PREDICTION ON DRUG ABUSED BEFORE QUARANTINE OF THE PANDEMIC COVID-19 IN MALAYSIA (2000-2019)

MOHAMAD IQBAL BIN ABDULLAH
CA19010

BACHELOR OF COMPUTER SCIENCE
(COMPUTER SYSTEM & NETWORKING)
WITH HONOURS

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : MOHAMAD IQBAL BIN ABDULLAH

Date of Birth

Title : PREDICTION ON DRUG ABUSED BEFORE QUARANTINE OF THE PANDEMIC COVID-19 IN MALAYSIA (2000 – 2019)

Academic Session : SEMESTER II (2022 / 2023)

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☑	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

____
(Student's Signature)

____
(Supervisor's Signature)

Dr. NABILAH FILZAH BINTI MOHD RADZUAN

____
New IC/Passport Number
Date: 23 FEBRUARY 2023

____
Name of Supervisor
Date: 23 FEBRUARY 2023

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

SUPERVISOR'S DECLARATION

"I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Computer Science (Computer System & Networking) With Honours

____

(Supervisor's Signature)

Full Name : Dr. NABILAH FILZAH BINTI MOHD RADZUAN

Position : SENIOR LECTURER

Date : 23 FEBRUARY 2023

____

(Co-supervisor's Signature)

Full Name :

Position :

Date :

UMP
Universiti
Malaysia
PAHANG
Engineering • Technology • Creativity

STUDENT'S DECLARATION

"I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions"

____

(Student's Signature)

Full Name : MOHAMAD IQBAL BIN ABDULLAH

ID Number : CA19010

Date : 20 JANUARY 2023

PREDICTION ON DRUG ABUSED BEFORE QUARANTINE OF THE PANDEMIC COVID-19 IN MALAYSIA (2000-2019)

MOHAMAD IQBAL BIN ABDULLAH

Thesis submitted in fulfillment of the requirements
for the award of Bachelor of Computer Science (Computer System & Networking)
With Honours

Faculty of Computing

UNIVERSITI MALAYSIA PAHANG

JANUARY 2023

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God, the Almighty that has given me guidance, showered me his blessings and given me the opportunity to be alive on this face of this planet and given me time to complete this research successfully without any major obstacles.

I am very grateful and would like to express my sincere gratitude to my supervisor Dr. Nabilah Filzah Binti Mohd Radzuan, for her invaluable guidance, continuous encouragement, and constant support in making this project possible. I really appreciate her guidance from the initial to the final level that enabled me to develop an understanding of this research thoroughly. Without her advice and assistance, it would be a hard journey to complete this project. I also sincerely thanks Dr. Nabilah Filzah for the time spent proofreading and correcting my mistakes.

I also would like to express very special thanks to Dr. Danakorn Nincarean A/L Eh Phon for his tutor and clear explanation during this entire process of completing this project.

My sincere thanks go to all lecturers and staff of the Faculty of Computing in UMP that helped me in many ways and made my education journey at UMP pleasant and unforgettable memories. Many thanks go to my Degree member group for their excellent co-operation, inspiration, and supports during this study. I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream, and sacrifice throughout my life. I am thankful for their sacrifices, patience, and understanding that was inevitable to make this work possible. Their sacrifices had inspired me from the day I learned how to read and write until what I have become now. I cannot find the appropriate words that could properly describe my appreciation for their devotion, support, and faith in my ability to achieve my dreams.

Lastly, I would like to thank all people that contributed to my final year project directly or indirectly. I would like to acknowledge their comments and suggestions, which were crucial for the successful completion of this study.

ABSTRACT

In 2020, a new virus made a big impact on all people around the world, with the first cases identified on the 12th of December, 2019 in Wuhan City, Hubei Province in China which known as the Coronavirus Virus Disease 2019 that forced almost all nations to make a lockdown to control the virus from the spread and causes more damages not only to the industry but also the mental health of people. This is quite worrisome as it might cause the increase of drug abusers among Malaysian because, over the five years, the number of drug abuse decreases very slowly and might be increasing during the lockdown due to the pandemic. This study aims to predict drug abuse before the quarantine of the pandemic COVID-19 in Malaysia. In this context, drug abuse means the person who uses the drug either prescription or over-the-counter by other means from what of their purposes. To test the prediction on what drug will be abused by the Malaysian, a data mining method has been used to get the result. The method used are Linear Regression and Random Tree, an Unsupervised Machine Learning from the Regression to predict drug abuse which is how many drug abusers will be in Malaysia that required variables from datasets such as year, value and etc. The results show that the prediction used is almost 80% accurate from the datasets that have been used. These results suggest that the potential drug abuse is more likely to be unchanged or might increase in response to the many factors, especially mental health, by the COVID-19 and its variants.

ABSTRAK

Pada tahun 2020, virus baharu memberi impak besar kepada semua orang di seluruh dunia, dengan kes pertama dikenal pasti pada 12 Disember 2019 di Bandar Wuhan, Wilayah Hubei di China yang dikenali sebagai Penyakit Virus Koronavirus 2019 yang memaksa hampir semua negara. untuk membuat penutupan untuk mengawal virus daripada penyebaran dan menyebabkan lebih banyak kerosakan bukan sahaja kepada industri tetapi juga kesihatan mental orang ramai. Ini agak membimbangkan kerana ia mungkin menyebabkan peningkatan penyalahguna dadah di kalangan rakyat Malaysia kerana, dalam tempoh lima tahun, jumlah penyalahgunaan dadah berkurangan dengan sangat perlahan dan mungkin meningkat semasa perintah berkurung akibat pandemik. Kajian ini bertujuan untuk meramal penyalahgunaan dadah sebelum tempoh kuarantin pandemik COVID-19 di Malaysia. Dalam konteks ini, penyalahgunaan dadah bermaksud orang yang menggunakan dadah sama ada preskripsi atau tanpa kaunter dengan cara lain daripada tujuan mereka. Untuk menguji ramalan apakah dadah yang akan disalahgunakan oleh rakyat Malaysia, kaedah perlombongan data telah digunakan untuk mendapatkan hasilnya. Kaedah yang digunakan ialah Linear Regression dan Random Tree, Pembelajaran Mesin Tanpa Pengawasan daripada Regresi untuk meramalkan penyalahgunaan dadah iaitu berapa ramai penyalahguna dadah akan berada di Malaysia yang memerlukan pembolehubah daripada set data seperti tahun, nilai dan sebagainya. Keputusan menunjukkan bahawa ramalan yang digunakan adalah hampir 80% tepat daripada set data yang telah digunakan. Keputusan ini menunjukkan bahawa potensi penyalahgunaan dadah berkemungkinan besar tidak berubah atau mungkin meningkat sebagai tindak balas kepada banyak faktor, terutamanya kesihatan mental, oleh COVID-19 dan variannya.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

COVID-19	Corona Virus Disease 2019
ATS	Amphetamine Type Stimulant
WHO	World Health Organization
AADK	Agensi AntiDadah Kebangsaan
MCO	Movement Control Order
CMCO	Conditional Movement Control Order
RMCO	Recovery Movement Control Order
EMCO	Enhanced Movement Control Order
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus-2
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
KDD	Knowledge Discovery in Database
PDRM	Royal Malaysia Police
SOP	Standard Operation Procedure
RSSE	Root Relative Squared Error
RAE	Relative Absolute Error
RMSE	Root Mean Squared Error
MEA	Mean Absolute Error

CHAPTER 1

INTRODUCTION

1.1 Introduction

Coronavirus disease 2019 (COVID-19) has affected many people worldwide, both physically and mentally. Malaysia's first lockdown, dated on 18th of March 2020 remarks or indicated the first fear on the COVID-19 case began on the 4th of February as a 41-year-old man returned from Singapore. (Elengoe, 2020). During this lockdown, all the working sectors are forced to close while leaving only the essentials to remain open, such as hospitals and groceries, to prevent COVID-19 cases.

The lockdown is a boredom phase as some people have lost their jobs, some need to cope with the new normalization, and some just do not have anything to do. Some people cannot cope with this new normalization and do not know how to handle it correctly. So, the solution to get out of this stress, mess situation is by taking drugs. It is the simplest solution with so many contradictory results but with a higher cost that might take someone life for some people that have given up.

This research aims to predict the drug abused by the Malaysian before the quarantine of the pandemic COVID-19 which are from year 2000 until 2019. The drug also has been misused and abused before the quarantine for people to forget the hardship they needed to go on through the entire phases to reach herd immunity in Malaysia. For example, according to research conducted in America, there is increasing use of Methamphetamine and Amphetamine Type Stimulant (ATS) drugs that are considered in the stimulant type of drugs. (Abramson, 2021).

This research helps improve, understand, and contribute later to visualize it in a graph to raise the awareness of Malaysian on health issues, especially the use of drugs before the pandemic of COVID-19.

1.2 Problem Statements

The lockdown or quarantine during pandemic COVID-19 has made a big impact on the people worldwide, with the first cases identified on 12th December 2019 in Wuhan City, Hubei Province in China. (World Health Organization, 2020). During the lockdown, the drug abuse made by the Malaysians has been steadily the same as before the lockdown in the stimulant type of drugs, according to a statistic run by the Agensi AntiDadah Kebangsaan Malaysia (AADK). It should have been declining since there is not a single person allowed to go out working during the lockdown as there are a lot of police and soldier patrol on the road because of the government measure on the Movement Control Order (MCO) unless with a valid reason. It is important to detect and predict drug abuse all the time because drugs not only affect someone's health, but also the growth of the economic sector of a nation. By making early detection and prediction, early countermeasures can be made to avoid future drug abusers at an early age such as 14-years-old. (Bach, 2017). This is important to reach a drug-free nation to achieve a healthier lifestyle and safer environment for the future generation.

Malaysian still has low awareness of the health issues during the quarantine time in the last year. Mental health looks a bit worrying as there is an increasing drastically in suicide cases recorded in 2021 reached 468 suicides in the first month rather than the entire 2020, which are 631. (Idrus, 2021). Most of the reasons for these suicides' cases are troubled family relationships, emotional pressure, and financial issues. Besides suicides, issues are also found increasing, such as pneumonia, depression, obesity because of lack of exercise, and many more. People need to take good care of both mental and physical health seriously, especially during these pandemics of COVID-19, as most of the movements are restricted and controlled by the police and soldiers based on government measurement. May this research and the Ministry of Health of Malaysia promote and encourage keeping a good healthier lifestyle during these quarantines of the pandemic COVID-19.

1.3 Objectives

There are three objectives covered in this research which are:

i. To study the trend of the drug abusers on Malaysian before the pandemic COVID-19 from year 2000 until 2019.

ii. To predict the usage of drug on Malaysian before pandemic COVID-19 by detecting earlier of the potential for future pandemic from year 2000 until 2019.

iii. To visualize the yield of study and raise awareness of Malaysian on health issues before the pandemic COVID-19.

1.4 Scopes

The research's scopes are as listed as follows based on the objective declared.

i. The data will be collected before the quarantine of the pandemic COVID-19 in the Malaysian region only.

ii. The data prediction and output will be displayed in an organized and more straightforward way so the reader can understand them very well.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

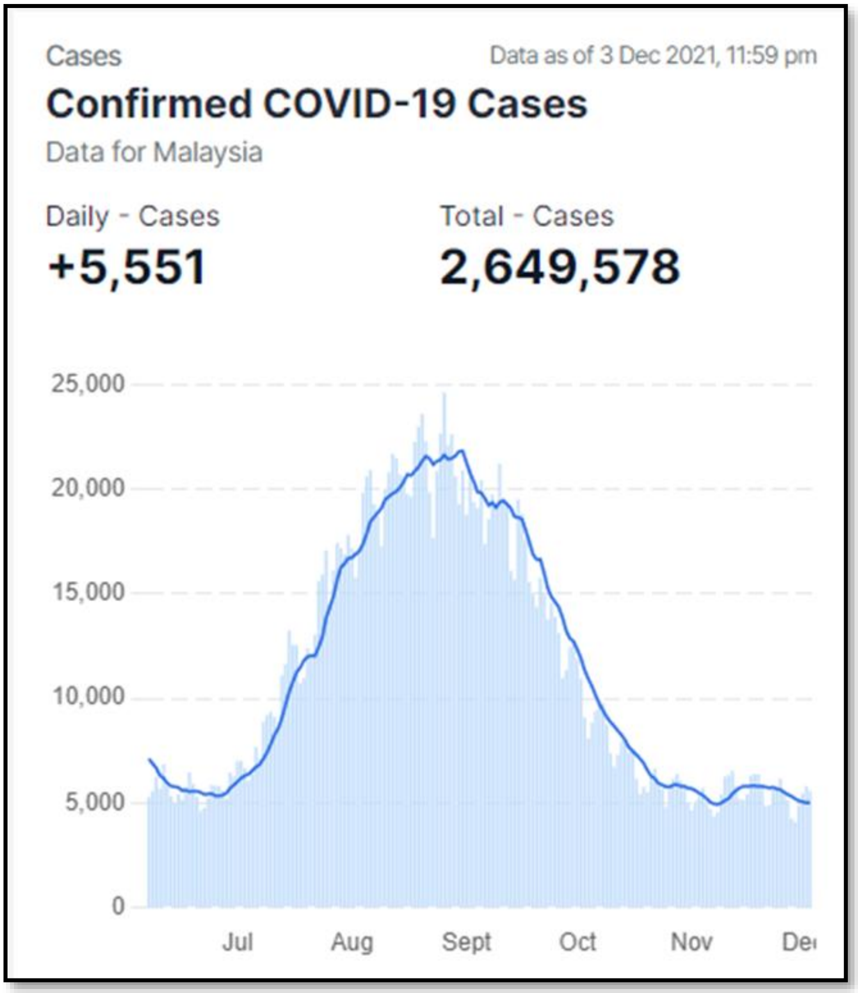
There are many kinds of research that have been done by people worldwide on the impact of the pandemic COVID-19 on many countries, including Malaysia itself. Most of the researches focus on the impact on a large scale that mainly on mental health and the economy. This research will focus on drug abuse prediction, and the study of existing researches has been classified as shown in Table 2.1. The part that will be covered in this chapter is about the Coronavirus Disease-2019 (COVID-19), the drug abused in Malaysia, quarantine of the pandemic COVID-19 in Malaysia, data mining (predictive and descriptive part), the comparison of the existing thesis and lastly the proposed study. The comparison is made based on the author, abstract, objective, pros and cons for the existing thesis.

The first part is about the origin of COVID-19 and its current situation on the worldwide view. Next is drug abuse in Malaysia, which has been a quite popular unsolved issue in Malaysia for a decade till the COVID-19 comes to say hello to humans. Then, the duration quarantine of the pandemic COVID-19 in Malaysia from the first until the latest movement control order from the government. Move on to the method used to get the prediction on the drug abused is the data mining and its comparison between both of it. Last but not least, a comparison between the existing thesis related to the COVID-19, drug abuse in Malaysia, and the quarantine of the pandemic COVID-19 in Malaysia that is based on its author, abstract, objective, pros, and cons.

2.2 Coronavirus Disease-2019 (COVID-19)

The Coronavirus Disease-2019 is the beginning of the nightmare for humans on earth. Starting in December 2019, a young adult in Wuhan, the capital city of Hubei province, was admitted to local hospitals was suspected of severe pneumonia with unknown cause. To make it worse, on the 31st of December, China has announced that the COVID-19 had made its outbreak to the World Health Organization (WHO), and on the 7th of January 2021, it was identified as a coronavirus that has >95% homology with bat coronavirus and more than 70% similarity with the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) (Singhal, 2020). Next, on the 11th of January, the first fatal case was reported as it was the massive migration of Chinese during the Chinese New Year Eve not only in China, but also to many other countries such as Thailand, Japan, and South Korea reported that they are from Wuhan.

The first case outside China is coming on to Thailand on the 8th of January 2020, according to the World Health Organization (WHO) (2020). She is a 61 years old woman who was a tourist from Wuhan and was hospitalized in Nonthaburi province. (STAT, 2020). The first COVID-19 case detected in Malaysia was on the 25th of January 2020 from 3 Chinese nationals who had close contact with the infected in Singapore travelled via Singapore to get into Malaysia and then were treated at Sungai Buloh Hospital Selangor (Elengoe, 2020). It is the beginning of the first pandemic that only God knows when it will end. Until today, the total case recorded in Malaysia has reached 2.64 million, with a median of 5200 cases over the past month, which was November.



