School of Computer Sciences

CDS590 – Consultancy Project & Practicum

Final Report

An Investigation of Machine Learning Usage on Speech Disorder Diagnosis Tools: Malay Preschool Language Assessment Tool (MPLAT)

[MUHAMMAD SYAZWAN BIN RUSDI]
[P-COM0085/19]

Supervisor: Profesor Madya Dr. Manmeet Kaur A/P Mahinderjit Singh

Mentor: Prof. Dr. Nor Adnan Yahaya

Practicum place: UniLangIT Academy Sdn. Bhd.

SEM 1 2021/2022

DECLARATION

“I declare that the following is my own work and does not contain any ***unacknowledged*** work from any other sources. This project was undertaken to fulfil the requirements of the Consultancy Project & Practicum for the Master of Science (Data Science and Analytics) program at Universiti Sains Malaysia”.

Signature	:	…………………………………
Name	:	Muhammad Syazwan bin Rusdi
Date	:	1st February 2022

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Professor Dr. Nor Adnan Yahaya, founding Director of LangIT Education & Consultancy and as known my client and mentor for providing me with an opportunity to do my internship at LangIT. I sincerely thank Dr Rogayah A Razak, another founding director for her guidance about speech disorder.

I would like to express my special thanks of gratitude to my supervisor Profesor Madya Dr. Manmeet Kaur as well as our lecturer who gave me the golden opportunity to do this project on the topic An Investigation of Machine Learning Usage on Speech Disorder Diagnosis Tools: Malay Preschool Language Assessment Tool (MPLAT), which also helped me in doing a lot of projects and I came to know about so many new things. I am thankful to them.

I also like to thank my parents and friends who helped me a lot in finishing this project within the limited time. I am making this project not only for marks but to also increase my knowledge.

ABSTRACT

Industrial training is an important phase of a student life. A well planned, properly executed, and evaluated industrial training helps a lot in developing a professional attitude. It develops an awareness of industrial approach to problem-solving using Data Science technique. UniLangit Sdn. Bhd. has their assessment tool to diagnose speech disorder which is Malay Preschool Language Assessment Tool (MPLAT). However, this tool is having some problems. They don’t know if there another subgroups or clusters beside non-disorder and disorder in MPLAT data set. MPLAT dataset doesn't has labelled/class features that identify the patients which unable to provide a good analysis. The objectives of this project are identifying the features essential for speech disorder diagnosis, determining the pattern of subgroup based on the dataset by using clustering algorithms, and evaluating clustering algorithms.

For this project, we start with selection features. We select features based on correlation between two features and variance. After that, we use K-means and hierarchical clustering to find optimal number of clusters and dividing data into several cluster. Finally, we evaluate clustering methods by using silhouette coefficient, Dunn index and Davies Bouldin index. As result of project, four features which are Grammatical Understanding (GU), Referential Meaning (REFMNG), Relational Meaning (RELMNG) and Early Literacy Skills (EL) are selected for clustering based on selection methods. Then, REFMNG and RELMNG influence on K-means clustering while REFMNG and EL influence on hierarchical clustering. After that, sampling method make a bit of improvement on clustering. Finally, K-means clustering is much better than hierarchical in term of density of each cluster based on Silhouette coefficient but hierarchical is a bit better than K-means in term of separation distances based on Dunn index.

ABSTRAK

Latihan industri merupakan fasa yang penting untuk perjalanan pelajar. Perancangan yang bagus, pelaksanaan yang lancar dan penilaian pada latihan industri banyak menolong dalam pembangunan sifat profesional. Latihan industri dapat meningkatkan kesedaran pada pendekatan industri untuk penyelesaian masalah dengan menggunakan teknik Sains Data. UniLangit Sdn. Bhd, mempunyai alat penilaian untuk diagnosis kecelaruan pertuturan iaitu MPLAT. Walau bagaimanapun, alat ini mempunyai beberapa masalah. Mereka tidak tahu jika terdapat sub-kumpulan atau kluster selain daripada kumpulan penghidap dengan tidak menghidap dalam set data MPLAT. Set data MPLAT tidak mempunyai atribut kelas untuk mengenal pasti pesakit di mana ketiadaan atribut ini tidak dapat menyediakan analisis yang bagus. Objektif projek ini ialah pengenalpastian ciri-ciri yang penting untuk diagnosis kecelaruan pertuturan, penentuan corak sub-kumpulan berdasarkan pada set data dengan menggunakan algoritma pengelompokan, dan penilaian pada algoritma pengelompokan. Untuk projek ini, kami memulakan dengan pemilihan ciri-ciri. Kami memilih ciri-ciri berdasarkan korelasi antara dua ciri dan variasi. Selepas itu, kami menggunakan pengelompokan K-purata dan berhierarki. Akhir sekali, kami menilai kaedah pengelompokan dengan menggunakan Koefisien Siluet, indeks Dunn dan indeks Davies Bouldin. Hasil daripada projek ini, empat ciri iaitu Pemahaman Tatabahasa (GU), Makna Rujukan (REFMNG), Makna Hubungan (RELMNG) dan Kemahiran Literasi Awal (EL) dipilih untuk pengelompokan berdasarkan kaedah pemilihan. Selepas itu, REFMNG dan RELMNG memberi kesan pada pengelompokan K-purata sambil REFMNG dan EL memberi kesan pada pengelompokan berhierarki. Akhir sekali, pengelompokan K-purata lebih baik daripada berhierarki dari segi ketumpatan setiap kluster berdasarkan pada Koefisien Siluet tetapi pengelompokan berhierarki lebih bagus daripada K-purata dari segi jarak pemisahan antara kluster-kluster berdasarkan indeks Dunn.

TABLE CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Background of Company

UniLangIT Academy Sdn. Bhd. or as known as LangIT Education & Consultancy is a Consultancy business that operate at Kota Damansara, Selangor. Based on its name LangIT, it is a words combination of language (Lang) and information technology (IT). This company provides speech therapist and IT training program. Their mission is to create a platform for educationists, health care providers such as speech-language therapists, psychologists, and occupational therapists. They also have technology experts to collaborate in training, research, development, and consultancy in the following areas. For project internship, this will be conduct under client which is director of LangIT and IT development department. They provide the consultation to everyone regarding Information and Communication Technology (ICT) skills, software engineering, statistic, and web development and computer graphics.

1.2. Background of Study

A speech disorder is a disability of the expression or communication in term of speech sounds, fluency, or voice. (Franciscatto, Trois, Lima, & Augustin, 2018). By using formal standardized language assessments tool, it can identify either children have speech disorder or not. One of UniLangIT's founders had create assessment tool for diagnosis speech disorder in Malay language which is Malay Preschool Language Assessment Tool (MPLAT). MPLAT is an assessment tool that can diagnose speech disorder by measuring receptive, expressive use of language and early literacy skills to preschool children.

For this project internship, client stated their problem. They investigate if machine learning can identify speech disorder through MPLAT assessment tool without therapist and if they can reveal more groups rather than just two groups. They ask to analyse MPLAT diagnoses dataset that had used to get standardized for MPLAT assessment tool. Unsupervised machine learning is purposed due to MPLAT dataset does not has labelled feature.

Unsupervised is one of two main types of machine learning which it can identify groups or pattern in multivariate data when those data only have less knowledge. (Shahapurkar & Sundareshan, 2004). It can help discover similarities and differences in information where can find ideal solutions for exploratory data analysis. (Education, 2020). It very suitable this is ideal for studies that want to explore subgroups or clusters that can be created in datasets that do not have a labelled class. (Hussain, Atallah, Kamsin, & Hazarika, 2018). There are two common unsupervised learning which are clustering and association rule learning. The clustering learning is chosen rather than association rule due to their objective which is grouping the data.

1.3. Problem Statement:

MPLAT assessment tool is used for identifying children with speech disorder. They also want to divide the data into several groups for further research, but they don't know how to divide because they lack understanding on important features and no pattern being defined within the dataset. Those problem are similar with some study cases. (Bóna & phonetics, 2019; Gardner-Hoag et al., 2021; Stevens et al., 2017). They also still wondering if this MPLAT is good enough. If they can find useful knowledge or pattern from analysing of MPLAT dataset, it can help the company expand MPLAT by adding more useful features or improve existing features that can help their therapists improve their diagnosis and screening process.

1.4. Objective:

1. To identify the features essential for speech disorder diagnosis
2. To determine the pattern of subgroup based on the dataset by using clustering algorithms
3. To test and evaluate several clustering algorithms.

1.5. Benefit of Project

This project give advantage to MPLAT Research if they can reveal new subgroups. This opportunity makes UniLangIT develop new treatment procedure that suitable to each subgroup in a session so they can solve their patients. If they satisfy with this unsupervised analysis, they can develop supervised machine learning and they can use to develop MPLAT screening application so that caregivers can perform screening easily and quickly.

CHAPTER 2

RELATED WORK

2.1 Speech disorder

The basic definition of speech is ability to express thoughts and feelings by articulate sounds. If someone had difficult to express through speak, it is called speech disorder. Speech disorder is defined as an extensive disability problem. It associated with poor long-term outcomes that affect individuals, families and children's academic achievement and it will affect vocational in adulthood. The author found some commonness that affected speech disorders specifically stuttering, voice, and speech-sound disorders in the study. The study showed that gender, parental education and by number of family member have significantly different between non speech disorder and speech disorder. Then, the study also mentioned that no significant difference across speech disorders and birth order, religion, and paternal consanguinity. (Karbasi, Fallah, & Golestan, 2011)

There two common types of motor speech disorder which are apraxia and dysarthria. Based on study by Duffy, dysarthria is a motor speech disorder which relate to of strength, speed, tone, steadiness, or accuracy of the movements of neuromuscular in face or mouth that affect the performance of speech. While apraxia is caused by damage to the parts of the brain related to speaking. (J. R. Duffy, 2000)

After that, the difference between speech delay and disorder is important to acknowledge for giving a suitable treatment. According to study done by Dodd. B, there are different between speech delay and speech disorder. The author used ANOVA test and the author found that there are significant difference between two groups. Speech disorder had small percent consonants correct and percent vowels correct. Besides that, speech disorder group made more error than delay group on consonants and vowels measures. After that, speech disorder had higher inconsistency scores then other group. The author also done Chi-square Test and its result showed children with speech delay are more to acquire non-linguistic rules than children with speech disorder. (Dodd, 2011)

2.2. Speech Disorders Diagnosis

Speech is the expression of or the ability to express thoughts and feelings by articulate sounds or speaking. If someone had difficult to express through speak, it is called speech disorder. A speech disorder is a disability of the expression or communication in term of speech sounds, fluency, or voice. (Franciscatto, Trois, Lima, & Augustin, 2018). Speech delay and speech disorder are difference. Speech delay happens when children was developing in a normal sequential pattern, but they developed later than typical while speech disorder happens when they cannot speech proper even, they developed in a normal sequential pattern. (Sense, 2002)

Some preschool children (between 4 to 6 years old) nowadays have speech delay but some of them may has speech disorder which its treating treatment is totally different with treating speech delay. To treat speech disorder, speech disorders diagnosis needs to be developed in special treatment. Recognition of a mixed dysarthria may help confirm expectations for a given disease. By using formal standardized language assessments tool, it can identify either children have speech disorder or not.

2.3. Assessment tools in Malaysia and MPLAT

Razak R.A. (2010,2018) mentioned that some therapists in Malaysia use formal standardized language assessments which were imported from other countries especially from United Kingdom and United State of America due to standardizing by healthcare organization such as Agency for Healthcare Research and Quality. But there are some problems when those assessment tools are applied in Malaysia. Firstly, those standardized language assessments were developed based on English language and western culture. English and Malay language are different in terms of the vocabulary, grammatical structures, logical reasoning, and inferences. Then, formal assessments may not give an accurate picture of the Malay child's language abilities due to different child developing's culture. Pearson (2004) supports a need for a fair language test in term of linguistically and culturally for evaluating speech and language disorders. (Razak, Madison, Siow, Aziz, & Hearing, 2010; Razak et al., 2018)

In Malaysia, there is one speech screening assessment tool in Malay language in Malaysia which is Malaysian Developmental Language Assessment Kit (MDLAK) which created in 1992. It widely used in Malaysia is based on developmental approximate norms but their issue which is it currently no standardized norm–referenced language assessment for preschool children in Malaysia. (Razak et al., 2010)

MPLAT were purposed to overcome those problems and it was successfully standardized. It contains six subtests which are picture vocabulary, grammatical understanding, sentence repetition, referential meaning, relational meaning, and early literacy skills. It has two versions which the full version (diagnostic) and the short version (screening). The diagnostic and screening versions can be used by speech-language therapists while the screening version can also be used by preschool teachers. This assessment tool has obtained a declaration of copyright invention for its intellectual content and form in 2014. Even this assessment tool be standardized, but its developer which is from UniLangIT still continue improving the speech disorder diagnosis and screening process. (Razak et al., 2010)

2.4. Unsupervised machine learning in speech disorder diagnosis

Machine learning has helped a lot in healthcare industries due to have a huge set of data of patients. Machine learning can help in term of disease diagnosis, drug discovery, clinical trial, and research. For speech therapy, machine learning can help a tachistoscopes who in charge diagnose speech disorders so they can increase diagnosis speed. For this problem, client want to determine which one have speech disorder and they want to know how many subclasses in MPLAT dataset. By using machine learning, it can analyse data to extract useful patterns and knowledge for dividing into several subclasses and then for further researching in speech disorder and therapy field. As mentioned before in background study, there are two commons unsupervised learning and clustering learning is chosen. It is because clustering learning focuses on grouping of data in same dataset based on similarity to each other while associated rule learning is used for discovering interesting relations between variables datasets. (Gardner-Hoag et al., 2021; Hussain et al., 2018; Sato, Izui, Yamada, & Nishiwaki, 2019)

There are not many studies related to the use of unsupervised in the field of speech disorder based on assessment tool data . (Chuchuca-Méndez et al., 2016). Mostly speech disorder studies are related to the analysis of voice data, or it called voice recognition. There is study about application situation awareness with machine learning to identify children's speech disorders. The author performed machine learning with 1,362 children and created a database containing the child's speech audio and its associated metadata. (Franciscatto et al., 2018).

