   No   No  Yes   No                 1
```

Figure 10. Final Data After the Clustering Process

After obtaining clustering results using the k-modes algorithm, the next step is to identify the characteristics of each cluster. To facilitate identification, the results of clustering can be visualized in advance in graphic form so that it is easier to analyze and understand. The following are the results of the identification of each cluster:

Table 2. Identification of each cluster

Cluster	Explanation
Cluster 1	Cluster 1 is a cluster that has the most members among other clusters. This cluster has 1667 members. This cluster contains groups of students, the

	majority of whom are from Singapore, Malaysia, Indonesia, Vietnam, and Myanmar. Compared to other clusters, cluster 1 has the highest number of female students compared to other clusters. In this cluster, female gender dominates. Judging from the year of study, this cluster is a collection of early-level students, the majority of whom take the school of art and social science department. Apart from these departments, there are also many students from the school of engineering and school of computer science departments who are included in this cluster. The stress level of students in this cluster is a mild level because the majority are at levels 1, 2, and 3. The main cause of stress in this cluster of students is academic problems. This cluster is the only group where the majority of the cause of stress is not financial.
Cluster 2	Cluster 2 is a cluster that has the fewest members, namely 393 members. This cluster contains groups from Singapore, Malaysia, Indonesia, Vietnam, and Thailand. This group is predominantly female. Judging from the year of study, this group of students is spread from beginning to end students. The department that dominates this cluster is the department school of science. After that there are quite a number of schools of art and social science, schools of engineering, schools of business, schools of design and environment. The stress level for the cluster is a high level, the majority is at level 6 and some are at level 9. The causes of stress in this cluster are due to adjustment, academic, financial, and friendship. This cluster is the only cluster whose cause of stress is on careers and on involvement in hostels, clubs, and society.
Cluster 3	Cluster 3 is a cluster that has 603 members. This cluster contains members from Singapore, Malaysia, Indonesia, and Vietnam. This cluster is dominated by male students. This cluster is dominated by students who are in the middle of their study period because it is seen from the year of study that the majority are in the second year. The level of stress in this cluster is medium. The main causes of stress at this level are academic and financial.

4. Conclusion

A clustering experiment has been carried out using the k-modes method on the data that causes stress in students. The data used for this study were 2663 data obtained from the Kaggle website in the form of .csv file format. The data used are the results of a survey regarding the ideal student life factors that are important in student life. Before clustering, the raw data is preprocessed so that later it can be clustered properly. In this study, a cluster was conducted with 3 clusters obtained from the implementation of the elbow method. Cluster 1 has 1666 members, cluster 2 has 393 members, and cluster 3 has 604 members.

Cluster 1 is a cluster with a mild stress level, the main cause of stress is academic issues. Cluster 1 is the only group where the majority of the cause of stress is not financial. Cluster 2 is a cluster with a high stress level which causes various stressors. However, cluster 2 is the only cluster where the cause of stress is on careers and on involvement in hostels, clubs, and society. Cluster 3 is a cluster with a medium stress level. This cluster is the only cluster dominated by male gender. The main cause of stress in this cluster is academic and financial.

References

[1] N. T. Lumban Gaol, “Teori Stres: Stimulus, Respons, dan Transaksional” *Buletin Psikologi*, Vol. 24, No. 1, pp.1-11, 2016

[2] R. D. Rahmayani, R. G. Liza, N. A. Syah, “Gambaran Tingkat Stres Berdasarkan Stressor pada Mahasiswa Kedokteran Tahun Pertama Program Studi Profesi Dokter Fakultas Kedokteran Universitas Andalas Angkatan 2017” *Jurnal Kesehatan Andalas*, Vol. 8, No. 1, pp.103-111, 2019

[3] L. Bismala, “Analisis Perbedaan Beban Stress Pada Mahasiswa Laki-laki dan Perempuan yang Sedang Menyusun Skripsi” *Jurnal Program Studi Akuntansi*, Vol. 1, No. 1, 2015

[4] Rahmawati, M. Faisal, “Analisis Cluster untuk Pengelompokan Desa Berdasarkan Indikator Penyakit Diare” *SAINTIFIK*, Vol. 5, No. 1, pp75-80, 2019

[5] F. S. Jumeilah, D. Pratama, “Identifikasi Cluster Penduduk Usia Kerja Pada Provinsi Sumatera Selatan Menggunakan K-Modes” *Jurnal Komputer Terapan*, Vol. 4, No. 1, pp.1-9, 2018

[6] Gustientiedinaa, M. H. Adiyaa, Y. Desnelitab, “Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru” *Jurnal Nasional Teknologi dan Sistem Informasi*, Vol. 5, No. 1, pp.017-024, 2019

[7] F. N. Marleny1, H.M. Junaidi, Mambang, “Penerapan K-Means Cluster Untuk Pengaruh Kecerdasan Emosi Dan Stres Terhadap Prestasi Belajar Mahasiswa” *Seminar Nasional Teknologi Informasi dan Multimedia*, 2015

[8] T. Yulianita, D. Istiawan, “Implementasi Algoritma K-modes untuk Penentuan Prioritas Rehabilitasi Daerah Aliran Sungai Berdasarkan Parameter Lahan Kritis”, The 6th University Research Colloquium 2017, pp. 429-440

[9] F. A. Nugraha, N. H. Harani, R. Habibi, Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning, Bandung: Kreatif Industri Nusantara, 2020, pp. 165

[10] R. R. Rerung, “Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk” *Jurnal Teknologi Rekayasa*, Vol. 3, No. 1, pp. 89-98, 2018

[11] D. A. I. Cahya Dewi, D. A. K. Paramita, “Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali” *Jurnal Matrix*, Vol. 9, No. 3, pp. 102-109, 2019

[12] E. Muningsih, S. Krisnawati, “Sistem Aplikasi Berbasis Optimasi Metode Elbow Untuk Penentuan Clustering Pelanggan” *JOUTICA*, Vol. 3, No. 1, 2018

[13] Neus Llop Torrent, “The K-modes algorithm applied to Gender Analysis”, Treball de Fi de Grau, Barcelona, 2019

This page is intentionally left blank