Cases

Data as of 3 Dec 2021, 11:59 pm

Confirmed COVID-19 Cases

Data for Malaysia

Daily - Cases

+5,551

Total - Cases

2,649,578

Figure 2.1 Total Cases Recorded in Malaysia

COVID-19 is a beta coronavirus that contains four main structural proteins, which are Spike Glycoprotein (S), Membrane (M), Envelope (E), and Nucleocapsid (N) that used to neutralize the antibodies of humans, making them weak to viruses and can be transmitted via droplet or close contact with the infected person either symptom with such as fever, dry cough, and headache, or even asymptomatic person. Even though its mortality rate is lower, around 2.10%, it can increase proportionally with age and disease compared to SARS-CoV, which has around 10% mortality, and MERS-CoV that reached about 35% (Elengoe, 2020).

2.3 Drug Abused in Malaysia

Drug abuse in Malaysia has been very famous since the pre-independent era of Malaysia, which was around 1946s. During that time, the drug abused was Opium, which is a Depressant type of category that made the consumer have a chill and relaxing mood that depends on the doses consumed. It was consumed by China's immigrants, and it was introduced by the British colonialists that worked in Malaya Union (before Malaysia) (Rusdi AR). Later in the post-independence era around the 1960s, a new increasing use of drugs where the youngsters that get known to the Hippy subculture especially in Malays and in 1980s, the most popular drug that all known until today which was the heroin was becoming the national threat that forced the government to introduce national anti-drug task force to solve these problems.

In this thesis, the drugs will be predicted earlier by applying machine learning algorithm to the existing datasets. A depressant is the type of drug that can slow down the messages being transformed from the human's brain to the entire body and can affect their activities, such as concentrating on an easy job that even an ordinary person can do with ease. Next is the Stimulant Drugs, which is the reverse effect of the Depressant as it boosts up the messages transformed from the brain to the body. Stimulant's type can raise blood pressure, makes the heartbeat faster than an average person because of its effect on the organ in the human body as it basically stimulates the nerves to be more functional than usual. Hallucinogen's drug is affecting human's five senses, which are smell, touch, sight, hear and taste. This can be seen with someone taking the Hallucinogen, which will easily get confused and disorientated when interacting with them. Finally, the Doctor Prescription is a type of medicine (can be categorized as a drug) that serves as a painkiller or tranquillizer. This drug always gets wrong used by the addicted in terms of dosage because of the pain they feel, and the higher the dosage taken, the lower and slower the pain nerves will become. Figures 2.2 and 2.3 show the drug's types with explained details collected cases from 2016 till 2020.

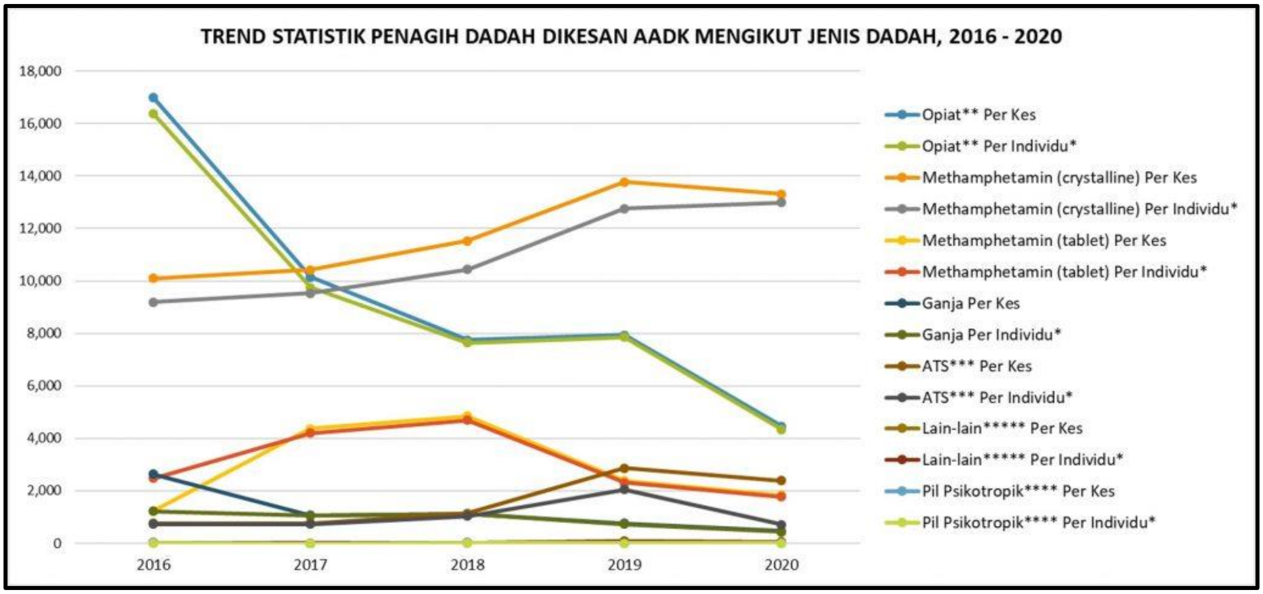
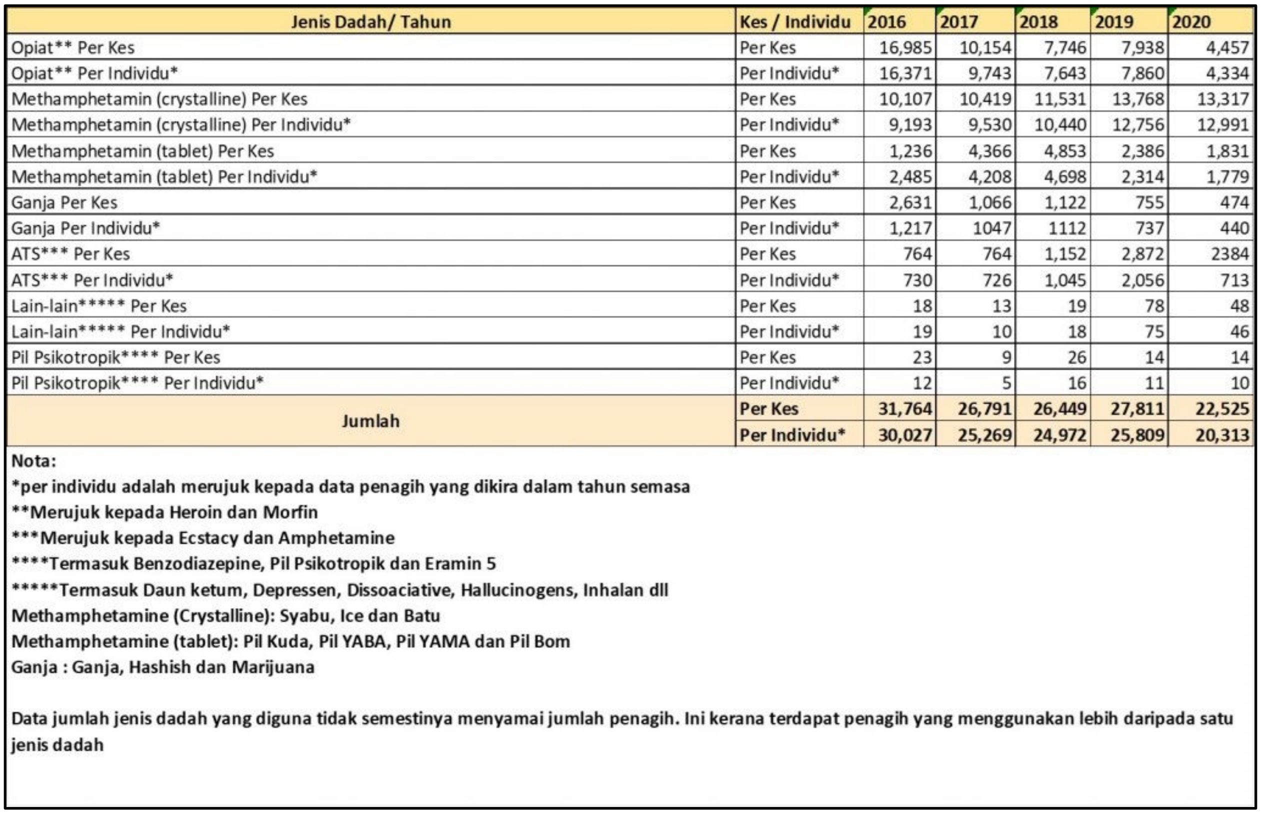


Figure 2.2 The Statistics of Drugs in Malaysia from 2016-2020

The drugs collected are divided into two parts: cases and per individual from 2016 until 2020 from the AADK’s website. There are many other types of drugs, but in these cases, these are the famous ones and mostly being consumed by Malaysians regardless of race, gender and age. The drugs are Opiate, Methamphetamine in crystalline and tablets, Marijuana, ATS, Psychotropic Pills and etc. Figure 2.3 below shows the number collected and certain drugs collected with their name.



Jenis Dadah/ Tahun	Kes / Individu	2016	2017	2018	2019	2020
Opiat** Per Kes	Per Kes	16,985	10,154	7,746	7,938	4,457
Opiat** Per Individu*	Per Individu*	16,371	9,743	7,643	7,860	4,334
Methamphetamin (crystalline) Per Kes	Per Kes	10,107	10,419	11,531	13,768	13,317
Methamphetamin (crystalline) Per Individu*	Per Individu*	9,193	9,530	10,440	12,756	12,991
Methamphetamin (tablet) Per Kes	Per Kes	1,236	4,366	4,853	2,386	1,831
Methamphetamin (tablet) Per Individu*	Per Individu*	2,485	4,208	4,698	2,314	1,779
Ganja Per Kes	Per Kes	2,631	1,066	1,122	755	474
Ganja Per Individu*	Per Individu*	1,217	1047	1112	737	440
ATS*** Per Kes	Per Kes	764	764	1,152	2,872	2384
ATS*** Per Individu*	Per Individu*	730	726	1,045	2,056	713
Lain-lain***** Per Kes	Per Kes	18	13	19	78	48
Lain-lain***** Per Individu*	Per Individu*	19	10	18	75	46
Pil Psikotropik**** Per Kes	Per Kes	23	9	26	14	14
Pil Psikotropik**** Per Individu*	Per Individu*	12	5	16	11	10
Jumlah	**Per Kes**	**31,764**	**26,791**	**26,449**	**27,811**	**22,525**
	Per Individu*	**30,027**	**25,269**	**24,972**	**25,809**	**20,313**

Nota:
***per individu adalah merujuk kepada data penagih yang dikira dalam tahun semasa**
****Merujuk kepada Heroin dan Morfin**
*****Merujuk kepada Ecstacy dan Amphetamine**
******Termasuk Benzodiazepine, Pil Psikotropik dan Eramin 5**
*******Termasuk Daun ketum, Depressen, Dissoaciative, Hallucinogens, Inhalan dll**
Methamphetamine (Crystalline): Syabu, Ice dan Batu
Methamphetamine (tablet): Pil Kuda, Pil YABA, Pil YAMA dan Pil Bom
Ganja : Ganja, Hashish dan Marijuana

Data jumlah jenis dadah yang diguna tidak semestinya menyamai jumlah penagih. Ini kerana terdapat penagih yang menggunakan lebih daripada satu jenis dadah

Figure 2.3 The Statistics of Drugs in Malaysia from 2016-2020 2

This figure shows that Methamphetamine refers to the data collected in the year. Opiate drugs related to heroin and morphine. The ATS refer to ecstasy and amphetamine while the etc drugs are referred to the benzodiazepines, some psychotropic pills that different from the depressant, hallucinogen and inhalant in the psychotropic pills statistic. Table 2.1 below will give a brief understanding of the drugs, their effect, usage in 2020 and other examples of drugs in that category.

	Depressant	**Stimulant**	**Hallucinogen**	**Dr. Description**
Effect	• Feel relax • Chill • Lack of motivation	• Talkative • Energetic • Seizure • Insomnia • Anxiety	• Affect our 5 sensors • Produce visual distortion	• Dreamy • Rush of pleasure
Usage (2020) [3]	• Per case: 2733 • Per individual: 2685	• Per case: 17532 • Per individual: 15483	• Per case: 24 • Per individual: 23	• Per case: 2235 • Per individual: 2172
Example	•Marijuana / Cannabis •Opiate: -Heroine -Opium -Codeine -Morphine •Benzodiazepines /Tranquilizer -Midazolam -Lorazepam -Diazepam • Alcohol • Barbiturates	• Nicotine • Caffein • Amphetamine •Methamphetamine (Ice, Syabu) • Cocaine • New Psychoactive Substances (Cannabis Synthetic)	• Ketamine •LSD (Lysergic Acid) • PCP (Phencyclidine)	• Sleeping Pill/Tranquilizer •Benzodiazepines family • Pain Relief Opiate -Morphine -Pethidine -Tramadol

Table 2.1 The Types of Drugs

2.4 Quarantine of the Pandemic COVID-19 in Malaysia

The Malaysian government has announced its first lockdown named the Movement Control Order (MCO), starting from the 18th of March in 2020. This lockdown or quarantine is under the Prevention and Control of Infectious Disease Act 1988 that have a big impact not only on the movement control but also on main sectors such as the health and economy. During these quarantines, there are a lot of sectors that need to close, such as the business, religious, sports, social, cultural activities, and schools, except the essentials such as pharmacy, clinics, and supermarkets for the grocery items. There are four phases during this MCO, which are from Phase 1 until Phase 4 that start from 18th of March until the 3rd of May, 2020. If someone without any valid reasons breaks those MCO rules, they will be fined up to RM1000 and/or jailed for not more than six months or both. The task force that is responsible for this MCO lockdown is approximately 7000 militaries deployed from Royal Malaysia Police (PDRM) to conduct the roadblocks operation at the border of every state (Wikipedia contributors, 2021).

Phase	Date
Movement Control Order (MCO) (PKP) (18 March 2020 – 3 May 2020)	
Phase 1	18 March 2020 – 31 March 2020
Phase 2	1 April 2020 – 14 April 2020
Phase 3	15 April 2020 – 28 April 2020
Phase 4	29 April 2020 – 3 May 2020

Table 2.2 The First MCO By Phase During Pandemic

Then after the MCO, the government introduced a new movement control named the Conditional Movement Control Order (CMCO) and Recovery Movement Control Order (RMCO). This quarantine duration started from the 4th of May 2020 until the 31st of March 2021. This quarantine gives some flexibility on the rules to open up more sectors to raise the country's economic growth that has been lost due to the lockdown for about two months. In the CMCO, most of the economic sectors has been opened up but under the strong supervision from the worker with certain Standard Operation Procedure (SOP) from the government such as scan body temperature, log in to MySejahtera application and always maintain of 1-meter social distancing. In RMCO, the rules in CMCO is being lightened more as more things can be done by the Malaysian such as religious activities being allowed again with some restrictions and the resuming operations of the academic sectors such as daycares, school, high school and universities. The thing that most people are waiting for is the interstate travel allowed from the 10th of June except for those areas remain under the Enhanced Movement Control Order (EMCO). Table 2.3 shows the following duration of CMCO and RMCO by Phase during the pandemic.

Phase	**Date**
Conditional Movement Control Order (CMCO) (PKPB) (4 May 2020 – 9 June 2020)	
Phase 1	4 May 2020 – 12 May 2020
Phase 2	13 May 2020 – 9 June 2020
Recovery Movement Control Order (RMCO) (PKPB) (10 June 2020 – 31 March 2021)	
Phase 1	10 June 2020 – 31 August 2020
Phase 2	1 September 2020 – 31 December 2020
Phase 3	1 January 2021 – 31 March 2021

Table 2.3 The Following Duration Quarantine by Phase During Pandemic

This thesis will cover duration only until year 2019. This is because of the insufficient datasets provided by the Malaysian's government that provide datasets of drug abusers up until year 2020.

2.5 Data Mining

The process used for this thesis is by data mining of sorting through a lot of data sets taken from various sources to get the prediction on the drug abused as output (Stedman & Hughes, 2021). There are two categories of techniques in data mining, one is Descriptive, and another one is Predictive. In data mining, there are several steps from collecting the raw data until it becomes useful, targeted data or new knowledge that will be presented to the people out there.

The process of getting the knowledge first is the selection of data. The selection of data starts from collecting data from certain phenomena or, in this case, is from the data abuser and before the new pandemic of COVID-19 within the quarantine duration rules by the Malaysian government. Data can be collected through various available websites such as www.kaggle.com, where all kinds of datasets are available for the public use.

Next is how the targeted data go to preprocessing to get preprocessed data. It is one of the important operations here. An algorithm is used here to detect a certain point called outlier to set a goal for the targeted data. For example, the dataset obtained from www.kaggle.com can use the algorithm to detect the wanted values in the distribution of datasets. From the preprocessed data to go into transformed data, it is to normalize the dataset such as its attributed from 0 to the desired value. Another step is discretizing the dataset in order to grouping the datasets according to the required categories.