There is study that purpose application machine learning in meta-analysis on voice disorder, which a subset of speech disorder. The author used three datasets to perform machine learning. First dataset is Saarbruecken Voice Database which is collection of voice recordings by over 2000 people. Next dataset is Massachusetts eye and ear infirmary (MEEI) that contains over 1,400 vocal tests of the long vowel. And third dataset is Arabic voice pathology database that collected voices of the patients by experienced pheneticists. (Chuchuca-Méndez et al., 2016)

Even so, there still study that used unsupervised machine learning especially clustering algorithm to analyse on structure dataset which transformed from voice recording. Bóna J. had done study about clustering of disfluencies in typical, fast, and cluttered speech. The main theme of this study is to determine patient that faced cluttered speech which it is a speech fluency disorder. The type of fluency is determined based on frequency of disfluencies which measured based on ability of person to speak both in typical and atypical speech. The more frequent occurrence of disfluencies, more possible that person has cluttered speech. The author found that previous research had difficult to find significant difference between the typical and cluttered groups in the frequency of disfluencies. In this study, participants which from cluttered, fast, and typical speakers were selected. They were asked to speak and repeat the same topic and those speaking were recorded. The expert in speech-language pathologist measured speech rate, and articulation rate from voice recording and determined the value of six features. The author used statistical analyses Kruskal–Wallis-test, Mann–Whitney-test, and UniANOVA to extract analyse value. Then, those result is used for clustering sample by using Tukey post hoc test. The result of clustering was evaluated using external clustering measurement and it showed only 66.7% of fast, 52.2% of cluttered and 71.4% of typical speech were correct which showed the accuracy of clustering model was not enough. The author concluded that this study still cannot identify the difference between those fluency groups. The author mentioned that it occurred due to lack of sample and statical analyse was not utilized properly by the author. (Bóna & phonetics, 2019)

There is study about unsupervised clustering analytics in wearable internet of things (IoT) which also transform voice recording into structure data. The study’s aim is using IoT to diagnose patient with speech disorder and applying unsupervised clustering analytics to identify speech disorder and dividing data into several groups. The authors used the SmartFog architecture and created SmartFod wearable device which it records voice and those collected data will be transferred to end user for analysing. End user will analysis and extracts the structures data or score from voice recording by using Praat scripting language. The authors used 164 speech samples and three features which is average fundamental frequency and intensity. They used K-means clustering to divide data into several groups. The result of clustering with *k* value equal to two showed the minimum separation between two clusters is huge and it means the clustering method is worked well. Then, the authors used three and four as *k* value. The problem with this study is they did not validate the clustering for three and four as *k* values and even did not validate two as *k* values in detail. (Borthakur, Dubey, Constant, Mahler, & Mankodiya, 2017)

Due to limitation of a study of unsupervised machine learning in speech diagnosis assessment tool, a study of unsupervised in autism spectrum disorder (ASD) will be used as another reference. ASD is a neurodevelopmental disease that affected to the delay of social and communication behaviours. (Baadel, Thabtah, Lu, & Care, 2020). Not all speech disorder patients have ASD, but ASD patients have speech and communication disorder. The different is ASD has speech and behavioural challenges such as they like to repeat do something, but normal speech disorder just only have speech challenges. (Hailpern, Karahalios, DeThorne, &

Halle, 2010). Most of ASD studies use assessment tool data which consist of speech challenges and behavioural screening.

Said et al. had developed predictive model for analysing autism to identify either clustering can improve accuracy for supervised machine learning algorithm. Author purposed an applied machine learning framework called Clustering-based Autistic Trait Classification (CATC) when clustering is used in pre-processing phase. Datasets consist of 1875 participated. OMCOKE algorithm is a type of clustering model used by him. The OMCOKE algorithm is based on K-means but the difference with initial K-means is OMCOKE can consider outlier or noise data in the dataset by building outlier clusters quickly and separating these outlier or noise data into those clusters. After clustering, cluster feature is created, and it is compared with true labelled class. If data's cluster is matched with true class, it will be kept from dataset. Otherwise, it will be removed from dataset. Then, he used those clustered dataset using supervised machine learning algorithm to evaluate the performance of the clustering phase and compared between models which use non-clustered dataset and model which using clustered dataset. The result showed that model using clustered dataset is better that model without clustering. During comparing Cluster and true Class after clustering phase, those data that has cluster which does not match with true class will be remove from clustering dataset. If the accuracy of clustering is high or more than 90 percent or amount of unmatched data is small, he can remove those unmatched data. If not, unfitting will be occurred. (Baadel et al., 2020).

The study about clustering machine learning analysis to reveal subgroups or clusters from autism spectrum disorder data based on behavioural phenotypes was purposed. Author released the research has yet to purpose machine learning to identify phenotypes on a large sample size to improve examine treatment response across such subgroups. 2400 clinical record were chosen for this study which are collected within 36 months. Authors use data science framework Gaussian Mixture Models which provide a statistical framework for learning clustering algorithm from data that allows for probabilistic analyses. They use Hierarchical Agglomerative Clustering to identify behavioural characteristics of ASD. The result of clustering showed there are 16 subgroup or cluster on dataset. Hierarchical Agglomerative Clustering created clusters based on similarity of skill signature, but the magnitude of each skill is variance. Based on clustering, the crucial differences of 16 cluster are mean age and degrees of severity across developmental domains. Therefore, the study found that high amount of variance give high advantages to clustering provides. The more of dataset size and feature, more knowledge and clusters will be revealed. However, this study does not focus on evaluation of this clustering machine learning. And data from dataset are collected from standardized assessment tools which authors have difficult to compare previous clustering research who use unstandardized assessment tools. (Stevens et al., 2019)

The same author had purpose study about clustering in ASD by analysing challenging behaviour features. Author did this study to evaluate the topography and functions maintaining challenging behaviours to reduce the harmful effects on skill acquisition and their life. Author used applied behaviour analysis (ABA) for this study. Author used SKILLS^{TM} database which store of treatment data that provided by a large national provider of autism treatment services. The target features for this clustering are eight challenging behaviour's score and gender feature. K-means clustering is used to analyse common behaviour profiles across challenging behaviours. Author determined k value by changing k from 1 to 20 and then selecting the value of k that match to a level in intra-cluster sum of squared distances. The result of clustering showed that most clusters are defined by a single dominant challenging. The cluster whose created by relative strengths of two challenging behaviours only cluster 5 and 6, the rest are not due to relative strengths are small. Author repeated clustering algorithm by removing gender feature and with same *k* value. The result showed the seventh cluster is good clustering result when there are five challenging behaviours have high relative strength. Those result could be occurred if mostly patient have single dominant challenging which not in accordance with the objectives of the study. The relative strengths of challenging behaviours can occur in all clusters if author consider about outline. (Stevens et al., 2017)

This study case was repeated by Hoag J.G. with difference method. (Gardner-Hoag et al., 2021). The author's purpose was the same which revealed types of ASD based on different challenging behaviours and identify differences in treatment response between groups. The difference with previous study is the author used k-means clustering and multiple linear regression and comparing those algorithms as another goal. The author used same framework which is ABA and used the dataset from SKILL^{TM} but only 854 samples were selected due to criteria. For k-means clustering, the author using different k value and those value which has smallest sum of squared errors was selected. However, the author did not encourage to use number of features as *k* value to prevent overfitting. After finding a optimise k value, seven was chosen. A multiple linear regression model analyses the correlation between the target variable by finding the weights of the equation to explain the correlation between the explanatory variable and the target variable. The mean squared error analysis were used to evaluate an efficient of regression. The result of lustring showed that only Cluster 4 and 7 had high relative strength between challenging behaviours and the other cluster showed single dominant challenging. For regression, the model showed 67% of the variance of exemplar mastery. For mean squared error, cluster 3 (elopement) and cluster 6 (aggression) had value of error below than 0.24. The highest mean squared error is cluster 4 (stereotypy low rate) which 0.37.

Table 2.1: Comparison of study cases' methods

Author	Framework	Method	Result	Weakness
Bóna & phonetics, 2019	none	Tukey post hoc test	- 66.7% of fast, 52.2% of cluttered and 71.4% of typical speech were correct	- Lack of sample - Statical analyse was not utilized properly by the author
Borthakur, Dubey, Constant, Mahler, & Mankodiya, 2017	SmartFog architecture	K-means clustering	-For k=2, minimum separation between two clusters is huge - Clustering method is worked well - Use k=3 and 4	- They did not validate the clustering for k= 3,4 values - Did not validate k=2 in detail
Baadel et al., 2020	CATC	Pre-processing: OMCOKE clustering Predictive: RIPPER, PART, Random Forest, Random Tree, ANN	Model using clustered dataset is better that model without clustering	- Clustering makes data sample decrease which it will be used for supervised - Do not checked whether cluster dataset is imbalanced
Stevens et al., 2019	Gaussian Mixture Models	Hierarchical Agglomerative Clustering	- 16 clusters created - The differences between clusters are	- Does not focus on evaluation of this clustering machine learning

			mean age and degrees of severity across developmental domains -Only six clusters (out of 16) with non-dominant	
Stevens et al., 2017	ABA	K-means clustering	Only one cluster (out of seven) with no single dominant challenging behaviour	- Mostly cluster have single dominant challenging which
Gardner-Hoag et al., 2021	ABA	K-means clustering and regression	- For K-means, only two clusters (out of seven) with non-dominant challenging behaviour Regression: - 67% of the variance of exemplar mastery -mean squared error between 0.21 to 0.37	- Still limited to the existing data in the data set

CHAPTER 3

METHODOLOGY

3.1. Dataset Description

The dataset was collected from 300 participants which recruited from preschools in the rural area of Gua Musang and the urban area of Kota Bahru located in the East Coast state of Kelantan in 2010. MPLAT diagnose version (full version) were used for this experiment.

Table 3.1: List of features in MPLAT dataset

Feature	Type	Description
AGEGROUP	Ordinal	Group of every six months (which start from 4 years 0 month until 6 years 11 months) of children who participate the MPLAT diagnostic. Groups are 1, 2, 3, 4, 5 and 6.
Ageinmonth	Integer	Age (in month) of children who participate the MPLAT Diagnostic.
PV	Integer	Picture Vocabulary. One of subtests in MPLAT. Participants are asked to point to the picture that represents a word that told by examiners.
SR	Integer	Sentence Repetition. One of subtests in MPLAT. Participants are asked to listen to a sentence read by the examiner and repeat the sentence exactly as they heard it
GU	Integer	Grammatical Understanding. One of subtests in MPLAT. Participants are asked to choose the best picture out of three choice pictures represent the grammatical content that told by examiners.
REFMNG	Integer	Referential Meaning. One of subtests in MPLAT.

		Participants are asked to explain the meaning of words (object, person, feature) that told by examiners.
RELMNG	Integer	Relational Meaning. One of subtests in MPLAT. Participants are asked to explain the similarity of two different words (object, person, feature).
EL	Integer	Early Literacy Skills. One of subtests in MPLAT. It evaluates copying, reading, and writing skills. Participants are tested on phoneme-grapheme correspondence when asked to relate letters with sounds and syllables, spell words and copy words and sentences
TOTAL	Integer	Total of whole test’s scores

This data doesn’t have labelled features that makes client has difficult on grouping/classify data. And this dataset cannot be used on supervised machine learning classifier.

3.2. Data Science Process: Knowledge Discovery in Databases (KDD)

For this project, the data science process is implement using KDD Framework. KDD is a process of finding meaningful knowledge and patterns from a data, and it is very useful when machine learning methods is included. (Piatetsky-Shapiro, Fayyad, Smith, & mining, 1996)

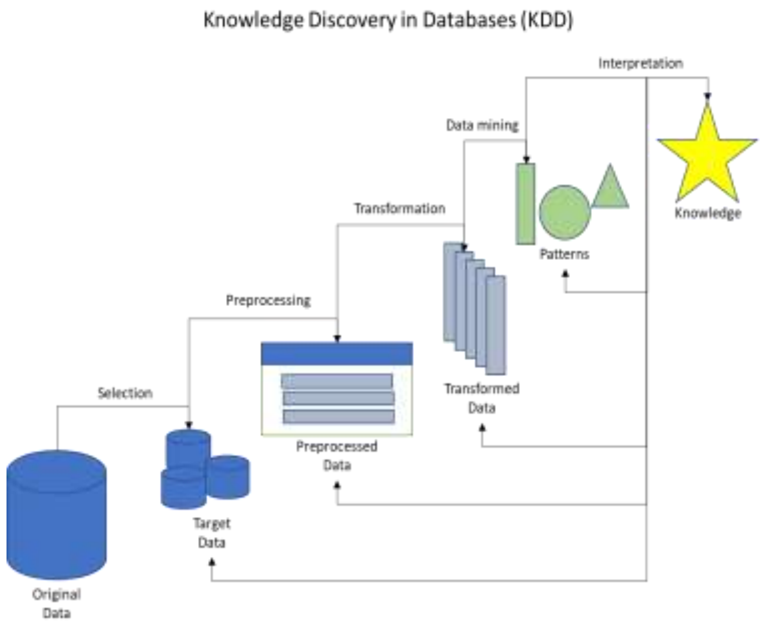


Figure 3.1. Knowledge Discovery in Databases (KDD)

There are five processes in KDD which are selection, pre-processing, transformation, data mining and interpretation. Data mining and interpretation process need to be given more attention especially is interpreting mined patterns to gain important knowledge and those interpreted knowledge will be explained to client using easy to understand term. If interpretation process gives a bad result, it can revisit to previous process for doing correction.

3.3. Research Flow

Research flow charts is a diagram that show the progress of research from the beginning of exploration to the end of a conclusive understanding. This flow charts can be said as summary of KDD process.

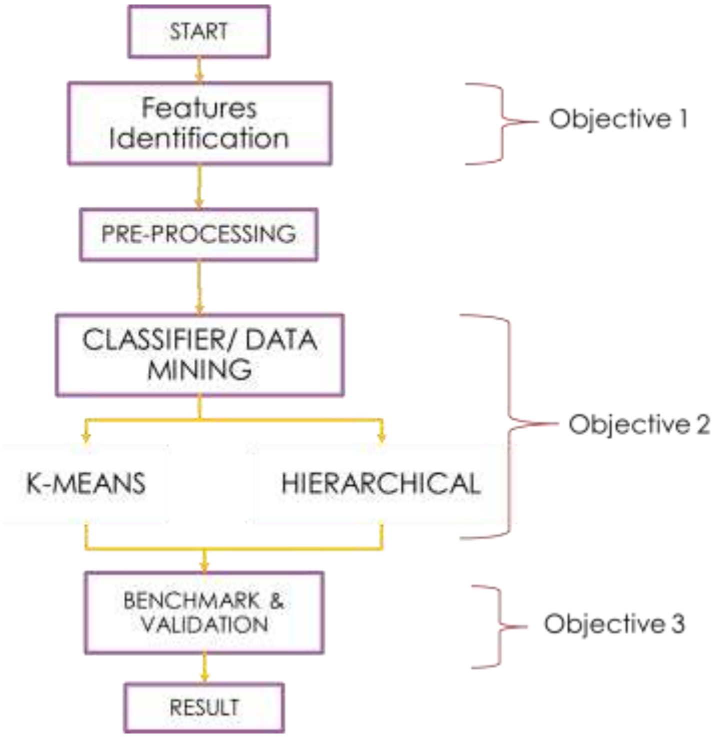


Figure 3.2. Research flow chart

Feature identification represent first objective which is to identify the features essential for speech disorder diagnosis, then pre-processing will be proceeded. The third stage is classifier which is created to reach second objective. Finally, we process benchmark and validation which it will achieve third objective.

3.3.1. Features Identification

Two criteria that need to be concerned for choosing features based on corelation and variance. The features are good to be chosen if it has low correlation with another features and has high variance. To simplify the selection, Principal Component Analysis (PCA) is used that include two those criteria.

a) Pearson correlation coefficient (PCC)

PCC is a statistical metric that calculate the strength of a correlation between two features. (Benesty, Chen, Huang, & Processing, 2008). There is formula of PCC as shown in below:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

$r = PCC$
$x_i = \ values\ of\ the\ x - variable\ in\ a\ sample$
$\bar{x} = mean\ of\ the\ values\ of\ the\ x - variable$
$y_i = \ values\ of\ the\ y - variable\ in\ a\ sample$
$\bar{y} = mean\ of\ the\ values\ of\ the\ y - variable$

A *r* value can take a range of values from 1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 represents a positive association, which is value of one variable will increases when another variable also increases. A value less than 0 represents a negative association which is if the value of one variable increases, the value of the other variable decreases. If variables with high correlation are chosen, the one-dimensional problem will be occurred during clustering, and it make difficult to analyse data due to the tightness of the clustering. To make this analysis easily, weight by correlation of features ae calculated. Weight by correlation is a weighting scheme is based upon correlation, and it returns the absolute or squared value of correlation as features weight. The higher the weight of an attribute, the more relevant it is considered.

b) Variances

Variance in simple definition is a measure that observe the spreading of data set. In mathematical term, it defined as the average of the squared differences from the mean. (Miller, 1966). Feature with high variance contains variety value in most sample while low variance may contain same value in most samples. Low variance will cause of high bias even the classifier is stable. (Breiman, 1996). Higher variance may contain more useful information, but the problem is some features with high variance may have low weighted by correlation. (Ferreira & Figueiredo, 2012). Therefore, considering both correlation and variance is important in features selection by choosing features with good value of both criteria.

c) Principal Component Analysis (PCA)

PCA is a mathematical algorithm that reduces the dimensionality of the data, and it also retains high variation and low correlation in the data set. (Ringnér, 2008). It is a common method that recently used for unsupervised machine learning. PCA identifies new the principal components, which are linear combinations of the original variables. First principal component accounts for the largest possible variance in the data set. The second principal component is calculated in the same way, with the condition that it is less correlated with the first principal component and that it accounts for the next highest variance. Next component also choose feature that has low correlation with previous component and has high variance. And the next component works with same procedure.

The number of components is determined based on cumulative variance. There are many rules to choose number of components, the most acknowledge is choose number components that has 95% of variance. After finding number of component and executing PCA is finished, the features that are important to each component are analysed based on the magnitude of the corresponding values in the eigenvectors of each component which it will be calculated.

3.3.2. Pre-processing and Transformation

After that, pre-processing process is searching for missing data and removing unnecessary matters that can reduce effectiveness process such as noisy, redundant, and low-quality data from the data. Transformation is the process that transform data into a form according to the require of data mining. Due to small sample data of MPLAT dataset, there study had suggested using progressive sampling to increase sample for increasing the accuracy and reducing bias. (Figueroa, Zeng-Treitler, Kandula, Ngo, & making, 2012)

3.3.3. Classifier

Classifier process is a most important process when it divides data into several groups and extracts useful patterns from data by using machine learning methods. For this project, unsupervised machine learning algorithms which is clustering will be used. Based on the project, there are two main goals which are to identify children with disorder and to reveal other subgroups rather than two group. So, there are two parts of mining phase:

i. using two cluster to identify speech disorder
ii. finding ideal number of clusters and analyse the relative strengths between subtest in each cluster.

Clustering is an unsupervised learning it works by cluster data into groups based on their similarity. The aim of this algorithm is to segregate groups with similar traits and assign them into cluster. There are two type of clustering that been used in this project, K Means Clustering and Hierarchical Method.

a) K Means Clustering

K Means is an iterative clustering algorithm whereby it aims is to find the local maxima in each iteration. It works by assigning the desired number of cluster K. It is a most acknowledge clustering because it less computationally intensive which it very good to be used with large datasets. Other that, it can use median or mean as a cluster centre to represent each cluster.

K-means clustering uses x as input and the number of clusters to be fit as k. The main goal of K-means is minimizing within-cluster variation, $W(C_k)$, when C_k is a cluster of k. K-means clustering algorithm calculation is shown as follows:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum\nolimits_{x}^{p} \left(x_{i\,j} - x_{i'j}\right)^2 \qquad (2)$$

Value of k is chosen by changing value k from 1 to 20 using *for* looping which suggested by Stevens E. (2019). However, using number of features as *k* value is not recommended so it can prevent overfitting. (Gardner-Hoag et al., 2021). There are some limitations of K-means. It is not only very dependent, but also dependent on the placement of the initial codebooks. (Shahapurkar & Sundareshan, 2004). If codebooks are too small, it might result in convergence to local minima and accuracy might be low.

The common challenging with this clustering is choosing and identify a good K value. The easily way to identify optimal *k* value is using the Elbow method. Basically, when the number of *k* increases, the sum of squared distances of samples to their closest cluster centre which called inertia decreases. To determine *k* value, the graph of correlation between number of k and inertia of clustering is plotted. The Elbow method works when the elbow point, which optimal *k* point is spotted when the inertia starts decrease slowly. (Syakur, Khotimah, Rochman, & Satoto, 2018). For example, the inertia from cluster *k* = 2 to *k* = 3 drastically decreases. Then, the inertia from cluster *k* = 3 to *k* = 4 also drastically decreases. But the inertia from cluster *k* = 4 to *k* = 5 starts slowly decrease and it continues to next *k* values. Therefore, *k* = 4 is the elbow point which is the optimal *k* value.

b) Hierarchical Clustering

Hierarchical clustering works by build the hierarchy of cluster. The algorithm is started to evaluate with all data point assigned to a cluster of their own. Next, the two nearest cluster are merged into the same cluster. This process is iterated until there is only a single cluster left. This clustering is one of common because it very easy to handle and parameter's adjustable, besides that, it can used even feature type in dataset is not an integer or float. Another advantage of hierarchical clustering which K-means does not has is hierarchical not need to guess how many numbers of clusters. To determine number of clusters, height of the branch points should be concerned which can be analyse Dendrogram. The less the height, the more information that can be gotten but more difficult to interpret. Results of agglomerative hierarchical clustering depend to a deal on the distance measure chosen. But distance measure chosen is very sensitive to noise such as noise and outliers in dataset. Those problem might be affected to the accuracy.

3.3.4. Validation: Clustering validation measures

According to previous studies, they evaluate the clustering algorithm by using external cluster validation. This approach is done by comparing the cluster class with the external information which is actual class. (Gupta & Panda, 2019). They can do so because they already have actual class data. For this project, internal cluster validation is used so cluster algorithm can be evaluated without external information. There is study that suggested using three internal validation measures to evaluate clustering (Gupta & Panda, 2019) (Petrovic, 2006) which are:

a) Silhouette analysis

Silhouette coefficient analysis is most common when involving internal. (Gupta & Panda, 2019). There are two ways to evaluate clustering in Silhouette coefficient. First method is calculating the mean distance amongst the clusters. The range value of silhouette coefficient value is between -1 and +1. If the value near to +1, it means the cluster is far from another clusters. When the value near to 0, all clusters close to each other while -1 explains that samples might have been assigned to the wrong cluster. Second methos is observing on Silhouette Plot. This plotting shows closeness of the points in one cluster to another clusters. The performance of the clustering model is good if the width of all clusters in Silhouette Plot are similar, and their scores are more than average silhouette scores.

b) Dunn index

Dunn index is a simple validation for clustering algorithm. It is calculated the lowest inter-cluster separation as minimum separation between two clusters and then it is divided by the highest intra-cluster distance which is maximum diameter between points in any cluster. (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). The clustering solution is evaluated as good solution when the Dunn Index value is high. (Misuraca, Spano, Balbi, & Methods, 2019). This calculation of Dunn index (D) shows that this index is direct proportion to inter-cluster separation, and it also inverse proportion to intra-cluster distance. The calculation is shown as follow:

$$D = \frac{min.separation}{max.diameter} \tag{3}$$

c) Davies Bouldin Index

The Davies-Bouldin index is based on the almost of the distances between clusters and their cluster centre to obtain a final value that represents the quality of the partition. (Karo, MaulanaAdhinugraha, & Huda, 2017). The score is calculated as the average similarity of each cluster with a cluster most like it. The clustering is performed good in term of separation if average similarity is low. This evaluation later become commonly because its computation of is simpler than that of Silhouette scores. (Halkidi, Batistakis, & Vazirgiannis, 2001). The calculation of average similarity, R_{ij} and Davies Bouldin Index are shown as below:

$$R_{ij} = \frac{s_i - s_j}{d_{ij}} \tag{4}$$

$i = \ cluster$
$j = \ cluster\ that\ most\ similar\ with\ cluster$
$s = the\ average\ distanc\ between\ each\ point\ of\ cluster\ and\ the\ centroid$
$d_{ij} = $ the distance between cluster i and j

$$DBI = \frac{1}{k}\sum_{i=j}^{k} \max R_{ij} \tag{5}$$

$k = \ the\ number\ of\ clusters$

CHAPTER 4

RESULTS

4.1. Feature Selection

Firstly, the client wants to separate data based on subtests which are PV, SR, GU, REFMNG, RELMNG, and EL. The AGEGROUP and Ageinmonth are ignored for this clustering because the marking standard or methods of all subtests were different based on age group. For example, 50 score of RELMNG of participant from group 3 are different from 50 score of RELMNG of participant from group 1. TOTAL is also dropped because it has high dependence to all subtests.

We use Principal Component Analysis (PCA) to analyse remaining features base on correlation and variance. The graph of correlation between number of component and cumulative variance is shown in below:

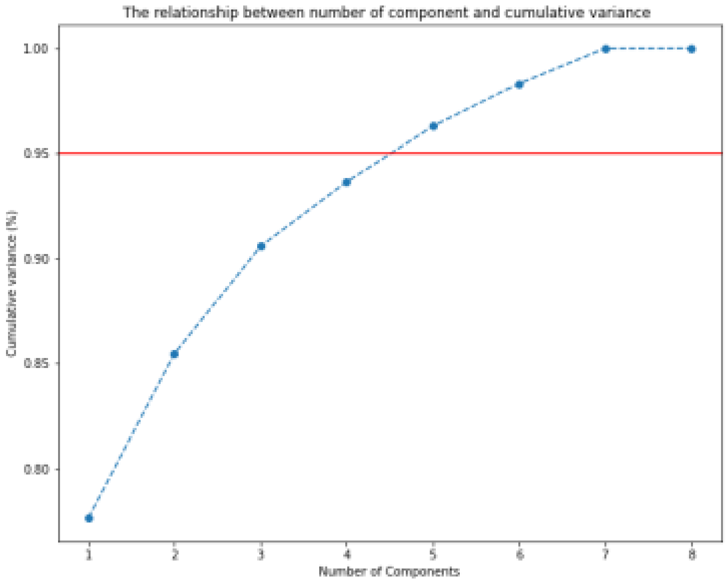


Figure 4.1: The correlation between number of component and cumulative variance

There are many rules to choose number of components, the most acknowledge is choose number components that has 95% of variance. Based on graph, we can see that four is an optimal number of components. To analyse which features that are important to each component, the magnitude of the corresponding values in the eigenvectors of each component are calculated.

Table 4.1: Importance features in principle components

Component	PV	SR	GU	REFMNG	RELMNG	EL
1	0.2935	0.2715	0.145	0.3243	0.6446	0.5464
2	0.2344	0.0707	0.1866	0.1067	0.6809	0.6559
3	0.5223	0.2937	0.0267	0.6001	0.2681	0.4377
4	0.491	0.0142	0.5734	0.5766	0.1593	0.2686

The summary of calculation is showed in Table 4.2:

Table 4.2: Summary of importance features in principle components

Component	Importance Features
1	RELMNG
2	RELMNG and EL
3	REFMNG
4	REFMNG and GU

For this project, we choose RELMNG, EL, REFMNG and GU features to develop clustering. For further research, weight by correlation and variances of all selected features are calculated as show in Table 4.3 and Table 4.4

Table 4.3: Weight by correlation of selected features

Feature	Weighted By Correlation
GU	1.0
REFMNG	0.539
RELMNG	0.455
EL	0.437

Table 4.4: Variances of selected features

Features	Variances
RELMNG	104.943266
EL	81.527313
REFMNG	35.382598
GU	14.255017

4.2. Sampling

Due to size of dataset is small, we plan to use random sampling method to increase dataset's size. However, we need to determine how much to sample this dataset. So, we use for looping with range of (300,900) to apply K-means and calculate the silhouette score of each size. The result is shown on below:

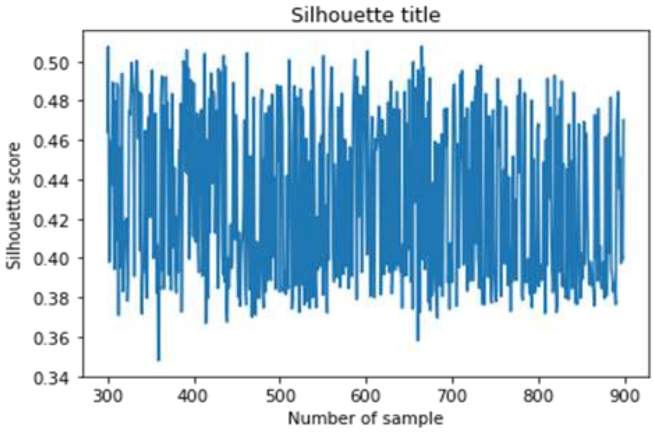


Figure 4.2: The correlation between number of sample and silhouette score

Based on the graph, we can identify that dataset with 665 samples is the highest silhouette score. Therefore, we choose 665 as sampling size and we apply random sampling method to increase dataset from 300 to 665.