Then, transformed data will be mined to get a new pattern which we want to apply. To get the pattern, the user needs to choose between the supervised and unsupervised computational methods (javatpoint, n.d.). The differences between these supervised and unsupervised are that in supervised data mining, the algorithm learns from what data the user has input in order to predict the unknown result, while on the other hand, the unsupervised want to learn about the pattern based on the correlation of the datasets while the user already known on the result (GeeksforGeeks, 2021). This thesis will be used as the unsupervised one since drug prediction is the goal of this thesis, and there are a lot of drugs available out there in and off the market.

Lastly, it is interpreted from the patterns to obtain knowledge given by that pattern and see if there are irrelevant patterns from the data. From the entire process, from getting the raw data until acquiring the knowledge, the user can understand steps by steps on how the data is being processed to get valuable outcomes which is drug prediction based on this thesis.

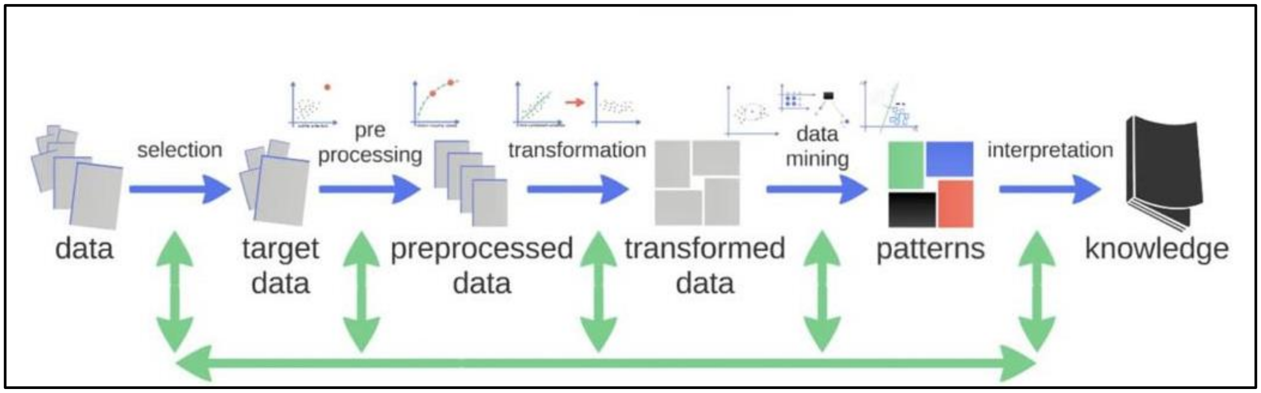


Figure 2.4 The Process of Data Mining

Since this thesis uses the unsupervised learning techniques to get higher accuracy for the outcomes, there are many regression algorithms available such as the Linear Regression, Support Vector Machines, Neural Networks and Random Tree. Table 2.4 below shows the advantage and disadvantages of these algorithms (GeeksforGeeks, 2020).

Algorithm	Advantages	Disadvantages
Linear Regression	•The estimate of unknown parameters obtained from linear are optimal •Use data efficiently •Good results obtained from relatively small datasets •Easy to interpret	• Output can lie outside the range •Sensitive to outlier
Support Vector Machines	•Not biased by outliers •Flexibility in choice of form threshold	•Not suitable for large dataset attributes •Lack of transparency results
Neural Network	•Adaptive learning •Preferred for recognition •Highly fault tolerant	•Computationally expensive learning process •Scalability
Randoms Tree	• It is easier to understand the result because of its visualization •Able to evaluate numerical and nominal attributes	•Sometimes generate complex tree •Prediction is not continuous

Table 2.4 The Comparison of Existing Algorithm of Regression’s Unsupervised

2.6 Comparison of the Existing Thesis

In order to complete this thesis, a comparison between the article journal needs to be analyzed to get a brief on what the article journal wants to let the user knows about. This comparison is crucial for the development and completion of this thesis. The comparison is made based on the journal's title, author, abstract, pros, and cons. Table 2.1 shows the comparison of the existing thesis that is related to this thesis's study.

An Innovative Data Mining and Dashboard System for Monitoring of Malaysian Dengue Trends

A study about data mining and dashboard system for monitoring Malaysian's dengue trends (Jamil & Shaharanee, 2016) used the time-series approach of data mining techniques such as Exponential Smoothing, Moving Average and Simple Linear Regressions to get the correct prediction about the dengue outbreak issues in Malaysia from 2010 until 2015. The dataset used for this research was obtained from The Public Sector Open Data Portal (data.gov.my), which acted as an online one-service centre linked with the Malaysian government during that time with implemented the KDD Data Mining methodology. The preprocessing methods were done by implementing data cleaning, data integration, data transformation, data reduction and data discretization on the datasets. The parameters used in this project were ID, Negeri, Minggu 1, Minggu 2, Minggu 3 … Minggu 260, and Value. Then it visualizes the summarization of the datasets to give the user a clearer picture of the dengue outbreak in Malaysia. After getting each result for the model applied to the datasets, the best model will be selected based on the best fit to explain the relationship between variables. Last but not least, the results of the model training will be displayed using Dashboard System for Monitoring of Malaysian Dengue Trends made by the author.

Pros	Cons
- Datasets used is certified and trusted - Using SAS Enterprise Miner Workstation 13 for developing the data mining technique - Clear presentation of forecast dengue trend from 2010 until 2020. - One related used of data mining techniques for this thesis - Similar methodology used which is KDD methodology. - Similar attributes such as State and Year	- Unknown approaches used for data preprocessing's part - Prediction about dengue outbreak on last 10 years results - Test techniques did not mention

Application of data mining techniques and logistic regression to model drug use transition to injection: a case study in drug use treatment centers in Kermanshah Province, Iran

A study about data mining techniques and logistic regression to investigate the risk factors between drug use to injection. This article used the classification models to determine the factors related to drug use transition to injection in order to construct an effective prevention method by identifying PWID (Person Who Inject Drugs) earlier (Najafi-Ghobadi et al., 2019). Datasets were collected through the treatment centres in Kermanshah by following the rules and getting permission under the proper guidance of the Kermanshah University of Medical Sciences, consisting of 2098 records. The attributes used were Gender, Marital Status, Occupational Status, Housing Status, Education, The First Used Drug, Motivation for Starting The Drug Use, Family History of Drug Use, History of Taking Opium, History of Taking Hallucinogens, History of Taking Crystal, History of Taking Heroin, History of Taking HashHish, History of Taking Sap, History of Taking Cocaine, History of Taking Sedative, History of Taking Amphetamine, History of Taking Methadone, History of Taking Cigarette, History of Taking Alcohol, History of

Overdose, History of Suicide, History of Mental Disorder, History of Mental Disorder in Family, History of Prisons, Number of Referrals to Drug Treatment Centres. During the preprocessing, the missing data and outliers were checked using CART regression trees to deal with the missing data. Anomaly detection was used to detect the outlier, and then it was deleted as the outlier did not affect the result. Four data mining models were implemented: the Neural Network, Logistic Regression, Support Vector Machine and lastly, Decision Tree and were tested using 70% of training data and 30% of testing data, and also 90% cross for training while 10% for validation. The results of these models were displayed using the software.

Pros	Cons
- Datasets used is certified and trusted under law - Missing value is settle by using CART regression tree - Four famous data mining models used which were Decision Tree, Neural Network, Logistic Regression and Support Vector Machine - Testing techniques were cross validation and split training and testing data	- Unknown approaches used for data preprocessing's part - Unknown used of software - Unknown methodology used for this article - Mostly of the attributes were nominal -

2.7 Proposed Study

This study intends to use data mining and visualization techniques to get the best prediction on the drug abused by Malaysian. The best duration to measure this prediction is by the last nineteen years. However, it is too long since this thesis has only two semesters which is quite struggling to produce an impressive result, and that is why the duration taken is short from the start duration of the MCO until the end of RMCO. After determining the duration, the scope is set onto Malaysia, and the process begins. The first thing needed to be done to understand the dataset implemented in the study. The dataset must be relevant to the study, or otherwise, it will disturb the process later in the algorithm, which will affect the result and consume the time and cost needed to redo all the processes back. After the dataset has been confirmed, it is time to do the processing part. An algorithm will be used to detect a pattern of the dataset and collect valuable information from it, and that is why the linear regression from the regressions of the supervised is being used to predict a discrete response, which means that the data is being divided into many categories. Unsupervised learning is used in this study because the input and output are already known, and that is why the dataset is being trained to be reasonable to respond to that new data output and prediction on which drug. After getting the purpose of this study using the desired algorithm, the data will be presented using the Power BI application by Microsoft. It is a very powerful tool that can be tweaked according to the user's desires.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter discusses the concept of the methodology used for doing the study on the prediction on drug abused before quarantine of the COVID-19 pandemic in Malaysia. A data mining methodology is a set of methods, policies, and procedures that a development team employs to carry out a data mining methodology. The challenges in choosing and selecting the most appropriate methodology, as well as following that methodology, are to do so wisely so that high-quality data mining that meets all requirements can be delivered while staying on schedule, which is critical for business success, and avoiding steps that waste time, money, resources, and squander productivity.

3.2 Research Framework

The table shows the framework of the research and its components, which include data collecting, preprocessing, implementing modelling, and integrating the result with visualization software. The research starts with acquiring a drug dataset from reliable resources based in Malaysia. For this task, the unified database must be from trusted such as government companies. The following is to preprocess the data. There are many processes happening here, such as data cleaning. Once the preprocessing data has met the standard requirements in terms of statistical analysis, the researcher can proceed to the data transformation. In data transformation, there are two techniques: discretization and normalization. These two techniques are essential to make the preprocessed datasets able to be used by the model later. The first technique, data discretization, means grouping the data into several groups, while the other technique, data normalization, is transforming data into another format to minimize or maximize it into scaling datasets. Then, the model is applied after these datasets are transformed into a new dataset that can be used for data mining. In the selection of a model of data mining, this is crucial as the researcher must understand the model that will be used on the transformed dataset. The configuration, the value, and the output of the model must be understood for the researcher to predict the model's performance. The model's performance is vital as it can affect the result of the output made by the model later. This result is used in the visualization to let the user know what is going on with the knowledge gained from the entire process. The visualization used in this research is by using Power BI, software from Microsoft, to enrich the presentation of the knowledge discovered during the whole process**.**

Dataset	Preprocessing	Data Transformation	Model Mining	Evaluation / Visualization
Dataset collected from website	Dataset preprocessing	Dataset transformation	Data mining with selected model	Data evaluation and visualization

Table 3.1 The Research Process Workflow of Proposed Study

3.2.1 KDD Methodology

The Knowledge Discovery from Data (KDD) methodology is a popular analytic data mining process used in the cross-industry. The KDD methodology has four main parts: Data Selection, Preprocessing, Transformation, Data Mining, and Evaluation (University of Regina DBD, n.d.). In each phase, there are several things that need to be done in order to move on to the next phases. Each of the phases needs to do carefully before proceeding to the next phase. Once entering the next phase, the researcher cannot return back to the previous phase to fix something that caused the current phase to make the results not good enough to meet the expectation of the model evaluation. This is because it is unwise to do so, as the time spent for each phase is long enough. After all the process has met its expected result, then lastly can validate the knowledge gained from the entire phases. There are many methodologies that are available out there to be used, but in this case, it is preferable to use the KDD as the business understanding in this thesis is not being used, thus making the Cross Industry Standard Process for Data Mining (CRISP-DM) not likely a better choice to be used (Hazarhun, 2022). The business understanding of the CRISP-DM methodology is really consuming time that might affect the timeline set for this thesis.

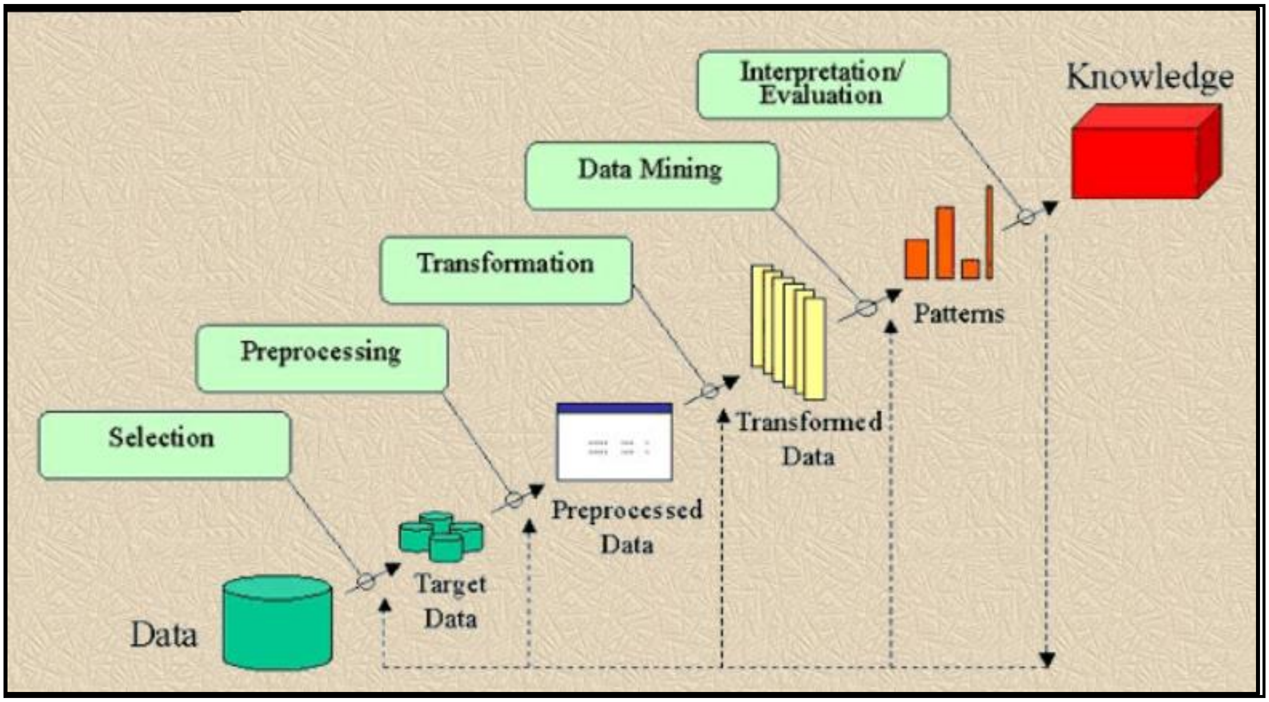
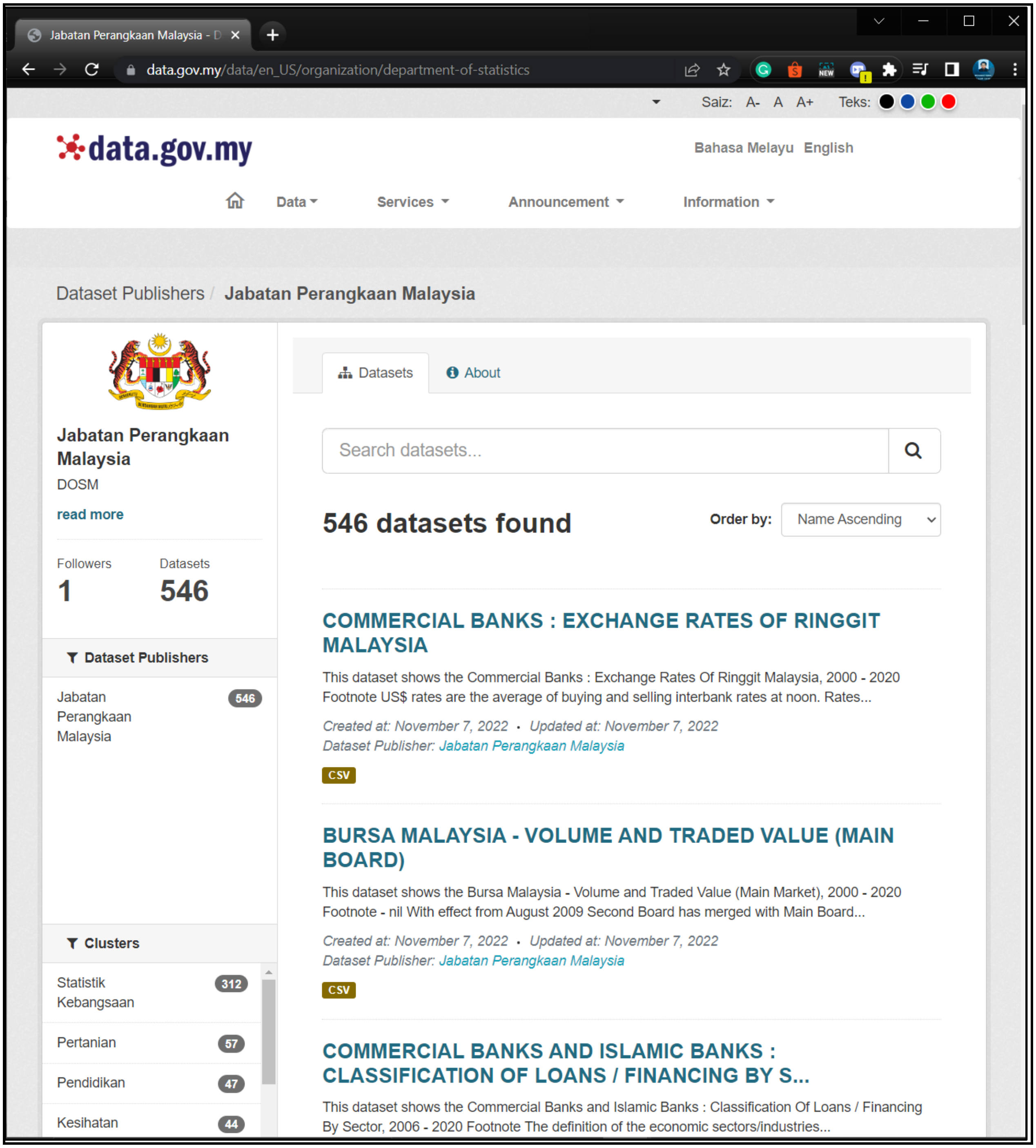