4.3. Clustering

We use two common clustering which is K-means and Hierarchical Clustering. In this section, clustering for all correlation between two features are developed. There are two information that will be analysed which are separation pattern, amount of data for each cluster and the clustering rules. (Waisakurnia, 2020). The clustering rule is a way to describe each cluster. The rules are obtained by using supervised machine learning algorithm which is decision tree. This algorithm is used because it can extract the information which is feature importance and its condition. This rule is operated by using features as input and labelled clusters as output.

4.3.1. Number of clusters

For this project, we analyse on two values of clusters, *k* which are 2 and optimal number clusters. For finding optimal number of clusters in K-means, number of clusters is determined by using Elbow method. Basically, when the number of *k* increases, the sum of squared distances of samples to their closest cluster centre which called inertia decreases. To determine *k* value, the graph of correlation between number of k and inertia of clustering is plotted. The Elbow method works when the elbow point, which optimal k point is spotted when the inertia starts decrease slowly. Based on the graph, we determine that four is the Elbow point.

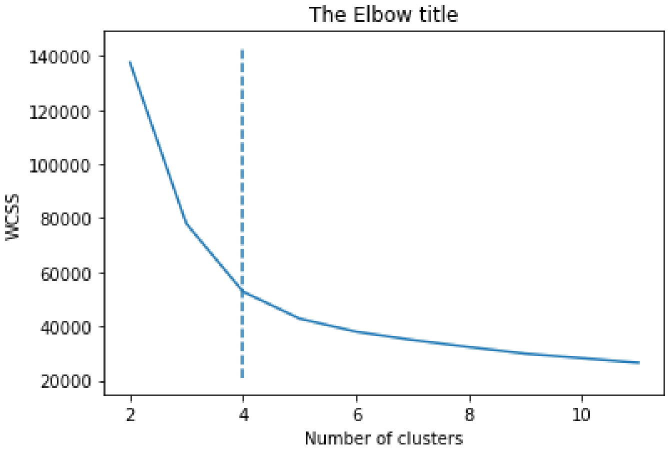


Figure 4.3: The elbow graph

For hierarchical clustering, to determine number of clusters, height of the branch points should be concerned which can be analyse Dendrogram. The less the height, the more information that can be gotten but more difficult to interpret. We choose based on our observation on the graph which not same with K-mean's Elbow method.

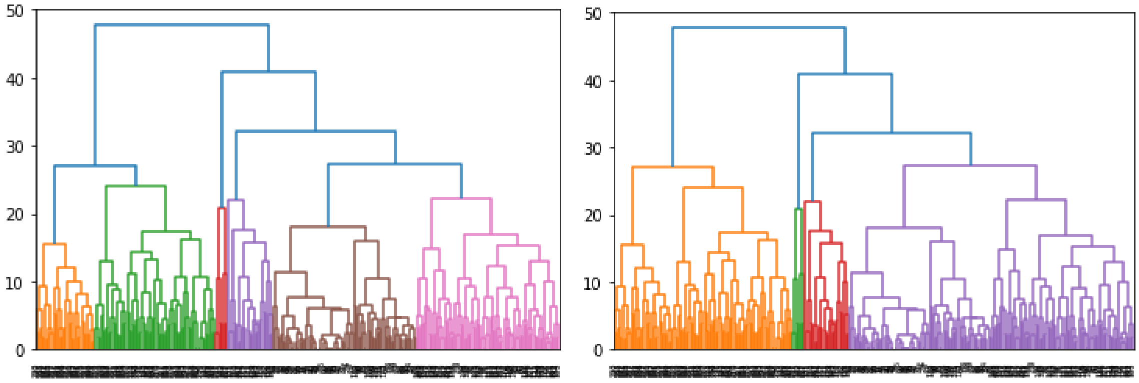


Figure 4.4: Dendrogram with height 25 (left) and 30 (right)

Based on this graph, we choose 25 and 30 of height. We found that if set height in 25 it shows six cluster and if we set height in 30, it shows four cluster. Six cluster shows better due to the height of each six tree are similar. However, if want to compare K-means and Hierarchical we choose four as number clusters.

4.3.2. Clustering with all selected features

The Table 4.3 shows the result of clustering for all selection features, and it also shows the features rule or feature range of each cluster.

Table 4.5: Result of K-means clustering for all selection features

Cluster	Count	Rules
0	192	• RELMNG <= 13.5 • REFMNG > 6.5
1	168	• RELMNG > 13.5 • REFMNG > 6.5 • EL <= 18.5
2	204	• RELMNG <= 13.5 • REFMNG <= 6.5
3	101	• RELMNG > 13.5 • REFMNG > 6.5 • EL > 18.5

Table 4.6: Result of hierarchical clustering for all selection features

Cluster when k=4	Count	Rules
0	210	• REFMNG > 17.5 • EL > 8.5
1	329	• REFMNG <= 8.5 • EL <= 6.5
2	11	• REFMNG <= 8.5 • EL > 17.5
3	115	• REFMNG > 8.5 and REFMNG <= 17.5

4.3.3. Number of features:

Based on some finding, if number of features that used on clustering is small, the clustering performance is better. Firstly, we use all selection features. The evaluation result of clustering all selected features and with four cluster is shown in below:

Table 4.7: Evaluation Score of Clustering for all selection features

Clustering	Silhouette Coefficient (higher is better)	Dunn Index (higher is better)	Davies Bouldin (lower is better)
K-means	0.4161	0.3509	0.8591
Hierarchical	0.3465	0.4201	0.9541

After that, we test clustering with two and three features in K-means clustering. The graph of K-means clustering performance for all, three and two features is shown in below:

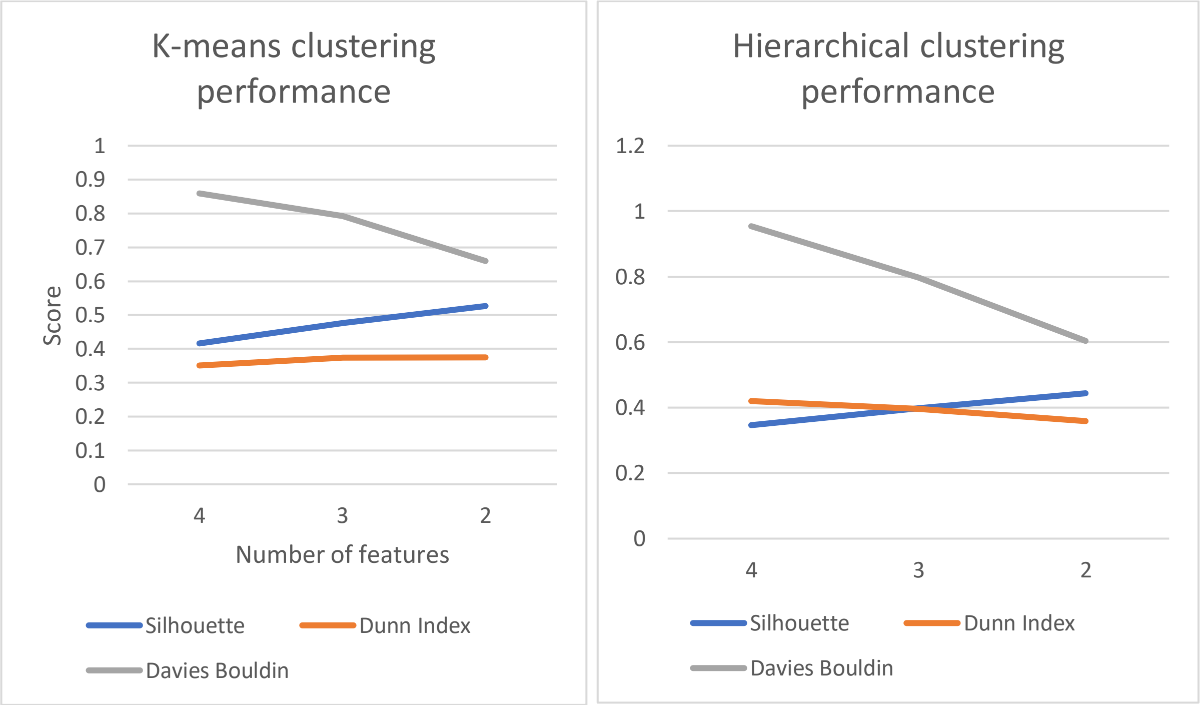


Figure 4.5: K-means and Hierarchical clustering performance

The graph show two features gives more performance of clustering. Therefore, we choose two features, and we test all two combination features.

4.3.4. Clustering with two features

This section also shows the result of clustering for *k*=2 and *k*=4 as shown in Figure 4.6.

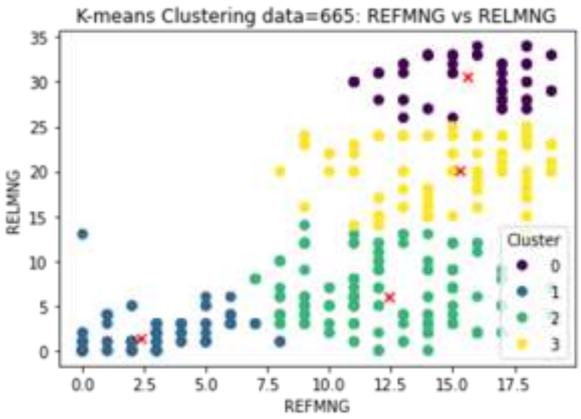


Figure 4.6: the scatter plot of K-means REFMNG and RELMNG

a) K-means Clustering:

The Table 4.8 show the result of clustering for all combination features, and it also show the features rule or feature range of each cluster.

Table 4.8: Result of K-means clustering for all correlation two features

Features	Cluster when k=4	Count	Rule
GU and REFMNG	0	186	• REFMNG > 5.5 and REFMNG <= 13.5 • GU > 9.5
	1	164	• REFMNG <= 7.5
	2	254	• REFMNG > 13.5 • GU > 9.5
	3	61	• REFMNG > 13.5 and GU <= 9.5
GU and RELMNG	0	155	• RELMNG > 16.5 and RELMNG <= 24.5
	1	316	• RELMNG <= 6.5
	2	75	• RELMNG > 24.5
	3	119	• RELMNG > 6.5 and RELMNG <= 16.5
GU and EL	0	162	• EL > 5.5 and EL <= 12.5
	1	99	• EL > 21.5
	2	303	• EL <= 5.5
	3	101	• EL > 12.5 and EL <= 21.5

REFMNG and RELMNG	0	75	• RELMNG > 25.5
	1	187	• REFMNG <= 6.5
	2	208	• REFMNG > 6.5 • RELMNG <= 13.5
	3	195	• REFMNG > 6.5 RELMNG > 13.5 and RELMNG <= 25.5
REFMNG and EL	0	142	• REFMNG <= 6.5
	1	182	• REFMNG <= 6.5
	2	233	• REFMNG > 6.5 • EL <= 9.5
	3	108	• EL > 20.5
RELMNG and EL	0	157	• EL > 4.5 • RELMNG <= 16.5
	1	118	• EL > 17.5 • RELMNG > 16.5
	2	118	• EL <= 4.5 • RELMNG <= 5.5
	3	272	• EL <= 4.5 • RELMNG <= 5.5

a) Hierarchical Clustering:

The Table 4.9 show the result of clustering for all combination features, and it also show the features rule or feature range of each cluster.

Table 4.9: Result of hierarchical clustering for all correlation two features

Features	Cluster when k=4	Count	Rule
GU and REFMNG	0	292	• REFMNG > 8.5 and REFMNG <= 14.5 • GU > 12.5
	1	174	• REFMNG <= 8.5 • GU > 8.5
	2	164	• REFMNG > 8.5 and REFMNG <= 14.5 • GU <= 12.5
	3	35	• REFMNG > 8.5 • GU <= 8.5

GU and RELMNG	0	401	• RELMNG <= 14.5
	1	175	• RELMNG > 14.5 and RELMNG <= 27.5 • GU > 10.5
	2	27	• RELMNG > 14.5 • GU < =10.5
	3	62	• RELMNG > 27.5
GU and EL	0	230	• EL > 6.5 and EL <= 20.5
	1	108	• EL > 20.5
	2	71	• EL <= 6.5 • GU <= 6.5
	3	256	• EL <= 6.5 • GU > 6.5
REFMNG and RELMNG	0	196	• REFMNG <= 6.5
	1	159	• RELMNG > 11.5 and RELMNG <= 22.5
	2	134	• REFMNG > 6.5 • RELMNG > 22.5
	3	176	• REFMNG > 6.5 • RELMNG <= 11.5
REFMNG and EL	0	205	• EL > 12.5 • REFMNG > 7.5
	1	191	• EL <= 12.5 • REFMNG <= 7.5
	2	5	• EL > 12.5 • REFMNG <= 7.5
	3	264	• EL <= 12.5 • REFMNG > 7.5
RELMNG and EL	0	144	• RELMNG > 20.5 • EL > 11.5
	1	148	• RELMNG > 10.5 • EL <= 11.5
	2	16	• RELMNG <= 10.5 • EL > 11.5
	3	357	• RELMNG <= 10.5 • EL <= 11.5

4.4. Evaluation

For evaluation, there are four comparison that be presented which are comparison between features, comparison between original and sampling dataset and comparison between K-means and hierarchical clustering. And for comparison features and clustering, sampling dataset are used.

4.4.1. Comparison between correlation of two features

For this section, we compare all features correlation in K-means and hierarchical clustering:

a) K-means clustering

Table 4.10: Evaluation score for K-means Clustering

Features	Silhouette	Dunn Index	Davies Bouldin Score
GU and REFMNG	0.4569	0.3071	0.8362
GU and RELMNG	0.4762	0.2926	0.7782
GU and EL	0.4619	0.2863	0.7528
REFMNG and RELMNG	0.5267*	0.3363*	0.6259**
REFMNG and EL	0.5013	0.2509	0.7162
RELMNG and EL	0.4766	0.1531	0. 7794

* Highest score for Silhouette and Dunn
**Lowest score for Davies Bouldin.

But as mentioned in Chapter 3, evaluate through silhouette score is just basic. The overall evaluation is analysing on Silhouette plot. This plotting shows closeness of the points in one cluster to another clusters. The performance of the clustering model is good if the width of all clusters in Silhouette Plot are similar, and their scores are more than average silhouette scores. In this plot, we want to observe which correlation between two features that perform well:

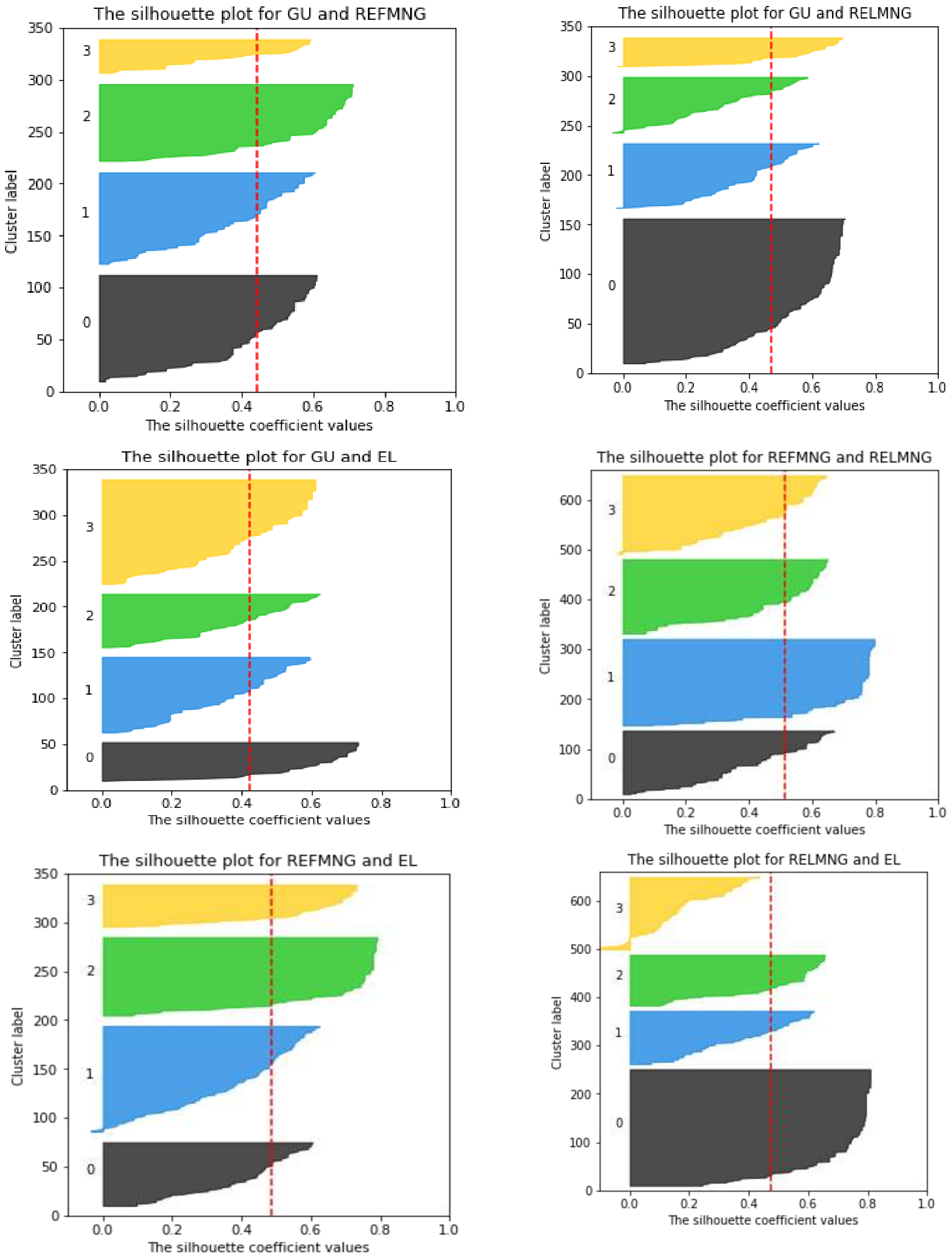


Figure 4.7: Silhouette plots for all features in K-means clustering

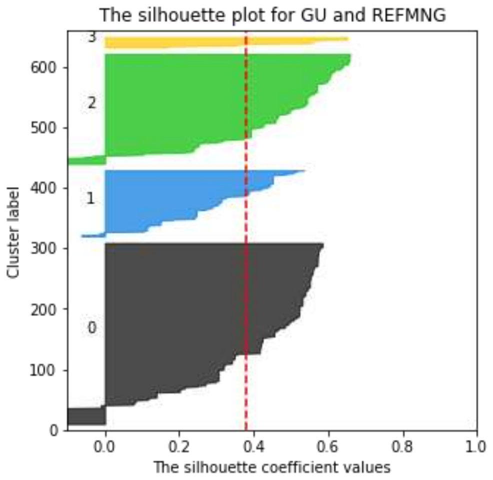
b) Hierarchical Clustering

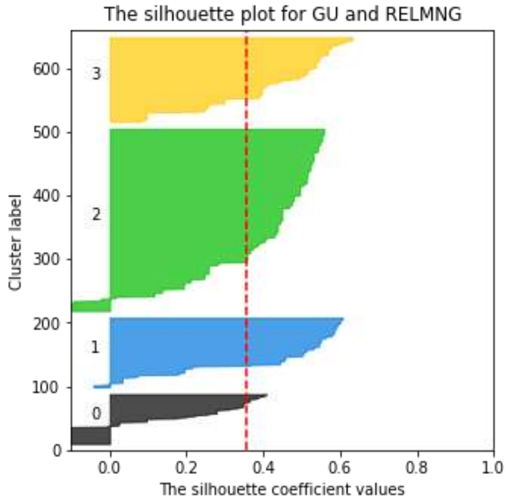
Table 4.11: Evaluation score for Hierarchical Clustering

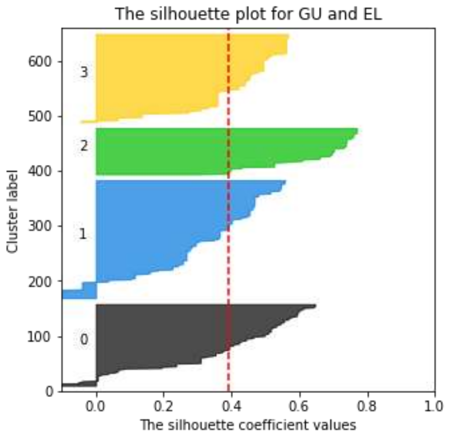
Features	Silhouette	Dunn Index	Davies Bouldin Score
GU and REFMNG	0.3811	0.2536	1.0988
GU and RELMNG	0.3545	0.2575	0.7987
GU and EL	0.3910	0.3276	0.7151
REFMNG and RELMNG	0.4807*	0.3363	0.7060
REFMNG and EL	0.4439	0.3589*	0.6039**
RELMNG and EL	0.4583	0.2936	0.7329

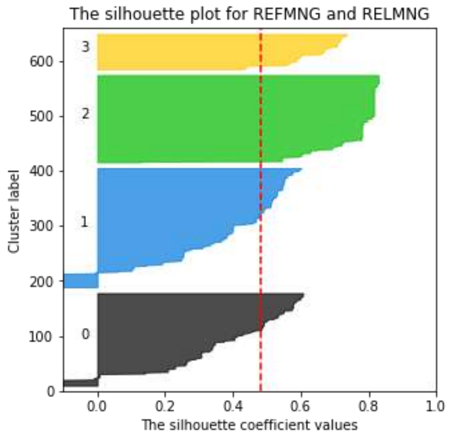
* Highest score for Silhouette and Dunn
**Lowest score for Davies Bouldin.

Then, we evaluate overall performance by analysing on Silhouette plot. In this plot, we want to observe which correlation between two features that perform well. The silhouette plot is shown in figure below:









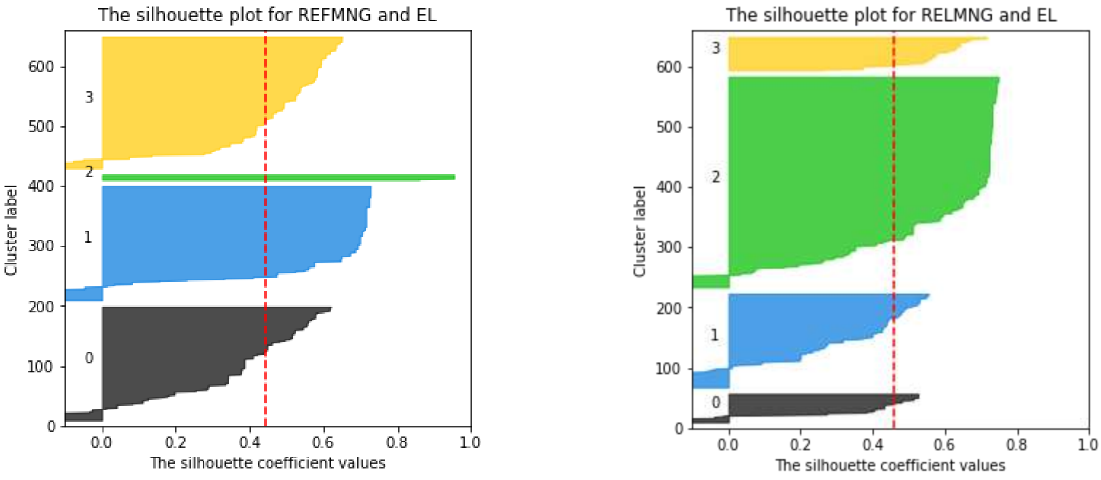


Figure 4.8: Silhouette plots for all features in Hierarchical clustering

4.4.3. Comparison Original Sample Size and Sampling Sample size

a) K-means clustering

There is evaluation performance of K-means clustering by using Silhouette Coefficient, Dunn Index and Davies Boldin Index, and we compare between original dataset and sampling dataset:

Table 4.12: Silhouette Score of K-means Clustering for Sampling and Non-sampling

Features	Silhouette Score	Sampling Silhouette Score
REFMNG and RELMNG	0.5016	0.5267*
REFMNG and EL	0.4875	0.5013
RELMNG and EL	0.5193*	0.4766
GU and RELMNG	0.4676	0.4762
GU and EL	0.4210	0.4619
GU and REFMNG	0.4410	0.4569

* The highest score for Silhouette

Table 4.13: Dunn Index Score of K-means Clustering for Sampling and Non-sampling

Features	Dunn Index	Sampling Dunn Index
REFMNG and RELMNG	0.3467*	0.3363*
GU and REFMNG	0.3071	0.3571
GU and RELMNG	0.2661	0.2926
GU and EL	0.2863	0.2863
REFMNG and EL	0.2509	0.2509
RELMNG and EL	0.1715	0.1531

*The highest score for Silhouette

Table 4.14: Davies Bouldin Index of K-means Clustering for Sampling and Non-sampling

Features	Davies Bouldin	Sampling Davies Bouldin
REFMNG and RELMNG	0.6916*	0.6259*
REFMNG and EL	0.7169	0.7162
GU and EL	0.7977	0.7528
GU and RELMNG	0.7838	0.7782
GU and REFMNG	0.7942	0.7793
RELMNG and EL	0.8799	0.8362

*The lowest score for Davies Bouldin.

To make analysing easily, the comparison between original dataset and sampling dataset for K-means clustering are plotted in Figure 4.9. Based on the graph, there is no significant difference between original dataset and sampling dataset.

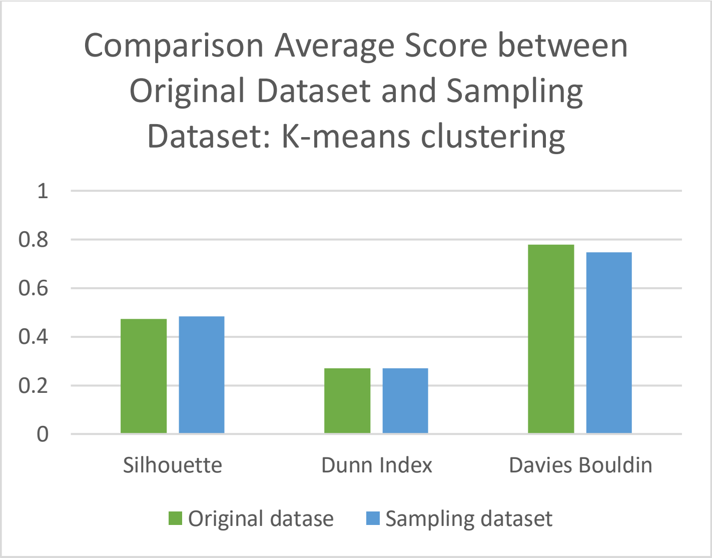


Figure 4.9: Comparison Average Score between Original Dataset and Sampling Dataset: K-means

i) Hierarchical Clustering

There is evaluation performance of hierarchical clustering by using Silhouette Coefficient, Dunn Index and Davies Boldin Index, and we compare between original dataset and sampling dataset:

Table 4.15: Silhouette Score of Hierarchical Clustering for Sampling and Non-sampling

Features	Silhouette Score	Sampling Silhouette Score
REFMNG and RELMNG	0.4685	0.4807*
RELMNG and EL	0.5071*	0.4583
REFMNG and EL	0.4903	0.4439
GU and EL	0.2988	0.3910
GU and REFMNG	0.2401	0.3811
GU and RELMNG	0.3992	0.3545

* The highest score for Silhouette

Table 4.16: Dunn Index Score of Hierarchical Clustering for Sampling and Non-sampling

Features	Dunn Index	Sampling Dunn Index
REFMNG and EL	0.3363	0.3536*
REFMNG and RELMNG	0.3386*	0.3363
GU and EL	0.2906	0.3276

RELMNG and EL	0.2936	0.2936
GU and REFMNG	0.2425	0.2589
GU and RELMNG	0.2774	0.2575

*The highest score for Silhouette

Table 4.17: Davies Bouldin Index of Hierarchical Clustering for Sampling and Non-sampling

Features	Davies Bouldin	Sampling Davies Bouldin
REFMNG and EL	0.5246*	0.5039*
REFMNG and RELMNG	0.7119	0.7060
GU and EL	1.1710	0.7151
RELMNG and EL	0.7618	0.7329
GU and RELMNG	0.9218	0.7987
GU and REFMNG	1.0296	1.0988

*The lowest score for Davies Bouldin.

To make analysing easily, the comparison between original dataset and sampling dataset for hierarchical clustering based on average scores are plotted in Figure 4.10.