Figure 3.2 The Five Phases in KDD Methodology

3.2.2 Data Selection

Starting with the selection of data, no matter from where it is sourced, as long as it is a reliable and trusted source. During this process, the user needs to understand the goals, application, and knowledge about the tasks. Creating a targeted dataset from the entire whole is based on the subset of variables of the data sample. This is to mainly focus on the required subset of data that are going to be mined rather than including the ones that do not need. After carefully selecting the subset of data from the entire data, we can then proceed to the next cycle, which is the preprocessing part. A good set of datasets can lead to good results in data mining after applying a model to it. It is advisable to pick only datasets with minimum attributes that will be used as it can reduce the time taken to reduce the preprocessing phase later. The goal is to identify the least number of columns as much as possible from the datasets that are related during building a model (Marlboro, 2022). Figure 3.2.1 above shows that there are about 546 datasets found in the Department of Statistics Malaysia. From there, to select the required subset of data, the user must find the specific keyword to get the desired dataset.



Dataset Publishers / Jabatan Perangkaan Malaysia

Jabatan Perangkaan Malaysia

DOSM

Followers 1

Datasets 546

Dataset Publishers

Jabatan Perangkaan Malaysia	546

Clusters

Statistik Kebangsaan	312
Pertanian	57
Pendidikan	47
Kesihatan	44

546 datasets found

COMMERCIAL BANKS : EXCHANGE RATES OF RINGGIT MALAYSIA

This dataset shows the Commercial Banks : Exchange Rates Of Ringgit Malaysia, 2000 - 2020 Footnote US$ rates are the average of buying and selling interbank rates at noon. Rates...

Created at: November 7, 2022 • Updated at: November 7, 2022
Dataset Publisher: Jabatan Perangkaan Malaysia

CSV

BURSA MALAYSIA - VOLUME AND TRADED VALUE (MAIN BOARD)

This dataset shows the Bursa Malaysia - Volume and Traded Value (Main Market), 2000 - 2020 Footnote - nil With effect from August 2009 Second Board has merged with Main Board...

Created at: November 7, 2022 • Updated at: November 7, 2022
Dataset Publisher: Jabatan Perangkaan Malaysia

CSV

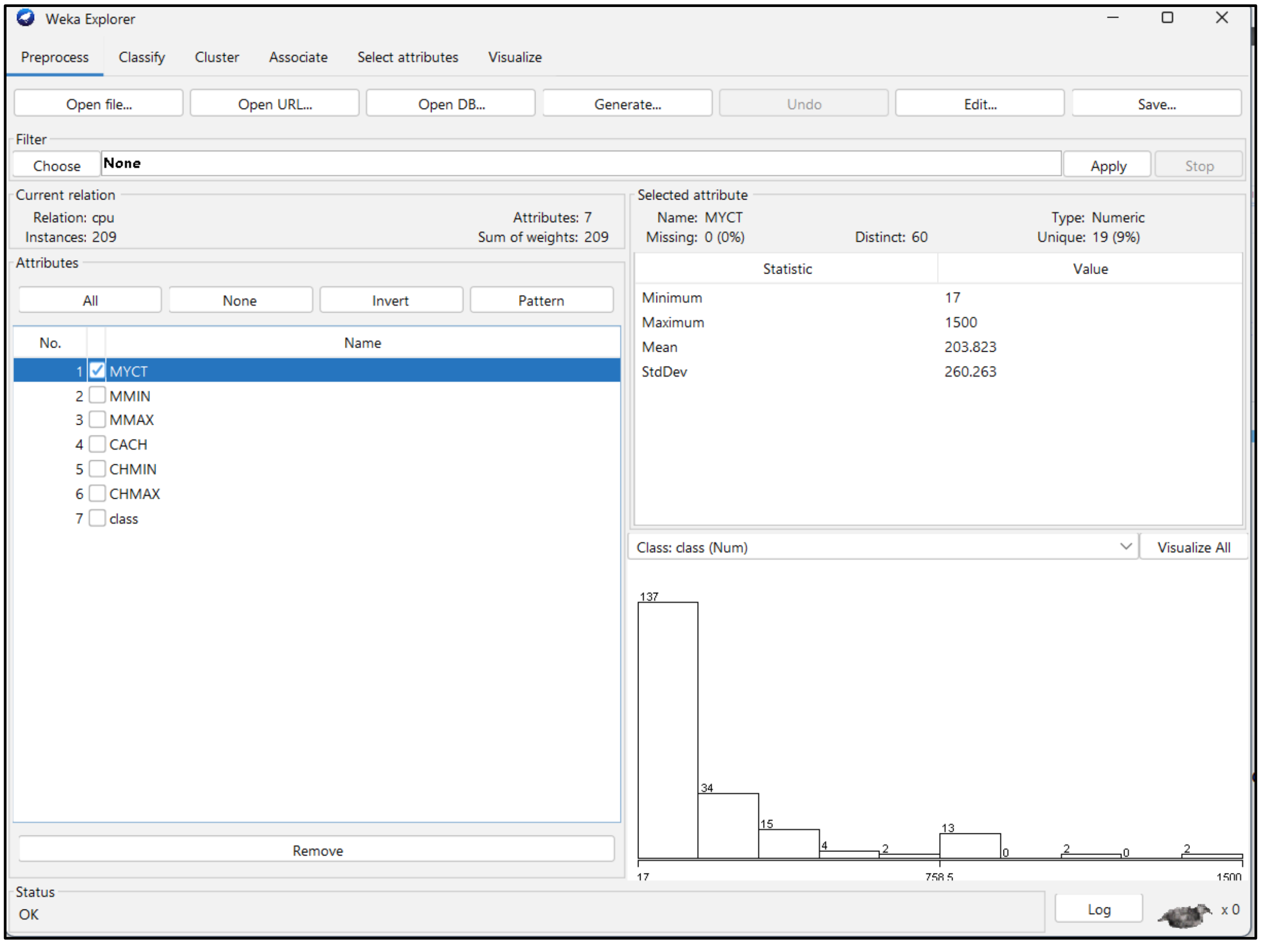
COMMERCIAL BANKS AND ISLAMIC BANKS : CLASSIFICATION OF LOANS / FINANCING BY S...

This dataset shows the Commercial Banks and Islamic Banks : Classification Of Loans / Financing By Sector, 2006 - 2020 Footnote The definition of the economic sectors/industries...

Figure 3.21 The Datasets in The Website

3.2.3 Data Preprocessing

This is where the part where the user needs to understand the data that they are going to use if it is matched to the problem that they are going to solve (Smart Vision Europe, 2020). In this phase, many things need to be done: data cleaning and data reduction. From the previous phase, the user already knows the dataset's information, such as its format type and field, value, the relationship between variables and the quality of each attributes that is related to the model that is going to be implemented. For example, gender and age, the most common data that will be used, must have a relationship in order to detect the drug abuse used by Malaysian. Figure 3.2.2 below shows how the Weka application presents the dataset from the CPU folder, which consists of seven variables, six of which are independent variables, while the last one, which is the class, is the dependent variable. The right tab of the application shows the minimum, maximum, mean and standard deviation values for each of the variables.



Filter: None

Current relation

Relation: cpu Attributes: 7
Instances: 209 Sum of weights: 209

Selected attribute

Name: MYCT Type: Numeric
Missing: 0 (0%) Distinct: 60 Unique: 19 (9%)

Statistic	Value
Minimum	17
Maximum	1500
Mean	203.823
StdDev	260.263

Attributes

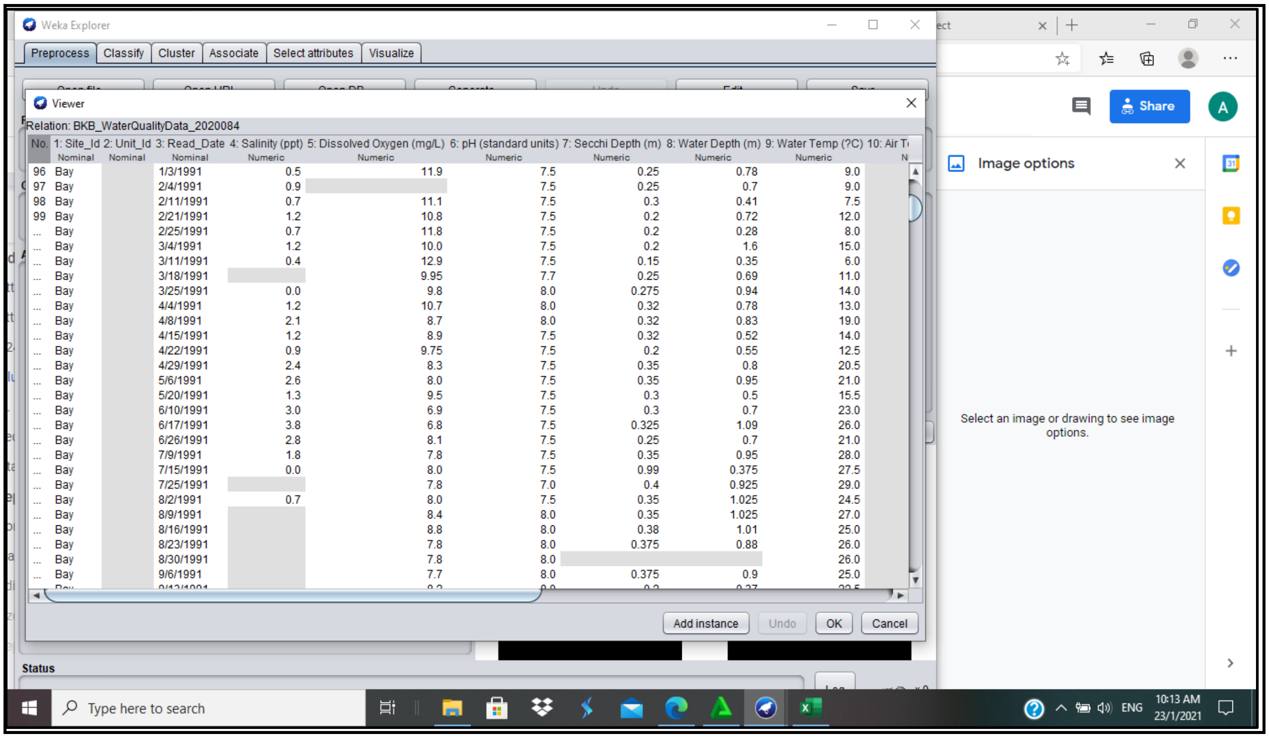
No.		Name
1	☑	MYCT
2	□	MMIN
3	□	MMAX
4	□	CACH
5	□	CHMIN
6	□	CHMAX
7	□	class

Class: class (Num)

Status
OK

Figure 3.2.2 Weka Application of Chosen Dataset of CPU

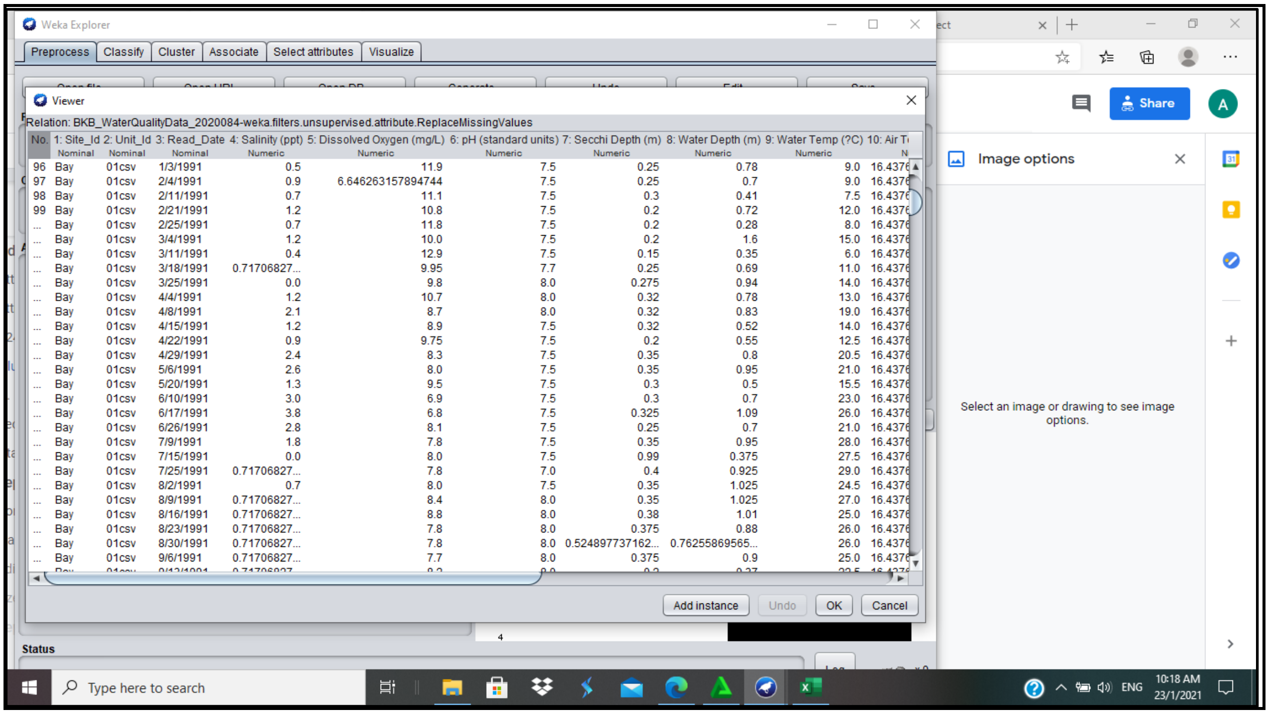
The first thing first is detecting the missing value. This situation is dire if it occurs in the datasets as it can damaging and affect the results of the model later. It can be that during the key in data, the person in charge forgets to enter it or does not know the right value of it, so there goes the blank space in certain datasets. The other common mistake made is the past datasets is corrupted because the person in charge before do not make proper maintenance to the datasets. To handle the missing values, there are two ways which are by ignoring the tuples and the second one is by filling them. Ignoring the tuples in missing values means totally ignoring the missing value and leaving it null if the dataset is too large that the model results later are not affected by it. The second choice is to fill in the missing value with the mean or median value of the attributes that occurred because of the data distribution. Suppose the data distribution of the datasets is symmetric. In that case, it is best to use the mean value, while the median is a better choice if the data distribution is skewed data distribution (Tamboli, 2022). This can be done in Weka by using the Unsupervised Attribute Filter in the ReplaceMissingValues option. This filter will fill all the missing values by means and mode values based on the configuration for both the nominal and numeric attributes. Figure 3.2.3 below shows the example of the missing value that can be found in certain uncompleted datasets, while Figure 3.2.4 shows the example of a filled missing value with the mean value in the datasets.



Relation: BKB_WaterQualityData_2020084

No.	1: Site_Id	2: Unit_Id	3: Read_Date	4: Salinity (ppt)	5: Dissolved Oxygen (mg/L)	6: pH (standard units)	7: Secchi Depth (m)	8: Water Depth (m)	9: Water Temp (?C)	10: Air T
	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	N
96	Bay		1/3/1991	0.5	11.9	7.5	0.25	0.78	9.0	
97	Bay		2/4/1991	0.9		7.5	0.25	0.7	9.0	
98	Bay		2/11/1991	0.7	11.1	7.5	0.3	0.41	7.5	
99	Bay		2/21/1991	1.2	10.8	7.5	0.2	0.72	12.0	
...	Bay		2/25/1991	0.7	11.8	7.5	0.2	0.28	8.0	
...	Bay		3/4/1991	1.2	10.0	7.5	0.2	1.6	15.0	
...	Bay		3/11/1991	0.4	12.9	7.5	0.15	0.35	6.0	
...	Bay		3/18/1991		9.95	7.7	0.25	0.69	11.0	
...	Bay		3/25/1991	0.0	9.8	8.0	0.275	0.94	14.0	
...	Bay		4/4/1991	1.2	10.7	8.0	0.32	0.78	13.0	
...	Bay		4/8/1991	2.1	8.7	8.0	0.32	0.83	19.0	
...	Bay		4/15/1991	1.2	8.9	7.5	0.32	0.52	14.0	
...	Bay		4/22/1991	0.9	9.75	7.5	0.2	0.55	12.5	
...	Bay		4/29/1991	2.4	8.3	7.5	0.35	0.8	20.5	
...	Bay		5/6/1991	2.6	8.0	7.5	0.35	0.95	21.0	
...	Bay		5/20/1991	1.3	9.5	7.5	0.3	0.5	15.5	
...	Bay		6/10/1991	3.0	6.9	7.5	0.3	0.7	23.0	
...	Bay		6/17/1991	3.8	6.8	7.5	0.325	1.09	26.0	
...	Bay		6/26/1991	2.8	8.1	7.5	0.25	0.7	21.0	
...	Bay		7/9/1991	1.8	7.8	7.5	0.35	0.95	28.0	
...	Bay		7/15/1991	0.0	8.0	7.5	0.99	0.375	27.5	
...	Bay		7/25/1991		7.8	7.0	0.4	0.925	29.0	
...	Bay		8/2/1991	0.7	8.0	7.5	0.35	1.025	24.5	
...	Bay		8/9/1991		8.4	8.0	0.35	1.025	27.0	
...	Bay		8/16/1991		8.8	8.0	0.38	1.01	25.0	
...	Bay		8/23/1991		7.8	8.0	0.375	0.88	26.0	
...	Bay		8/30/1991		7.8	8.0			26.0	
...	Bay		9/6/1991		7.7	8.0	0.375	0.9	25.0	

Figure 3.2.3 Missing Value in Dataset

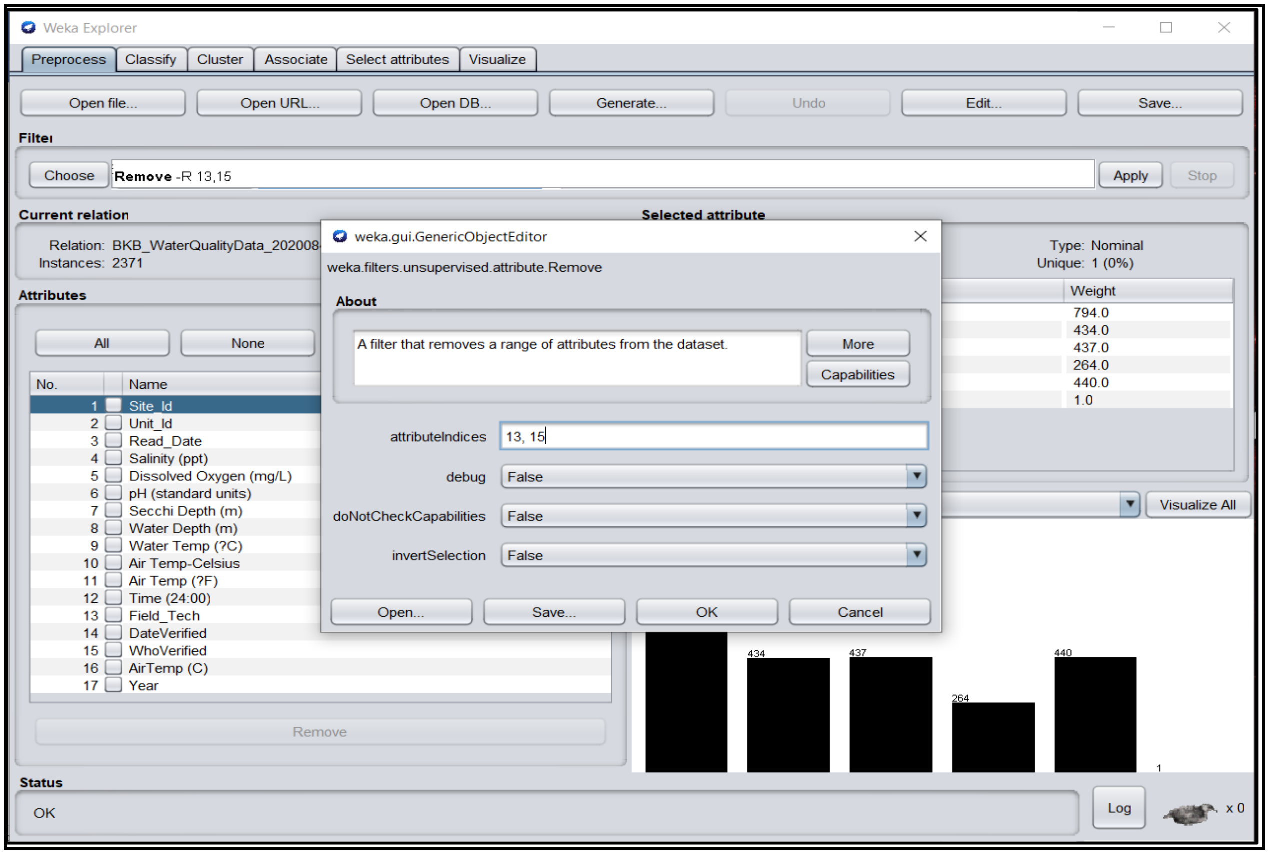


Relation: BKB_WaterQualityData_2020084-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: Site_Id	2: Unit_Id	3: Read_Date	4: Salinity (ppt)	5: Dissolved Oxygen (mg/L)	6: pH (standard units)	7: Secchi Depth (m)	8: Water Depth (m)	9: Water Temp (?C)	10: Air T
	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	N
96	Bay	01csv	1/3/1991	0.5	11.9	7.5	0.25	0.78	9.0	16.4376
97	Bay	01csv	2/4/1991	0.9	6.646263157894744	7.5	0.25	0.7	9.0	16.4376
98	Bay	01csv	2/11/1991	0.7	11.1	7.5	0.3	0.41	7.5	16.4376
99	Bay	01csv	2/21/1991	1.2	10.8	7.5	0.2	0.72	12.0	16.4376
...	Bay	01csv	2/25/1991	0.7	11.8	7.5	0.2	0.28	8.0	16.4376
...	Bay	01csv	3/4/1991	1.2	10.0	7.5	0.2	1.6	15.0	16.4376
...	Bay	01csv	3/11/1991	0.4	12.9	7.5	0.15	0.35	6.0	16.4376
...	Bay	01csv	3/18/1991	0.71706827...	9.95	7.7	0.25	0.69	11.0	16.4376
...	Bay	01csv	3/25/1991	0.0	9.8	8.0	0.275	0.94	14.0	16.4376
...	Bay	01csv	4/4/1991	1.2	10.7	8.0	0.32	0.78	13.0	16.4376
...	Bay	01csv	4/8/1991	2.1	8.7	8.0	0.32	0.83	19.0	16.4376
...	Bay	01csv	4/15/1991	1.2	8.9	7.5	0.32	0.52	14.0	16.4376
...	Bay	01csv	4/22/1991	0.9	9.75	7.5	0.2	0.55	12.5	16.4376
...	Bay	01csv	4/29/1991	2.4	8.3	7.5	0.35	0.8	20.5	16.4376
...	Bay	01csv	5/6/1991	2.6	8.0	7.5	0.35	0.95	21.0	16.4376
...	Bay	01csv	5/20/1991	1.3	9.5	7.5	0.3	0.5	15.5	16.4376
...	Bay	01csv	6/10/1991	3.0	6.9	7.5	0.3	0.7	23.0	16.4376
...	Bay	01csv	6/17/1991	3.8	6.8	7.5	0.325	1.09	26.0	16.4376
...	Bay	01csv	6/26/1991	2.8	8.1	7.5	0.25	0.7	21.0	16.4376
...	Bay	01csv	7/9/1991	1.8	7.8	7.5	0.35	0.95	28.0	16.4376
...	Bay	01csv	7/15/1991	0.0	8.0	7.5	0.99	0.375	27.5	16.4376
...	Bay	01csv	7/25/1991	0.71706827...	7.8	7.0	0.4	0.925	29.0	16.4376
...	Bay	01csv	8/2/1991	0.7	8.0	7.5	0.35	1.025	24.5	16.4376
...	Bay	01csv	8/9/1991	0.71706827...	8.4	8.0	0.35	1.025	27.0	16.4376
...	Bay	01csv	8/16/1991	0.71706827...	8.8	8.0	0.38	1.01	25.0	16.4376
...	Bay	01csv	8/23/1991	0.71706827...	7.8	8.0	0.375	0.88	26.0	16.4376
...	Bay	01csv	8/30/1991	0.71706827...	7.8	8.0	0.524897737162...	0.76255869565...	26.0	16.4376
...	Bay	01csv	9/6/1991	0.71706827...	7.7	8.0	0.375	0.9	25.0	16.4376

Figure 3.2.4 Missing Value Filled in Dataset

The next one is the noisy data. Noisy data here means meaningless data that is not related to the machine learning model to be computed later than found in the datasets. This can be seen either in class or in attributes. For example, in class, noise can be contradictory and also the mislabeled examples, while in the attributes, it can be missing value and also the error value that is outside of the expectation value. These class and attribute values have highly affected the quality of datasets because they can make the classifier performance become worse than usual performance, especially in classification problems. Removing the attributes is the best and easiest way to do the class and attributes. In Weka, there is an option to remove the unnecessary noisy data that might interrupt the model evaluation later. Be cautious, as removing the important attributes can affect the evaluation phase later while presenting the knowledge to the audience.



Filter

Remove -R 13,15

Current relation

Relation: BKB_WaterQualityData_202008

Instances: 2371

Selected attribute

Type: Nominal

Unique: 1 (0%)

Weight
794.0
434.0
437.0
264.0
440.0
1.0

Attributes

No.	Name
1	☐ Site_Id
2	☐ Unit_Id
3	☐ Read_Date
4	☐ Salinity (ppt)
5	☐ Dissolved Oxygen (mg/L)
6	☐ pH (standard units)
7	☐ Secchi Depth (m)
8	☐ Water Depth (m)
9	☐ Water Temp (?C)
10	☐ Air Temp-Celsius
11	☐ Air Temp (?F)
12	☐ Time (24:00)
13	☐ Field_Tech
14	☐ DateVerified
15	☐ WhoVerified
16	☐ AirTemp (C)
17	☐ Year

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Remove

About

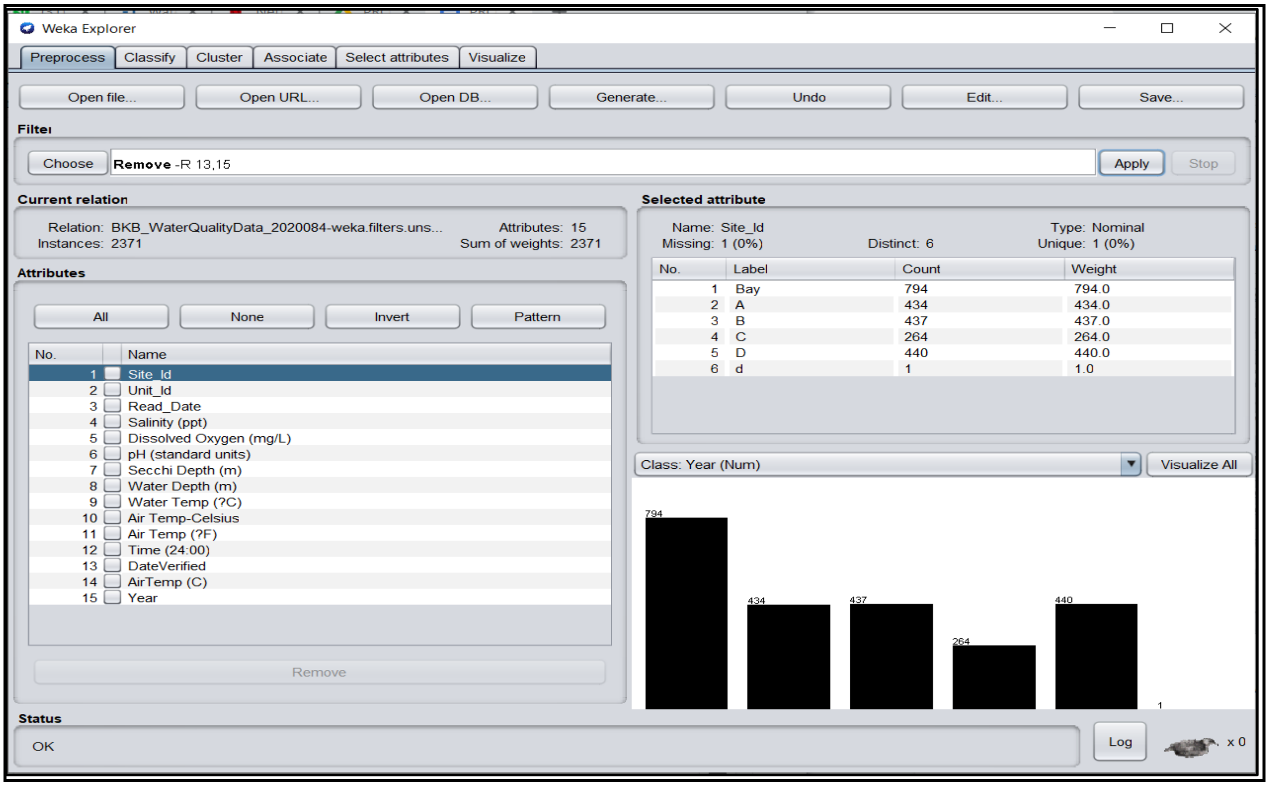
A filter that removes a range of attributes from the dataset.

attributeIndices	13, 15
debug	False
doNotCheckCapabilities	False
invertSelection	False

Status

OK

Figure 3.2.5 Removing Attributes in Weka



Filter

Remove -R 13,15

Current relation

Relation: BKB_WaterQualityData_2020084-weka.filters.uns... Attributes: 15
Instances: 2371 Sum of weights: 2371

Attributes

No.	Name
1	Site_Id
2	Unit_Id
3	Read_Date
4	Salinity (ppt)
5	Dissolved Oxygen (mg/L)
6	pH (standard units)
7	Secchi Depth (m)
8	Water Depth (m)
9	Water Temp (?C)
10	Air Temp-Celsius
11	Air Temp (?F)
12	Time (24:00)
13	DateVerified
14	AirTemp (C)
15	Year

Selected attribute

Name: Site_Id Type: Nominal
Missing: 1 (0%) Distinct: 6 Unique: 1 (0%)

No.	Label	Count	Weight
1	Bay	794	794.0
2	A	434	434.0
3	B	437	437.0
4	C	264	264.0
5	D	440	440.0
6	d	1	1.0

Class: Year (Num)

Status

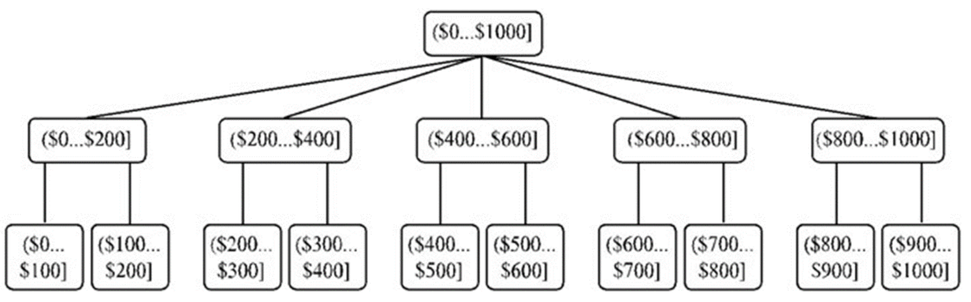
OK

Figure 3.2.6 Removing Attributes Results in Weka

3.2.4 Data Transformation

Data transformation is the next phase after the data preprocessing. As the name speaks for itself, data transformation is the process of transforming unstructured preprocessed datasets into new kinds of structured datasets to improve their reliability for later analysis (Karan, 2021). Data transformation includes the changes of type, value or structure of the data and transforming it into a clean and reliable set of datasets. There are many data transformation techniques, such as data discretization, data generalization and data normalization. These techniques can be used in any order and anytime when it is necessary to can improves the quality and reliability of the datasets, such as discretize only one set of attributes while generalize the others according to the needs of the learning models.