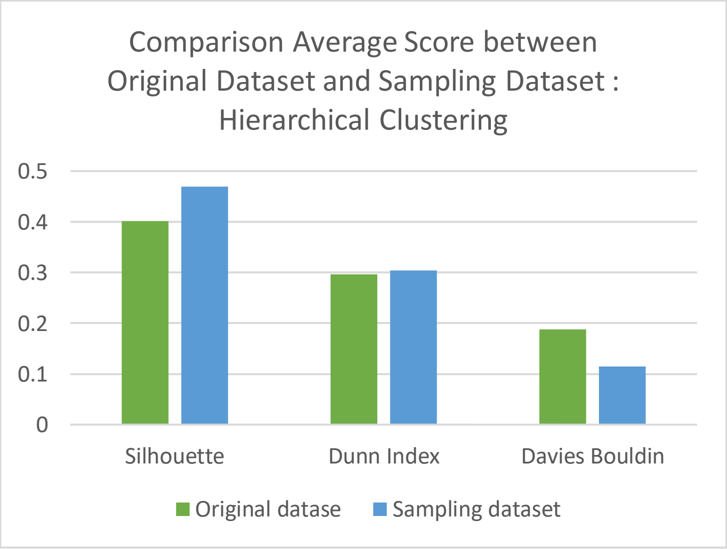


Figure 4.10: Comparison Average Score between Original Dataset and Sampling Dataset

4.4.4. Comparison K-means and Hierarchical Clustering

Score between K-means and Hierarchical Clustering are compared based on three evaluation metrics as show in Figure 4.11, Figure 4.12, and Figure 4.13:

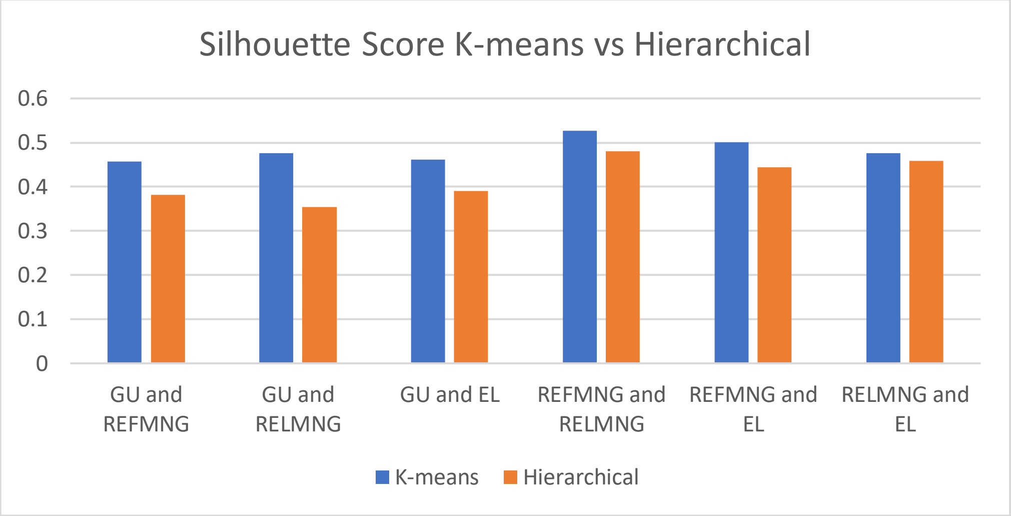


Figure 4.11.: Comparison K-means and Hierarchical (Silhouette)

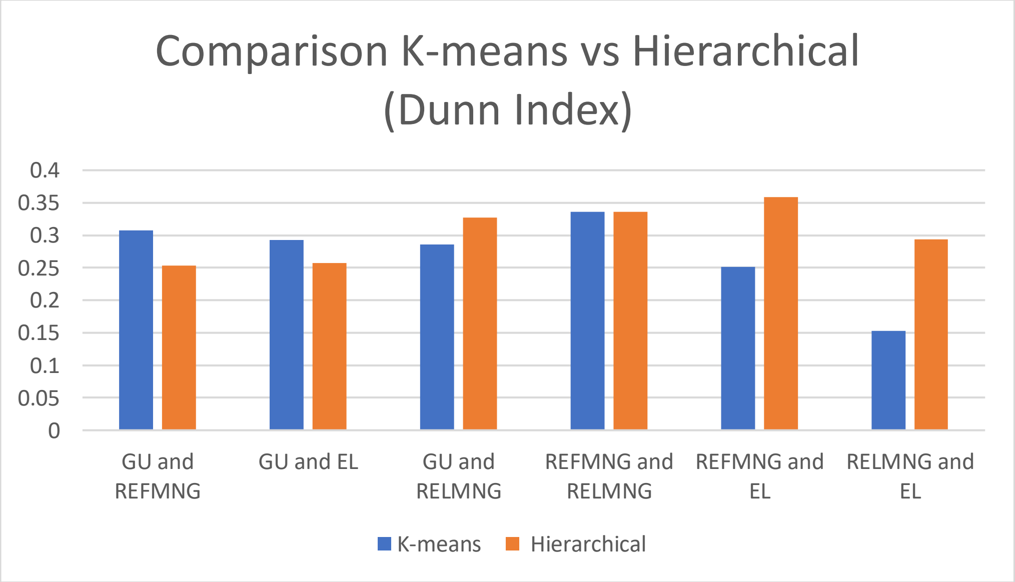


Figure 4.12: Comparison K-means and Hierarchical (Dunn Index)

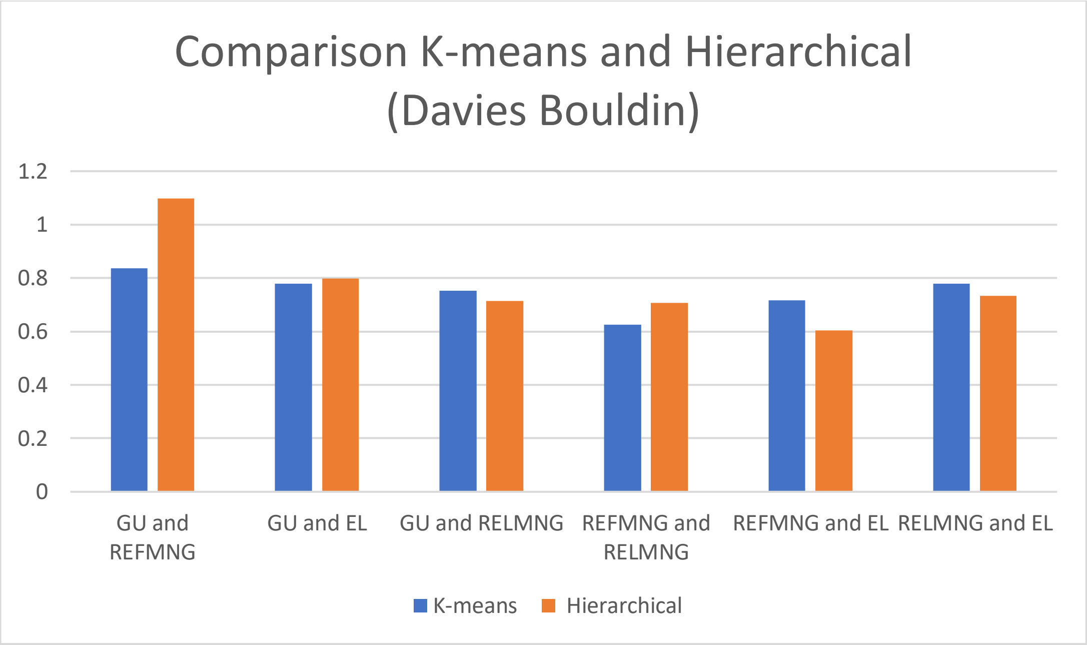


Figure 4.13: Comparison K-means and Hierarchical (Davies Bouldin Index)

The comparison between original dataset and sampling dataset for hierarchical clustering based on average scores are plotted in Figure 4.14 for make analysing easily.

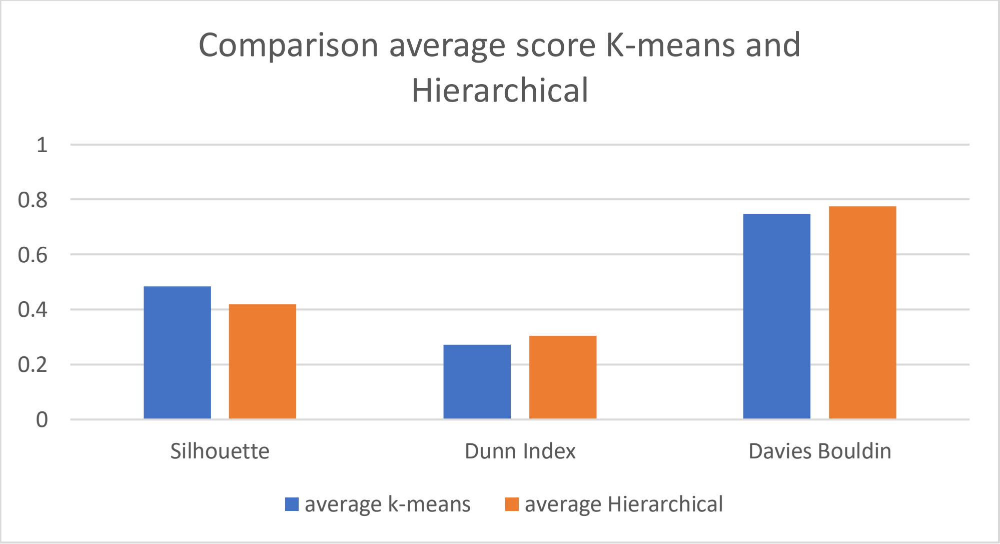


Figure 4.14: Comparison average evaluation score K-means and Hierarchical

4.4.5. Comparison with Others Paper

For justified this model is the best rather others, we compare with previous research paper. However, as I mention in Chapter 2 some studies did not focus on evaluation metric to measure the performance of clustering methods. Their authors just suggest which clustering methods are good based on the way those methods extract information.

Those research papers also discuss about which and how many features that dominant or affect to each cluster. Based on Table 4.5, two features that effect all clusters in K-means clustering which are REFMNG and RELMNG. Then based on Table 4.6, all clusters have two features that affect in all clusters in hierarchical clustering which are REFMNG and EL.

Table 4.18: Comparison Previous Research Paper

Author	Methods used	Result
Stevens et al., 2017	• K-means clustering	Only one (out of seven) cluster with non-single feature dominant
Stevens et al., 2019	• Hierarchical clustering	Only six (out of 16) cluster with non-single feature dominant
Gardner-Hoag et al., 2021	• K-means clustering	Only two clusters (out of seven) with non-single feature dominant

CHAPTER 5

DISCUSSION

5.1. Techniques

5.1.1 Sampling method

There is no significant difference between original and sampling dataset based on K-means clustering. But in hierarchical clustering, sampling dataset shows increasing of performance based on silhouette score and Davies Bouldin index. It is due of the duplicated sample which cannot change the variances of sample data when apply sampling methods. After applying sampling methods, RELMNG and EL still the only features that have high variances. If we want to overcome this, we need to find new data so the variances can increase.

5.1.2 Clustering methods

Based on the result, K-means is slightly better than hierarchical based on silhouette coefficient. Even silhouette scores of both clustering methods are slightly similar. But if we observe on silhouette plot in Figure 4.7 and Figure 4.8, it shows that K-means performance is very well and better than hierarchical clustering and hierarchical clustering performance are very poor.

In term of Dunn index and Davies Bouldin index, hierarchical clustering is slightly better than K-means clustering. Even so, as mentioned before, the density is poor. Based on silhouette plot for hierarchical clustering in Figure 4.8, all correlations show the performance are poorer than K-means. Even all clusters in all correlations are above of average score but the width of all clusters in all correlations. REFMNG and RELMNG is good enough if compare with another correlation features in hierarchical clustering. Even so, this correlation still not good enough in term of width of graph bar in silhouette plot. Then, Figure A.2 shows correlations (except REFMNG and RELMNG) have outlier clusters which it shows this clustering is poor in term of balance density. But we cannot conclude this hierarchical is not good due to lack of sample.

Overall, the whole performance of those clustering methods is good but not the best. It can be called best clustering if silhouette coefficient is near to 1 and Davies Bouldin index is near to 0. Based of this project, K-means is better than hierarchical. But the performance is affected by the structure of MPLAT dataset. As mentioned before, lack of dataset and some subtext features that have low variances can affect the performance of clustering. Besides than increasing dataset, we also can find another clustering methods that suitable for this dataset and can overcome low variances dataset.

5.2. Features

REFMNG and RELMNG is the good features for K-means clustering and this correlation is the best based on all evaluation methods. Based on silhouette plot for K-means clustering in Figure 11, all clusters in all correlations are above of score in term of average score. But in term of width, the correlation between REFMNG and RELMNG shows most of clusters are very similar rather than other correlations. the correlation between REFMNG and RELMNG shows perform well.

For hierarchical clustering, REFMNG and RELMNG is the good enough based on silhouette plot. In term of Dunn Index and Davies Bouldin Score, REFMNG and EL is most perform well rather than REFMNG and RELMNG. Even so if observe on scatter plot on Figure 5.1, the density of this correlation is not good because of cluster 2 is just an outlier.

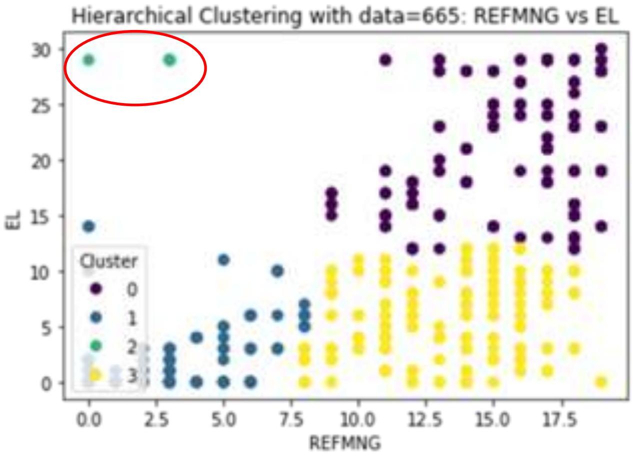


Figure 5.1: Scatter plot of hierarchical clustering of REFMNG and EL

If we observe on Table 4.6, REFMNG and EL affect the hierarchical clustering when all selected features are used. Even REFMNG and RELMNG is not the highest score Dunn Index and the lowest Davies Bouldin but at least it is just behind the highest Dunn Index and the lowest Davies Bouldin Index. The finding of both clustering shows a difference result. The possible factors that affect the result are:

i. RELMNG (104.943) and EL (81.527) have high variance values.

ii. For RELMNG, correlation with REFMNG is the highest (0.693).

iii. Correlation coefficient of EL and REFMNG (0.681) is lower than RELMNG and REFMNG.

iv. Even GU has low correlation with another selected features, its performance is lower than others (except GU and EL) due to lack of variance value (14.255).

v. GU and REFMNG is the lowest of performance even its correlation is low (0.356) due to REFMNG also has low variance (35.382).