The first data transformation technique is data discretization. Data discretization is the process of converting continuous data into several groups by dividing it according to a certain degree of value (Tim, 2022). The grouping itself is sorted in order and discrete value. Discretization is not only limited to numeric, as it also can be used on nominal instances. Discretization not only groups the instances, but it also can change the type of instances. For example, discretization can be used to change the attributes from numeric to nominal. In Weka, there are many options available to do discretization methods such as the number of groups wanted, which value to be group and etc.



A concept hierarchy for the attribute *price*, where an interval ($*X* . . . $*Y*] denotes the range from $*X* (exclusive) to $*Y* (inclusive).

Figure 3.2.7 Discretization Illustration

Next is data generalization. Data generalization is a process too. Data generalization is good to use when the distribution of the datasets is known or when it is a Gaussian distribution attribute. The huge differences between these attributes can affect the data model later, can cause the result of the data mining to be poor, and leaves a huge error such as the Mean Absolute Error (MEA) and the Root Mean Squared Error (RMSE) (GeeksforGeeks, 2019). Data generalization is good to use when the distribution of the datasets is unknown or when it is not Gaussian (a bell curve). However, in Weka software, there is no data generalization technique, so this step cannot be done in Weka.

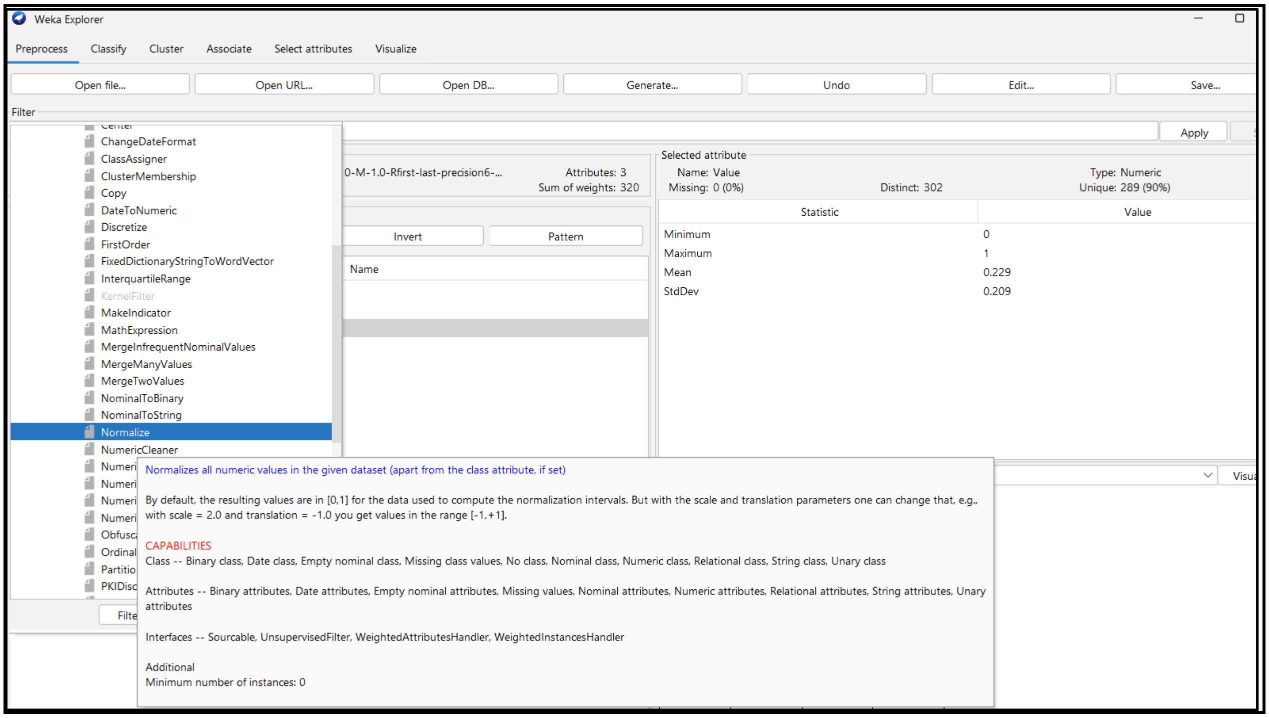
The last one is data normalization. Data normalization is used to make the instances of an attribute be scaled within a small value, such as from -1.0 to 1.0, for better data workflows. For example, the value of a salary in a year compared to the year of experience. When comparing these two attributes, it has a huge cap difference because of the amount of salary compared to the year of experience. Hence, normalization is required when dealing with this kind of huge scaling value. The MEA and RMSE also can be reduced if the preprocessed datasets apply the data normalization technique. Data normalization is also good to use when the distribution of the datasets is unknown or when it is not Gaussian (a bell curve). In Weka, the data normalization filter can use two techniques, which are the Standardize filter, which applies the Z-score normalization and the Normalize filter, that is applied the Min-max normalization. The calculation of Z-score normalization is based on how much a data point stray from the mean value of the datasets, while the Min-max normalization applies the linear transformation on the preprocessed datasets that preserves the relation of it (Tolety, 2023).

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

Figure 3.2.8 The Formula of Z-Score Normalization

$$\mathbf{v}' = \frac{\mathbf{v} - \mathbf{min(A)}}{\mathbf{max(A)} - \mathbf{min(A)}}(\mathbf{new\_max(A)} - \mathbf{new\_min(A)}) + \mathbf{new\_min(A)}$$

Figure 3.2.9 The Formula of Min-Max Normalization



0-M-1.0-Rfirst-last-precision6-... Attributes: 3
Sum of weights: 320

Selected attribute
Name: Value Type: Numeric
Missing: 0 (0%) Distinct: 302 Unique: 289 (90%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.229
StdDev	0.209

Normalizes all numeric values in the given dataset (apart from the class attribute, if set)

By default, the resulting values are in [0,1] for the data used to compute the normalization intervals. But with the scale and translation parameters one can change that, e.g., with scale = 2.0 and translation = -1.0 you get values in the range [-1,+1].

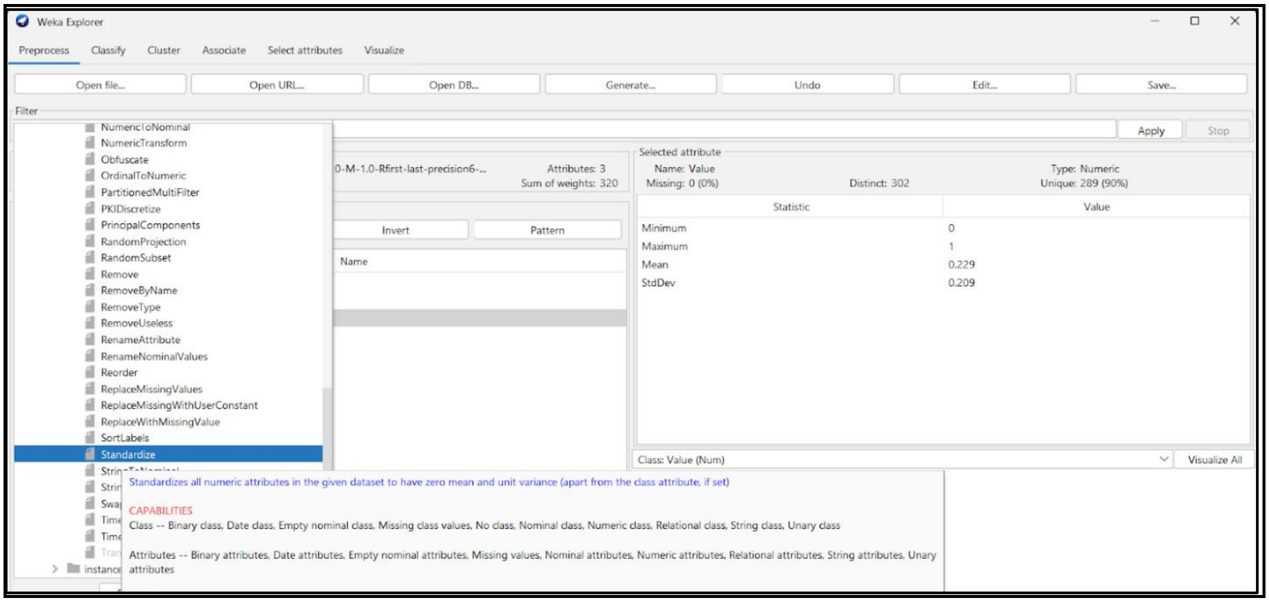
CAPABILITIES
Class -- Binary class, Date class, Empty nominal class, Missing class values, No class, Nominal class, Numeric class, Relational class, String class, Unary class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

Interfaces -- Sourcable, UnsupervisedFilter, WeightedAttributesHandler, WeightedInstancesHandler

Additional
Minimum number of instances: 0

Figure 3.2.10 Normalization Filter in Weka



0-M-1.0-Rfirst-last-precision6-... Attributes: 3
Sum of weights: 320

Selected attribute
Name: Value Type: Numeric
Missing: 0 (0%) Distinct: 302 Unique: 289 (90%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.229
StdDev	0.209

Class: Value (Num)

Standardizes all numeric attributes in the given dataset to have zero mean and unit variance (apart from the class attribute, if set)

CAPABILITIES
Class -- Binary class, Date class, Empty nominal class, Missing class values, No class, Nominal class, Numeric class, Relational class, String class, Unary class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

Figure 3.2.11 Standardize Filter in Weka

3.2.5 Modelling

The fourth phase is the fun part, which is the data mining phase. This is where the researcher needed to select a model that will be used which is appropriate to the requirement and objectives needed by solving the problem statements. There are so many models available to be chosen, and each of the models keeps improving day by day. The models chosen need to produce a good result that relates to the correlation between variables to mark it as a good model chosen. Why? This is because, when applying an irrelevant model to the preprocessed datasets, the results can be bad as it has a lower correlation coefficient and lower corrected prediction value. This led to the preprocess problem that needed the data to be preprocessed over and over again. That is why in Chapter 2, there is a comparison of data models being made, and the model chosen for this thesis is the Regression for prediction of value regardless of continuous or discrete value. For this research, the proper model that can be used are the Linear Regressions and also Random Tree. Linear Regression is one of the popular Data Mining Algorithms used to solve Regression problems, predicting values if the data have many numerical with some nominal attributes, which is the same as in this thesis that gives the probabilistic values while Random Tree is great for getting the non-linear relationship between input and output variables.