We can conclude that variance and affect the performance of both clustering methods and outlier also affect the performance of hierarchical clustering. The way to improve is huts remove outlier or do statistical method to overcome outlier. Even so, the MPLAT data is too small, it will affect if outlier is removed. So, we need to add new data and make MPLAT data and if there still outlier, we can remove without affect the performance.

5.3. Industrial Training's Challenges and Experiences

5.3.1. Challenges and solution

The challenge that was faced during internship is in data preparation. Data with small samples make become harder in clustering. This problem may be solved by adding more data which contains features that same with features in MPLAT dataset or using sampling method. Due to data that similar with MPLAT have not found, we decide to apply sampling method even the result shows slightly increasing of clustering performance.

The next challenge that was faced during internship is we do not familiar with this project domain which is speech disorder. To overcome this challenge, I need to have more discussion with other founder who is speech science specialist. Even so, there is another challenge that I have faced which is I only contact with my main client. If I want obtain data from his employee, I need get through my client which it takes a lot of time.

Then, the research paper that related with applying unsupervised machine learning in speech disorder. Most of paper related to speech disorder used supervised machine learning. As mentioned in Chapter 2, I referred research paper that related to unsupervised machine learning in ASD.

Lastly, lack of transparency is the challenge that faced by any employees especially interim or practical student which is me. Transparency for this project is an openness between client and interim student. It is already a habit if students are afraid to voice something to the mentors or clients due to several factors such as still not being close to them, anxiety, lack of confidence in self-sufficiency, fear of being scolded by them and more. If this challenge does not solve immediately, students will face resistances such as time, confusion, moralizing and more. There is an advice that can face this problem which is creating trust by being authentic with the client. Creating good collaboration with client is needed so that we can get to know them more closely and they too can get to know us, and it can create openness. To do so, I need to know the skills or expertise I have and need to know my limitation. If we do not know our capacity, the clients do not know what assignments they can give to us. Or if we overclaim our expertise, they have high expectation to us, and it will become problem if project not going well. Then, if we face some problems with project, just inform it to client immediately so they and I can have discussion about those problems.

5.3.2. Industrial Experiences

I can say that applying machine learning during practicum are totally difference with class. During class, the lecturers gave a fundamental concept of machine learning and data that given during programming laboratory section are very good in term of structure. And then, the student did not face constraints during explaining the project flow to lecturer due to some lectures are well with data science. But during practicum, we cannot expect to retrieve systematic data like the data used during the class and do not expect to have clients who know about data sciences. That challenge can help me to understand what the true situation. Those constraints help me to Besides that, practicum also teaches me how to apply machine learning in difference domain. During practicum, I had read research paper related to speech disorder and I learn a lot of this domain and I also learn about what method that suitable for this domain, why this method chosen and how to apply a method for this domain.

However, no one can avoid from facing professional and operational issues during practicum. Tele practice issues is the most common issue right now due to Covid -19 pandemic. To prevent this infection from getting worse at client's workplace, they just told me to do a practicum at home. Practicum at home is better in terms of flexibility and comfort, but the main weakness of this the communication gap between I and the client. The client also a university lecturer in Kuala Lumpur which he may face a busyness of time. We can only have a discussion during weekend. All discussions are made through online, which make the client feel uncomforting due to setup and internet connection problem. After that, the other issue which more problem than tele practice issue is socially constructed knowledge issue. Socially constructed knowledge can be defined as a concept when people develop knowledge by sharing assumptions and knowledges with each other. (Andrews, 2012). This process is quite difficult because the client and I are different fields. I struggle to explain about machine learning to the clients and they also struggle to explain about speech disorder.

5.4. Lesson Learn

During this project, many things that I learned for how the data science process work in organization which their domain are difference with my background field and develop unsupervised machine learning. I was unfamiliar with speech disorder domain because I am from different field. When project was started, I studied this domain by reading a lot of speech disorder paper and asked clients who specializes in this domain.

Conclusion, this industrial training teaches me how to solve the problem with applied the data science lifecycle and present in layman terms. Besides, this industrial training gives me a chance to show what I was learned from this course Data Science and Analytics.

CHAPTER 6

CONCLUSION

6.1. Conclusion

To simplify the conclusion by review the contribution on project’s objectives, Table 6.1 are shown in below:

Table 6.1: Contribution on objectives

Objective	Contribution
To identify the features essential for speech disorder diagnosis	• GU, REFMNG, RELMNG and EL are selected base on PCA • RELMNG and EL have high variance • GU is lowest correlate with other features
To determine the pattern of subgroup based on the dataset by using clustering algorithms	• Optimal number of clusters is found which is four clusters RELMNG and REFMNG affect K-means clustering • EL and REFMNG affect hierarchical clustering
To test and evaluate several clustering algorithms.	• Average silhouette coefficient of clustering K-means clustering is better than hierarchical clustering

	• Average Dunn index of hierarchical clustering is better than K-means clustering • Average Davies Bouldin Index of K-means is slightly better than hierarchical clustering. • Overall, K-means is suitable for this MPLAT dataset. • K-means is good if focus on balance density of clusters • Hierarchical is good if focus on separation. But still not recommend for small dataset

Overall, clustering should be considered in future studies of speech disorder classification. If clients want to create screening system, they should include the labelled features which assigned by clinicians so it can reduce error and increase accuracy. However, even machine learning algorithms perform very good, it still limits on screening phase. For fully diagnosis, it should be made by clinicians with special training in this area. (F. H. Duffy & Als, 2019)

6.2. Study limitation

Even sampling is used, it does not show rapidly improvement due to original dataset are too small. If number of samplings is set at large number, it will cause of too many duplicated and may cause under fitting or over fitting algorithm. This limitation also will affect the performance of clustering especially hierarchical.

Another limitation is features that have low variances. But those variance issues cannot be solved immediately because it is related to speech disorder field. There are some reasons why some features only have low variances.

6.3. Future Work

If we want to improve this machine learning in future, we need to collect more data to increase performance and accuracy. Next, we will request to client if they can put a labelled features if they can so we can develop supervised machine learning. Then, we will try to develop new machine learning that use another algorithm beside than K-means and hierarchical clustering so clustering performance become better.

REFERENCES

Andrews, T. J. G. t. r. (2012). What is social constructionism? , *11*(1).

Baadel, S., Thabtah, F., Lu, J. J. I. f. H., & Care, S. (2020). A clustering approach for autistic trait classification. *45*(3), 309-326.

Benesty, J., Chen, J., Huang, Y. J. I. T. o. A., Speech,, & Processing, L. (2008). On the importance of the Pearson correlation coefficient in noise reduction. *16*(4), 757-765.

Bóna, J. J. C. l., & phonetics. (2019). Clustering of disfluencies in typical, fast and cluttered speech. *33*(5), 393-405.

Borthakur, D., Dubey, H., Constant, N., Mahler, L., & Mankodiya, K. (2017). *Smart fog: Fog computing framework for unsupervised clustering analytics in wearable internet of things.* Paper presented at the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP).

Breiman, L. (1996). *Bias, variance, and arcing classifiers*. Retrieved from

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). - NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. - *61*(- 6).

Chuchuca-Méndez, F., Robles-Bykbaev, V., Vanegas-Peralta, P., Lucero-Saldaña, J., López-Nores, M., & Pazos-Arias, J. (2016). *An educative environment based on ontologies and e-learning for training on design of speech-language therapy plans for children with disabilities and communication disorders.* Paper presented at the IEEE CACIDI 2016-IEEE Conference on Computer Sciences.

Dodd, B. J. T. i. L. D. (2011). Differentiating speech delay from disorder: Does it matter? , *31*(2), 96-111.

Duffy, F. H., & Als, H. J. B. n. (2019). Autism, spectrum or clusters? An EEG coherence study. *19*(1), 1-13.

Duffy, J. R. (2000). Motor speech disorders: clues to neurologic diagnosis. In *Parkinson's disease and movement disorders* (pp. 35-53): Springer.

Education, I. C. (2020, 21 Sep). Unsupervised Learning. Retrieved from https://www.ibm.com/cloud/learn/unsupervised-learning

Ferreira, A. J., & Figueiredo, M. A. J. P. R. L. (2012). Efficient feature selection filters for high-dimensional data. *33*(13), 1794-1804.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., Ngo, L. H. J. B. m. i., & making, d. (2012). Predicting sample size required for classification performance. *12*(1), 1-10.

Franciscatto, M. H., Trois, C., Lima, J. C. D., & Augustin, I. (2018). *Blending situation awareness with machine learning to identify children's speech disorders.* Paper presented at the 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT).

Gardner-Hoag, J., Novack, M., Parlett-Pelleriti, C., Stevens, E., Dixon, D., & Linstead, E. J. J. M. I. (2021). Unsupervised Machine Learning for Identifying Challenging Behavior Profiles to Explore Cluster-Based Treatment Efficacy in Children With Autism Spectrum Disorder: Retrospective Data Analysis Study. *9*(6), e27793.

Gupta, T., & Panda, S. P. (2019). *Clustering validation of CLARA and K-means using silhouette & DUNN measures on Iris dataset.* Paper presented at the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

Hailpern, J., Karahalios, K., DeThorne, L., & Halle, J. (2010). *Vocsyl: Visualizing syllable production for children with ASD and speech delays.* Paper presented at the

Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. J. J. o. i. i. s. (2001). On clustering validation techniques. *17*(2), 107-145.

Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018). *Classification, clustering and association rule mining in educational datasets using data mining tools: A case study.* Paper presented at the Computer Science On-line Conference.

Karbasi, S. A., Fallah, R., & Golestan, M. J. A. M. I. (2011). The prevalence of speech disorder in primary school students in Yazd-Iran. 33-37.

Karo, I. M. K., MaulanaAdhinugraha, K., & Huda, A. F. (2017). *A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function.* Paper presented at the 2017 Second International Conference on Informatics and Computing (ICIC).

Miller, I. (1966). Probability, random variables, and stochastic processes. In: Taylor & Francis.

Misuraca, M., Spano, M., Balbi, S. J. C. i. S.-T., & Methods. (2019). BMS: An improved Dunn index for Document Clustering validation. *48*(20), 5036-5049.

Petrovic, S. (2006). *A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters.* Paper presented at the Proceedings of the 11th Nordic Workshop of Secure IT Systems.

Piatetsky-Shapiro, G., Fayyad, U., Smith, P. J. A. i. k. d., & mining, d. (1996). From data mining to knowledge discovery: An overview. *1*, 35.

Razak, R. A., Madison, C. L., Siow, Y. K., Aziz, M. A. A. J. A. P. J. o. S., Language, & Hearing. (2010). Preliminary content validity and reliability of a newly developed Malay preschool language assessment tool. *13*(4), 217-234.

Razak, R. A., Neelagandan, A. I., Yusuf, N. M., Woan, L. H., Ahmad, K., & Madison, C. J. M. J. o. P. H. M. (2018). The validation of the Malay Preschool Language Assessment Tool (MPLAT): The screening and diagnostic versions. *1*, 191-115.

Ringnér, M. J. N. b. (2008). What is principal component analysis? , *26*(3), 303-304.

Sato, Y., Izui, K., Yamada, T., & Nishiwaki, S. J. E. S. w. A. (2019). Data mining based on clustering and association rule analysis for knowledge discovery in multiobjective topology optimization. *119*, 247-261.

Shahapurkar, S. S., & Sundareshan, M. K. (2004). *Comparison of self-organizing map with k-means hierarchical clustering for bioinformatics applications.* Paper presented at the 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541).

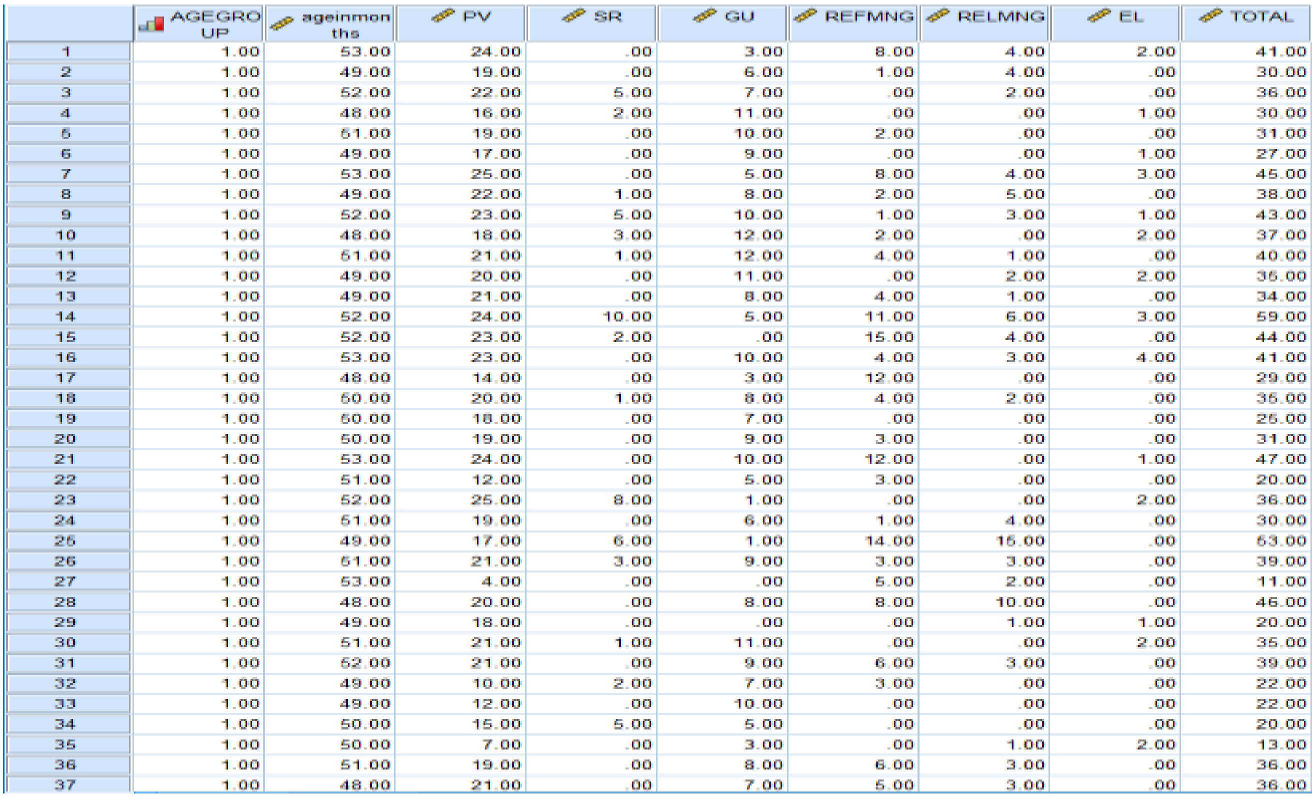
Stevens, E., Atchison, A., Stevens, L., Hong, E., Granpeesheh, D., Dixon, D., & Linstead, E. (2017). *A cluster analysis of challenging behaviors in autism spectrum disorder.* Paper presented at the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA).

Stevens, E., Dixon, D. R., Novack, M. N., Granpeesheh, D., Smith, T., & Linstead, E. J. I. j. o. m. i. (2019). Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *129*, 29-36.

Syakur, M., Khotimah, B., Rochman, E., & Satoto, B. D. (2018). *Integration k-means clustering method and elbow method for identification of the best customer profile cluster.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

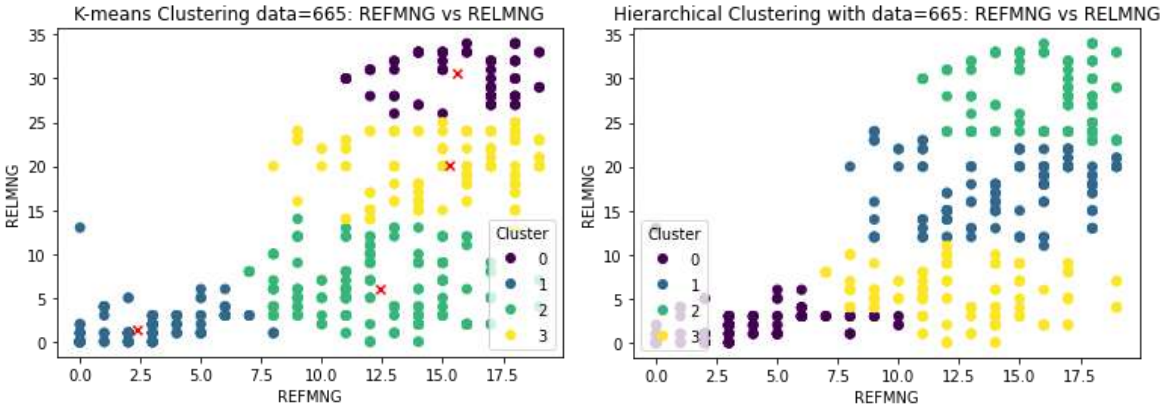
Waisakurnia, W. (2020). The Easiest Way to Interpret Clustering Result. Retrieved from https://towardsdatascience.com/the-easiest-way-to-interpret-clustering-result-8137e488a127

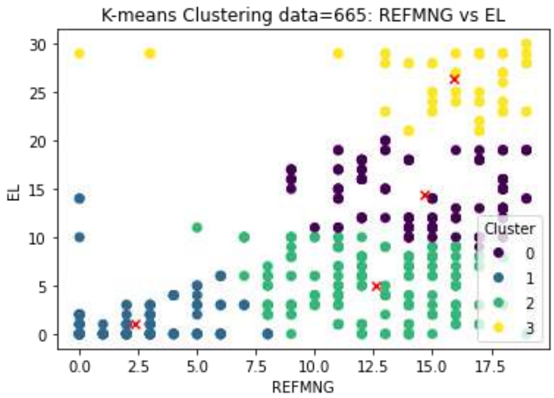
APPEDIX

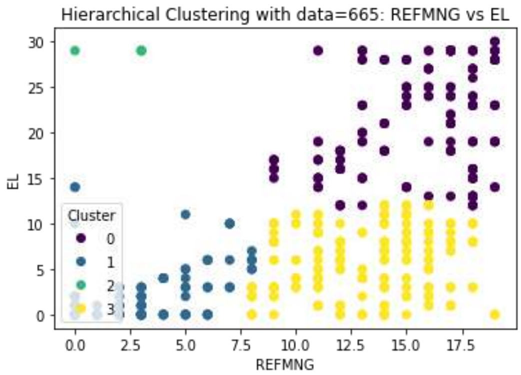


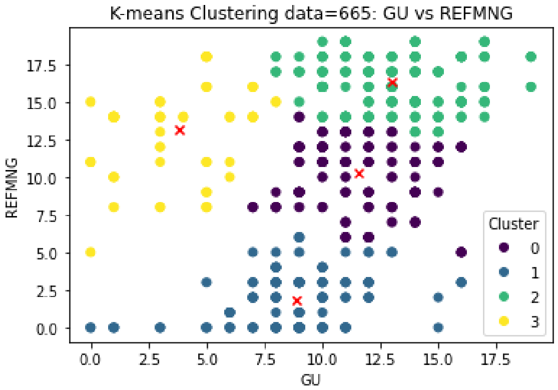
	AGEGROUP	ageinmonths	PV	SR	GU	REFMNG	RELMNG	EL	TOTAL
1	1.00	53.00	24.00	.00	3.00	8.00	4.00	2.00	41.00
2	1.00	49.00	19.00	.00	6.00	1.00	4.00	.00	30.00
3	1.00	52.00	22.00	5.00	7.00	.00	2.00	.00	36.00
4	1.00	48.00	16.00	2.00	11.00	.00	.00	1.00	30.00
5	1.00	51.00	19.00	.00	10.00	2.00	.00	.00	31.00
6	1.00	49.00	17.00	.00	9.00	.00	.00	1.00	27.00
7	1.00	53.00	25.00	.00	5.00	8.00	4.00	3.00	45.00
8	1.00	49.00	22.00	1.00	8.00	2.00	5.00	.00	38.00
9	1.00	52.00	23.00	5.00	10.00	1.00	3.00	1.00	43.00
10	1.00	48.00	18.00	3.00	12.00	2.00	.00	2.00	37.00
11	1.00	51.00	21.00	1.00	12.00	4.00	1.00	.00	40.00
12	1.00	49.00	20.00	.00	11.00	.00	2.00	2.00	35.00
13	1.00	49.00	21.00	.00	8.00	4.00	1.00	.00	34.00
14	1.00	52.00	24.00	10.00	5.00	11.00	6.00	3.00	59.00
15	1.00	52.00	23.00	2.00	.00	15.00	4.00	.00	44.00
16	1.00	53.00	23.00	.00	10.00	4.00	3.00	4.00	41.00
17	1.00	48.00	14.00	.00	3.00	12.00	.00	.00	29.00
18	1.00	60.00	20.00	1.00	8.00	4.00	2.00	.00	35.00
19	1.00	60.00	18.00	.00	7.00	.00	.00	.00	25.00
20	1.00	50.00	19.00	.00	9.00	3.00	.00	.00	31.00
21	1.00	53.00	24.00	.00	10.00	12.00	.00	1.00	47.00
22	1.00	51.00	12.00	.00	5.00	3.00	.00	.00	20.00
23	1.00	52.00	25.00	8.00	1.00	.00	.00	2.00	36.00
24	1.00	51.00	19.00	.00	6.00	1.00	4.00	.00	30.00
25	1.00	49.00	17.00	6.00	1.00	14.00	15.00	.00	53.00
26	1.00	51.00	21.00	3.00	9.00	3.00	3.00	.00	39.00
27	1.00	53.00	4.00	.00	.00	5.00	2.00	.00	11.00
28	1.00	48.00	20.00	.00	8.00	8.00	10.00	.00	46.00
29	1.00	49.00	18.00	.00	.00	.00	1.00	1.00	20.00
30	1.00	51.00	21.00	1.00	11.00	.00	.00	2.00	35.00
31	1.00	52.00	21.00	.00	9.00	6.00	3.00	.00	39.00
32	1.00	49.00	10.00	2.00	7.00	3.00	.00	.00	22.00
33	1.00	49.00	12.00	.00	10.00	.00	.00	.00	22.00
34	1.00	50.00	15.00	5.00	5.00	.00	.00	.00	20.00
35	1.00	50.00	7.00	.00	3.00	.00	1.00	2.00	13.00
36	1.00	51.00	19.00	.00	8.00	6.00	3.00	.00	36.00
37	1.00	48.00	21.00	.00	7.00	5.00	3.00	.00	36.00

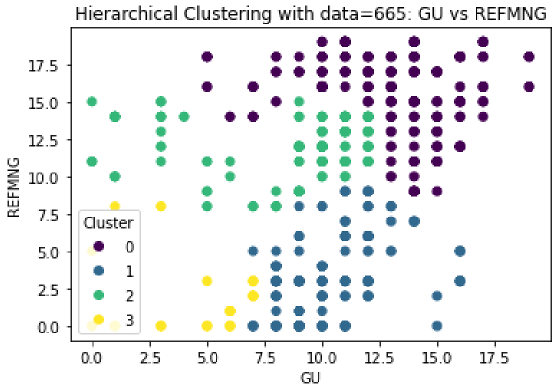
Figure A.1: MPLAT dataset

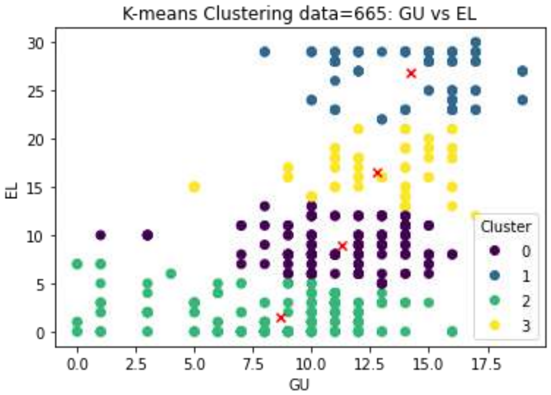


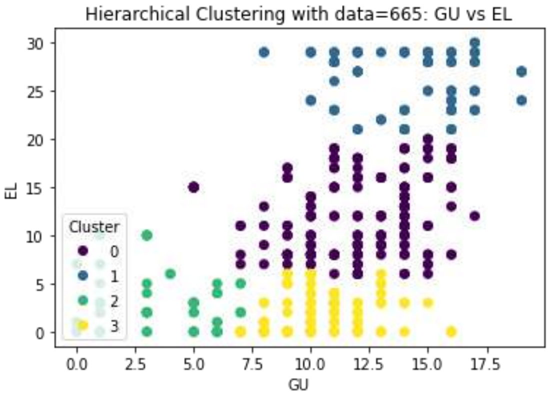

K-means Clustering data=665: REFMNG vs EL
EL
0
5
10
15
20
25
30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
REFMNG
Cluster
0
1
2
3

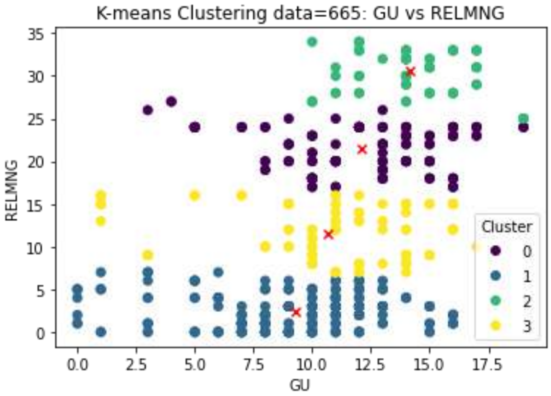

Hierarchical Clustering with data=665: REFMNG vs EL
EL
0
5
10
15
20
25
30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
REFMNG
Cluster
0
1
2
3

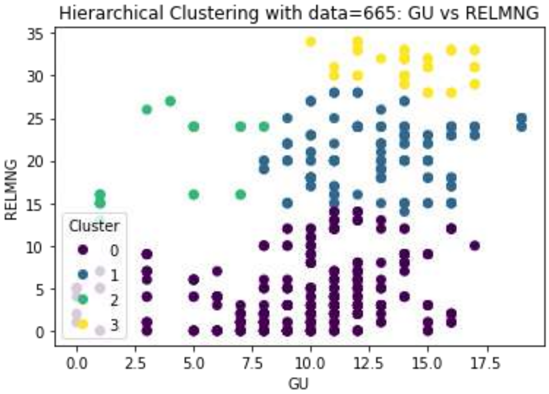

K-means Clustering data=665: GU vs REFMNG
REFMNG
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
GU
Cluster
0
1
2
3


Hierarchical Clustering with data=665: GU vs REFMNG
REFMNG
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
GU
Cluster
0
1
2
3


K-means Clustering data=665: GU vs EL
EL
0
5
10
15
20
25
30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
GU
Cluster
0
1
2
3


Hierarchical Clustering with data=665: GU vs EL
EL
0
5
10
15
20
25
30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
GU
Cluster
0
1
2
3


K-means Clustering data=665: GU vs RELMNG
RELMNG
0
5
10
15
20
25
30
35
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
GU
Cluster
0
1
2
3


Hierarchical Clustering with data=665: GU vs RELMNG
RELMNG
0
5
10
15
20
25
30
35
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
GU
Cluster
0
1
2
3

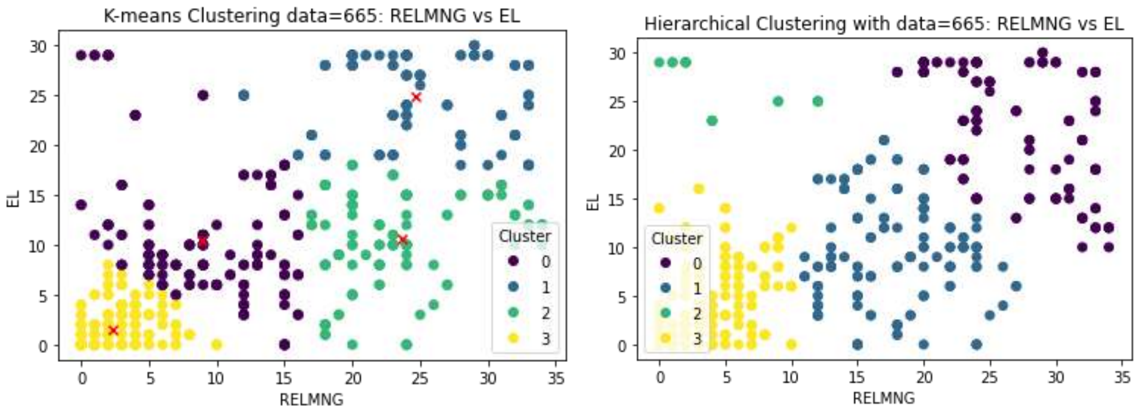


Figure A.2 Scatter plot of K-means and Hierarchical clustering's results