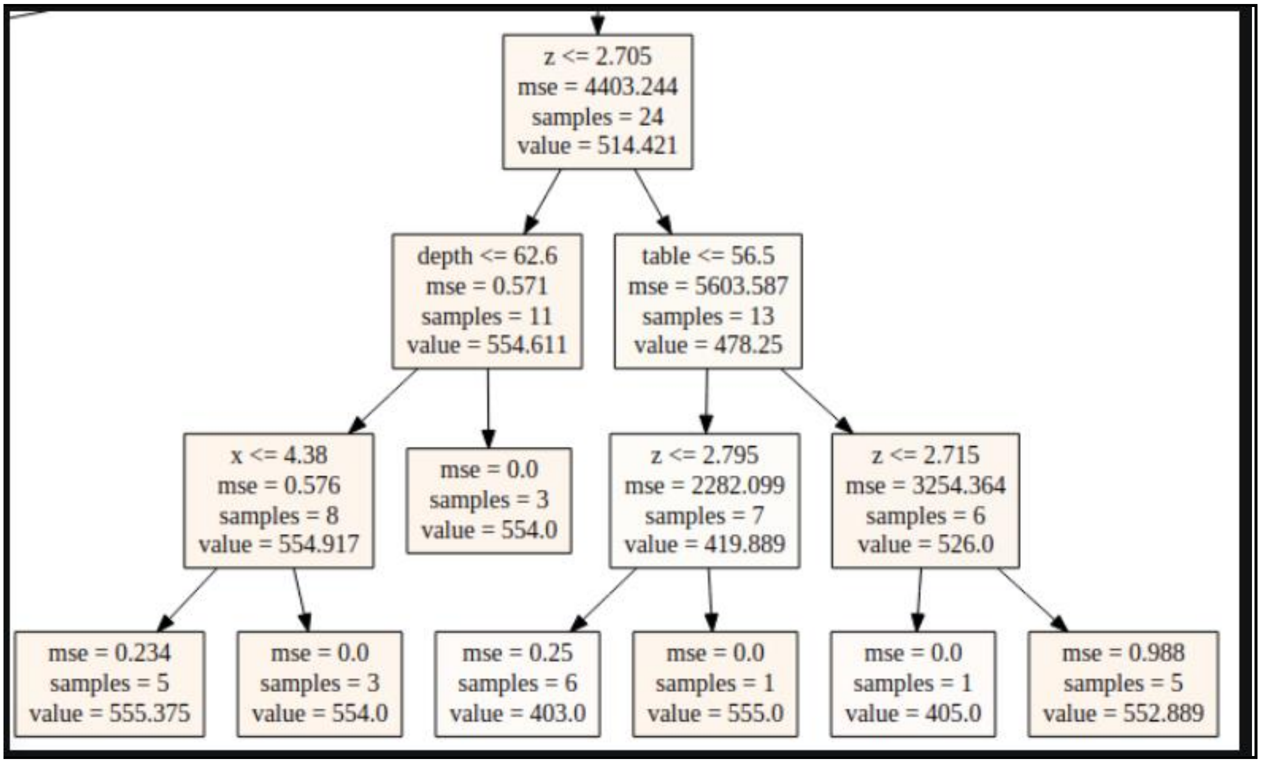


Figure 3.2.12 Random Tree Model in Data Mining

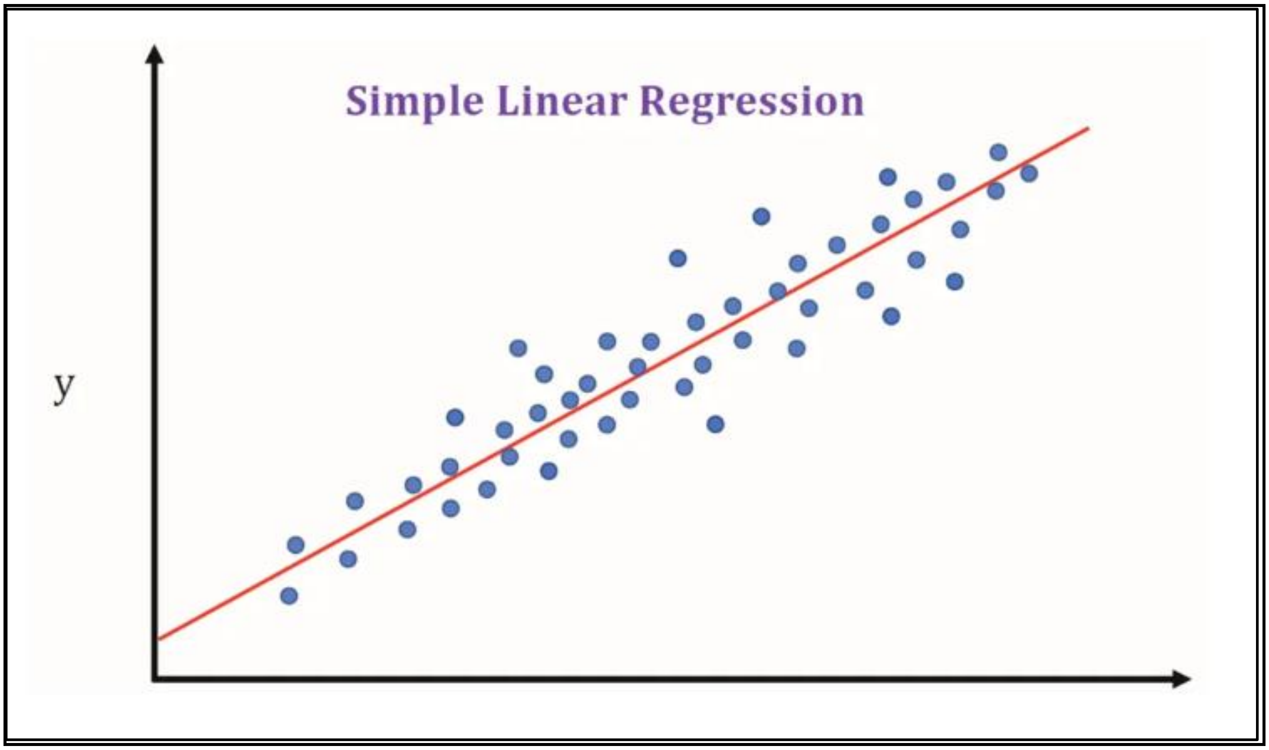


Figure 3.2.13 Simple Linear Regression Model in Data Mining

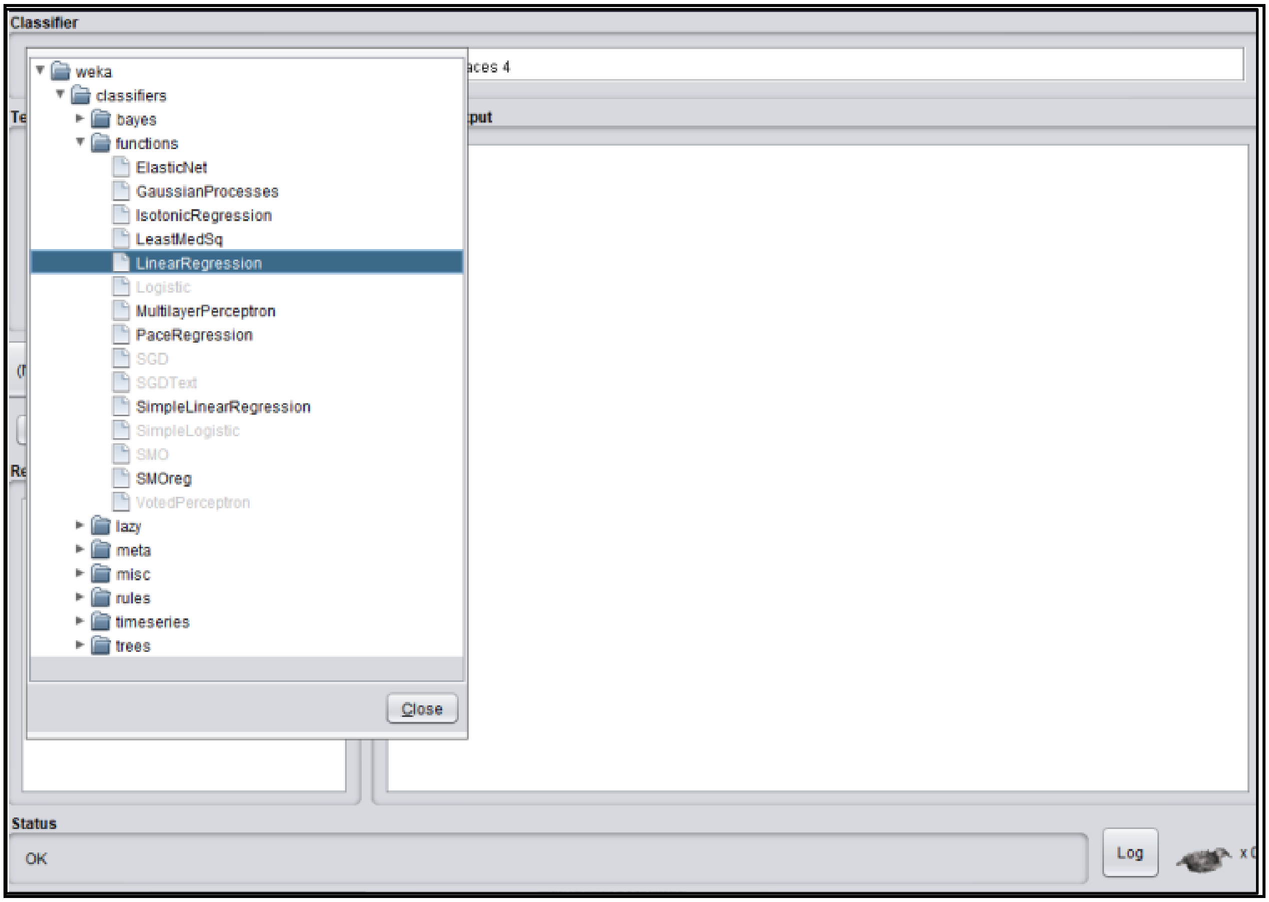
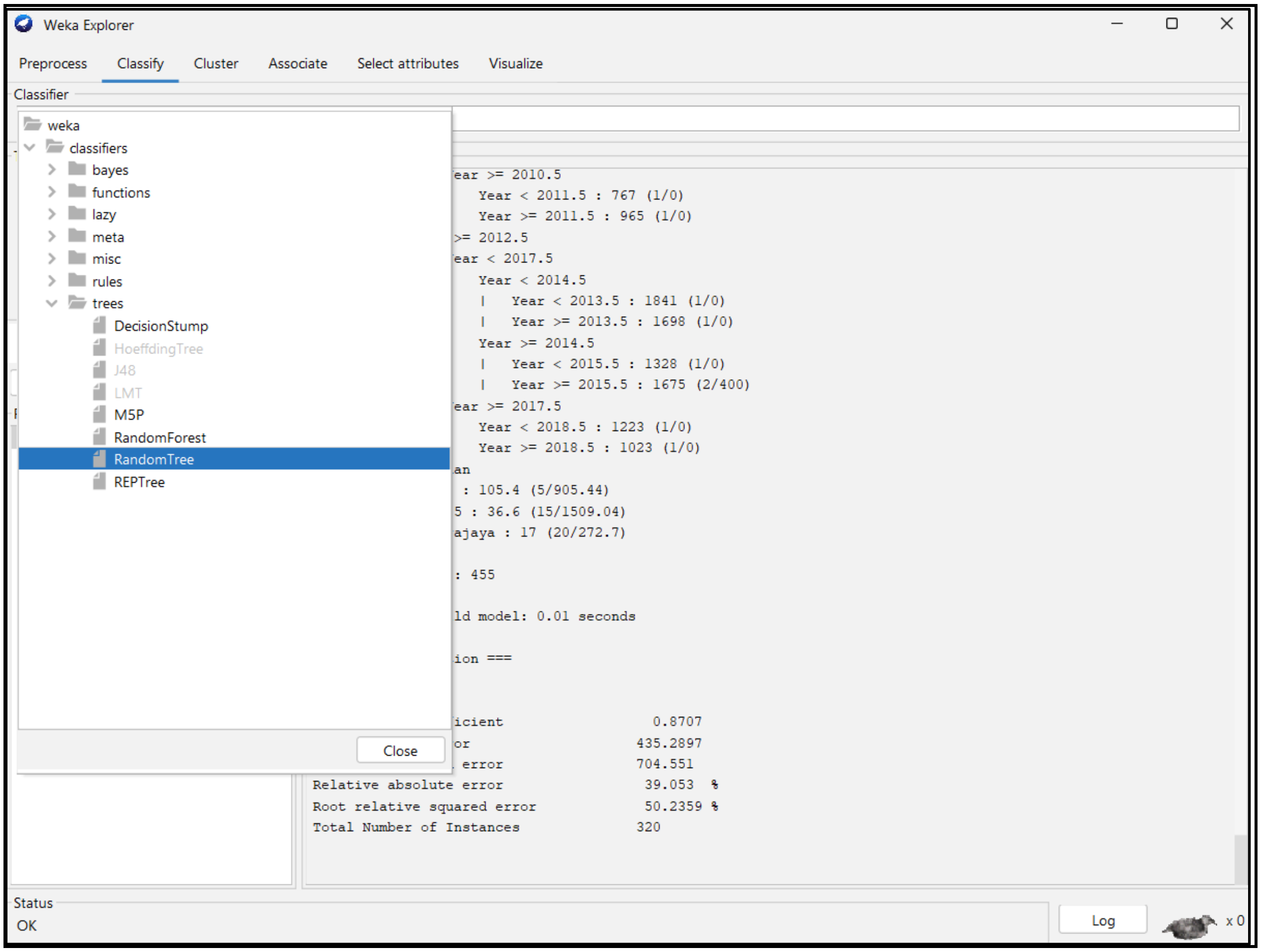


Figure 3.2.14 Choosing Linear Regression Model in Weka Application



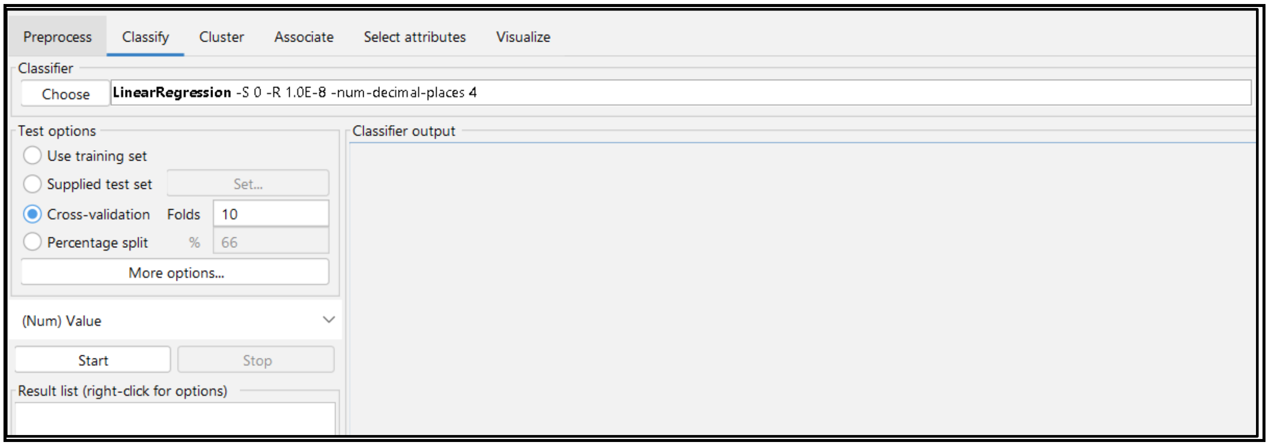
```
                 ear >= 2010.5
                    Year < 2011.5 : 767 (1/0)
                    Year >= 2011.5 : 965 (1/0)
                 >= 2012.5
                 ear < 2017.5
                    Year < 2014.5
                    |   Year < 2013.5 : 1841 (1/0)
                    |   Year >= 2013.5 : 1698 (1/0)
                    Year >= 2014.5
                    |   Year < 2015.5 : 1328 (1/0)
                    |   Year >= 2015.5 : 1675 (2/400)
                 ear >= 2017.5
                    Year < 2018.5 : 1223 (1/0)
                    Year >= 2018.5 : 1023 (1/0)
                 an
                  : 105.4 (5/905.44)
                 5 : 36.6 (15/1509.04)
                 ajaya : 17 (20/272.7)

                 : 455

                 ld model: 0.01 seconds

                 ion ===


                 icient                   0.8707
                 or                     435.2897
                  error                 704.551
Relative absolute error                  39.053  %
Root relative squared error              50.2359 %
Total Number of Instances               320
```

Figure 3.2.15   Choosing Random Tree Model in Weka Application

3.2.6 Evaluation

The evaluation process is about getting the knowledge from the model that has been applied. If it is against the requirement and expectation of the results, then it is advisable for the data analysis to go back to the first phase, which is the data selection phase. When doing the evaluation phase, the user will review the entire model and its results to detect any wrong that occurred during the data mining process. Therefore, when the evaluation goes wrong, the user needs to have a strong understanding of how the entire process happened from the first until the data modelling phase so that the result meets the main objectives of the research.

In the Weka application, the user can set any regression models wanted in the Classify tab by simply clicking on it and choosing the wanted algorithm. Next, in the test option, choose the training that is wanted. Let's pick on cross-validation and make it ten folds. This means by ten folds cross-validation is split into ten segments and then using nine segments in order to build the model and apply the model's algorithm prediction on the left-out segment. These ten folds mean that ten iterations will use nine different segments used to build the training dataset model while one segment is used as the test dataset. Another test option that will be used is the percentage split. This means randomly splitting the processed dataset into training and testing segments for each time Weka calculated the model used. This test option is good for getting a quick idea of the performance of the model. The default test option for data mining in Weka is cross-validation (Jason Brownlee, 2019). The percentage split is used in order to compare the accuracy of the model with the cross-validation test.



Classifier

LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

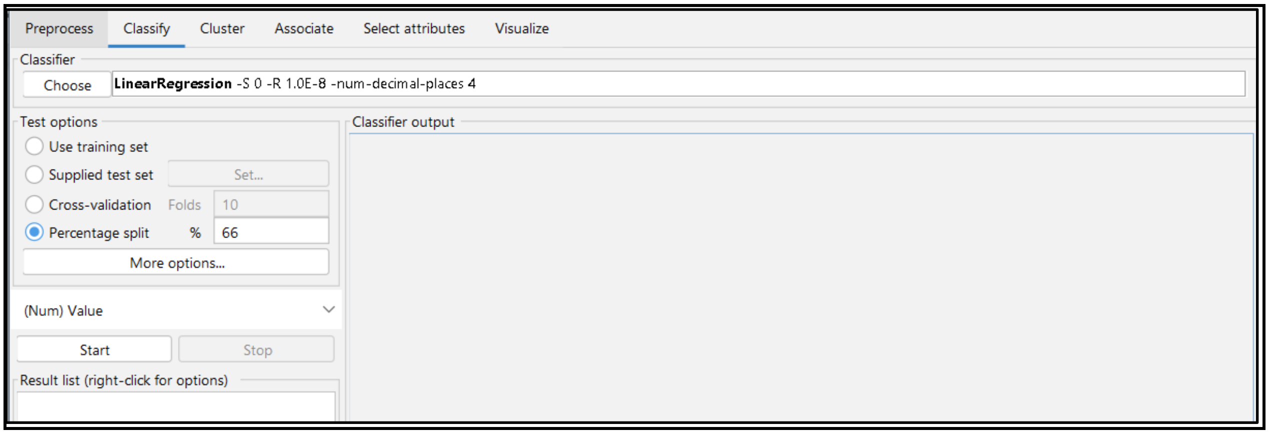
○ Percentage split % 66

(Num) Value

Result list (right-click for options)

Classifier output

Figure 3.2.16 Choosing Cross-Validation Test Option in Weka Application



Classifier

LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Num) Value

Result list (right-click for options)

Classifier output

Figure 3.2.17 Choosing Percentage Split Test Option in Weka Application

For the deployment of the results, this thesis will integrate the data mining discoveries from the start until the evaluation process into the main purpose or outcome wanted, which is to solve the regression problem. The visualization that will be used is from the Power BI application to give the audience a brief but impactful data visualization. There are many options on how the user wants the application to display the result, but for this thesis, it is easier to use the Power BI software as it can integrate with many types of files. The visualization is also neat and cheerful, which can be eye-catching to those who will view it.

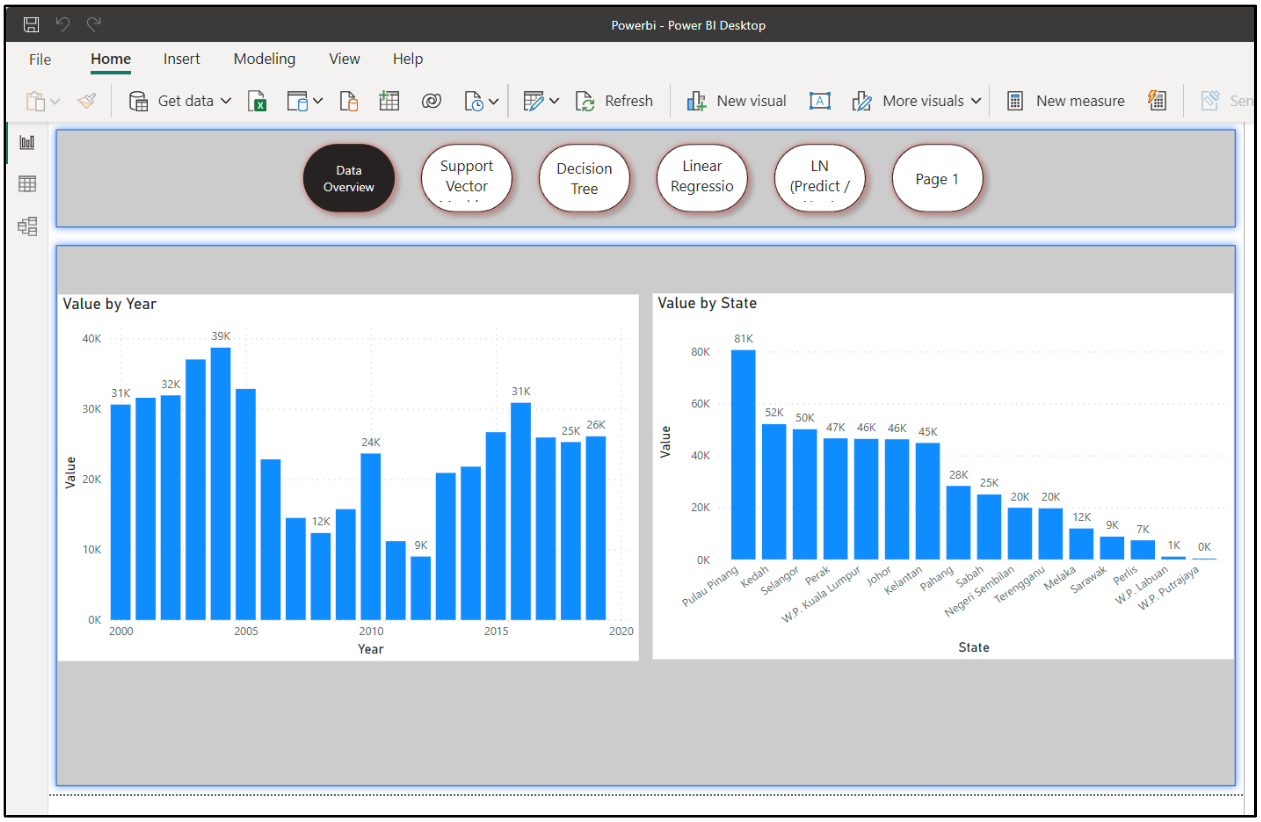


Figure 3.2.18 The Dashboard of the Power BI Desktop

3.3 System Proposed Design

The flow of the study must be relevant to the framework used since it can save time and cost in finishing this study. First, by preparing the dataset, variables, methods, and algorithms will be used. In the Weka application, select the prepared dataset and variables. Make sure to think twice about whether it is suitable for the study, and if not, it is advisable to find another for the best result. Next, choose an algorithm to process the data. This algorithm also needs to be observed and tested correctly. If the algorithm used is not suitable, turn back to choosing the algorithm that suits the case study, which is getting a prediction output. After doing the process of the dataset using a certain dataset, it will come out in two decisions, one is satisfying and vice versa. The study will go back into preparing all the kinds of stuff needed if the result is not satisfactory, regardless of the reasons. If the results are satisfying, the data will be integrated into the Power BI software to do the visualization process.

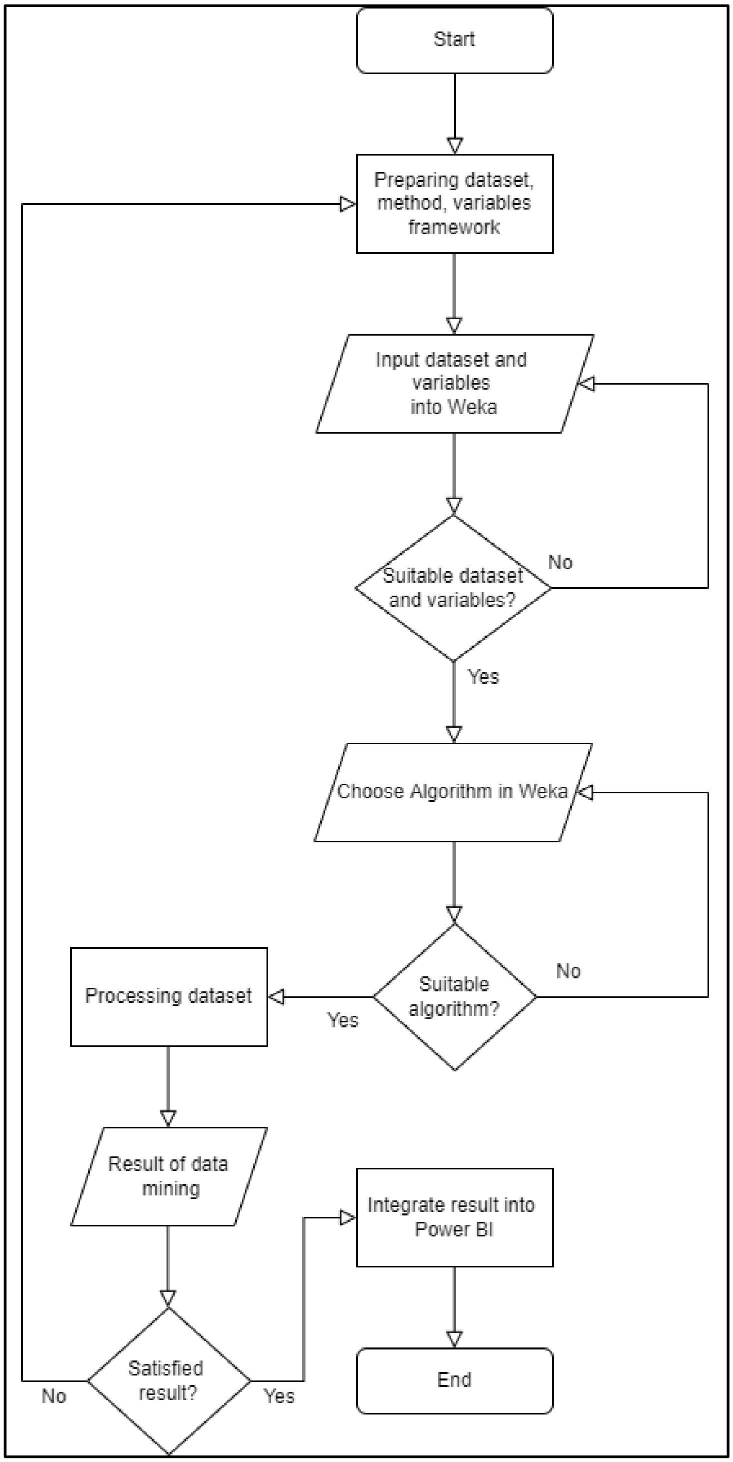

Start
Preparing dataset, method, variables framework
Input dataset and variables into Weka
Suitable dataset and variables?
No
Yes
Choose Algorithm in Weka
Suitable algorithm?
No
Yes
Processing dataset
Result of data mining
Integrate result into Power BI
Satisfied result?
No
Yes
End

Figure 3.3 Flowchart of The Proposed System

3.4 Database Design

The datasets that will be used in this study are very important as they can impact the entire calculation and thus can make the evaluation process error. So, when choosing from the beginning, which is the raw dataset, until getting the transformed and putting a suitable algorithm to get the pattern, all must be carefully picked, and a lot of tries and errors will be made during that time. The basic dataset that must have in the study is the State, Year and Value of the drug abuser. In the incoming Chapter 4, these datasets will be mostly changed according to the needs of the algorithm to get the most accuracy of the prediction on drug abuse.

Data Name	Data Type	Data Description
State	Nominal Data	The state in Malaysia
Year	Numerical Data	The year of the subject
Value	Numerical Data	The value of the drug abuser

Table 3.2 The Datasets Will Be Used

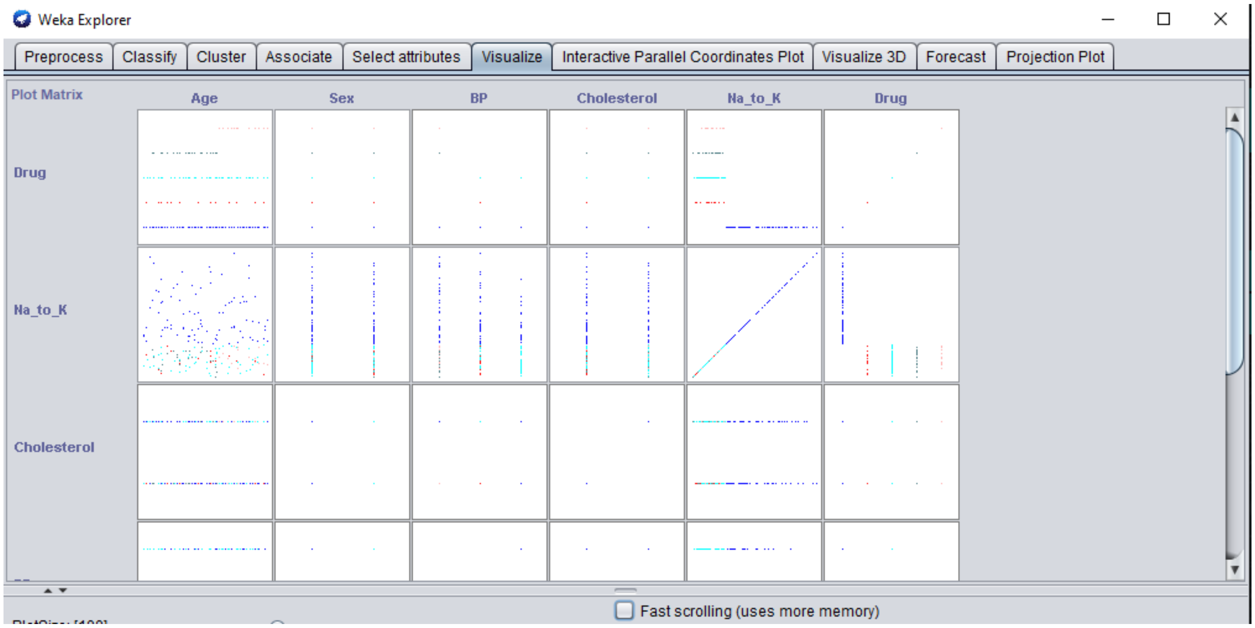
3.5 Hardware & Software Used

The study will use the hardware and software in the process of this are Computer, Microsoft Word, Power BI and lastly, Weka. Below is the important list of hardware and software used and their purposes for completing the process of this study. This software will be integrated with each other to produce an organized data visualization for the readers..

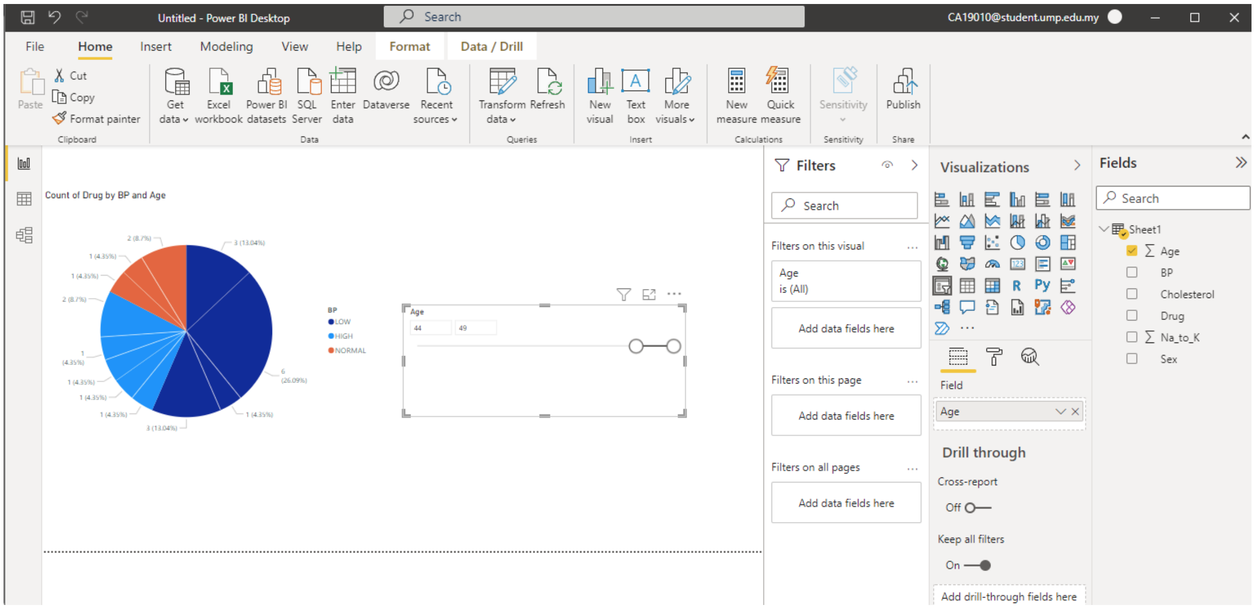
Hardware	
Items	Purpose
Computer	Main tools to do research, develop and visualize the drug prediction using data mining.
Software	
Items	Purpose
Power BI	An application used to connect from Weka datasets and visualize in neat and organized way for data insight.
Weka	An application used to do data mining task using the chosen supervised machine learning algorithms
Excel	An application used to store data and integrate later with Power BI

Table 3.3 The Hardware and Software Used

3.6 Proof of Concept / Prototype



Ratio of graph according to variables



Pie Chart and Slicer Visualization in Power BI

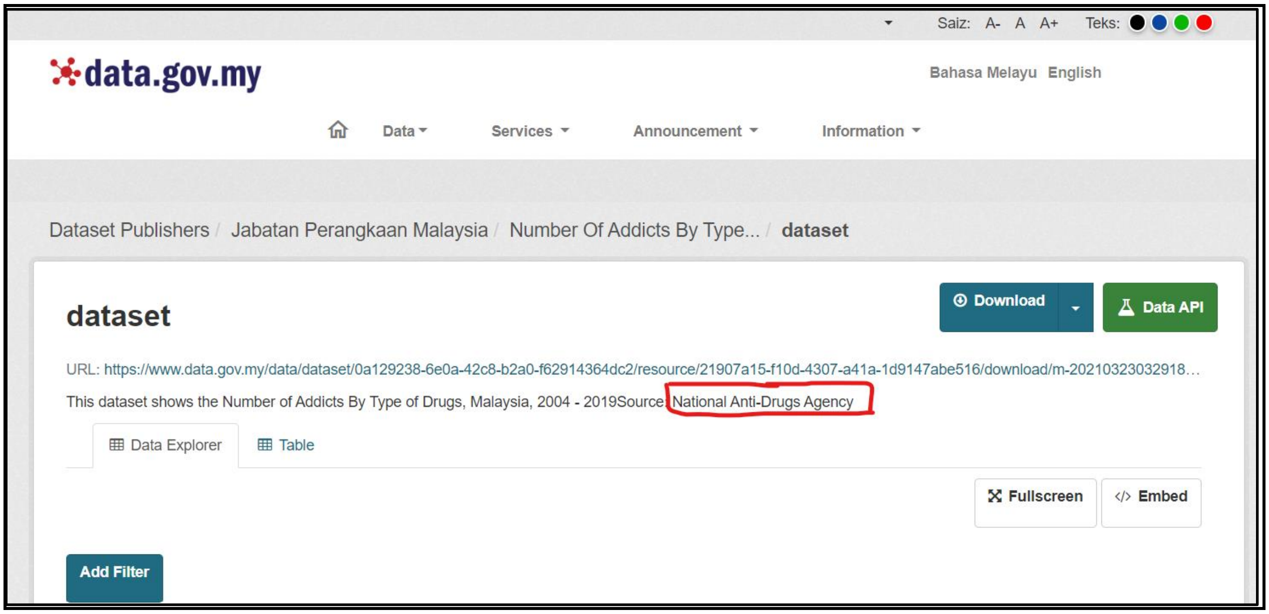
CHAPTER 4

IMPLEMENTATION, RESULT AND DISCUSSION

4.1 Introduction

The dataset for this project needs to be carefully examined and studied first before being through the preprocessing. There are various resources for collecting the dataset, such as the website, government agencies, and the private sector. For this project, the dataset will be collected through The Malaysian Administrative Modernisation and Management Planning Unit (MAMPU), which is the government agency. The reason is because of the authenticity of the dataset by the Malaysian Government itself, which is based on the AADK database. There are a lot of datasets out there on the website that provides various attributes such as nominal, binary, and even original attributes with a lot of instances in each attribute, but to get the datasets that meet this project job scope which are:

- Must be within Malaysia’s scope
- The duration of datasets is up to date of COVID-19 duration included its previous
- Must be an authentic and verified data



Dataset Publishers / Jabatan Perangkaan Malaysia / Number Of Addicts By Type... / dataset

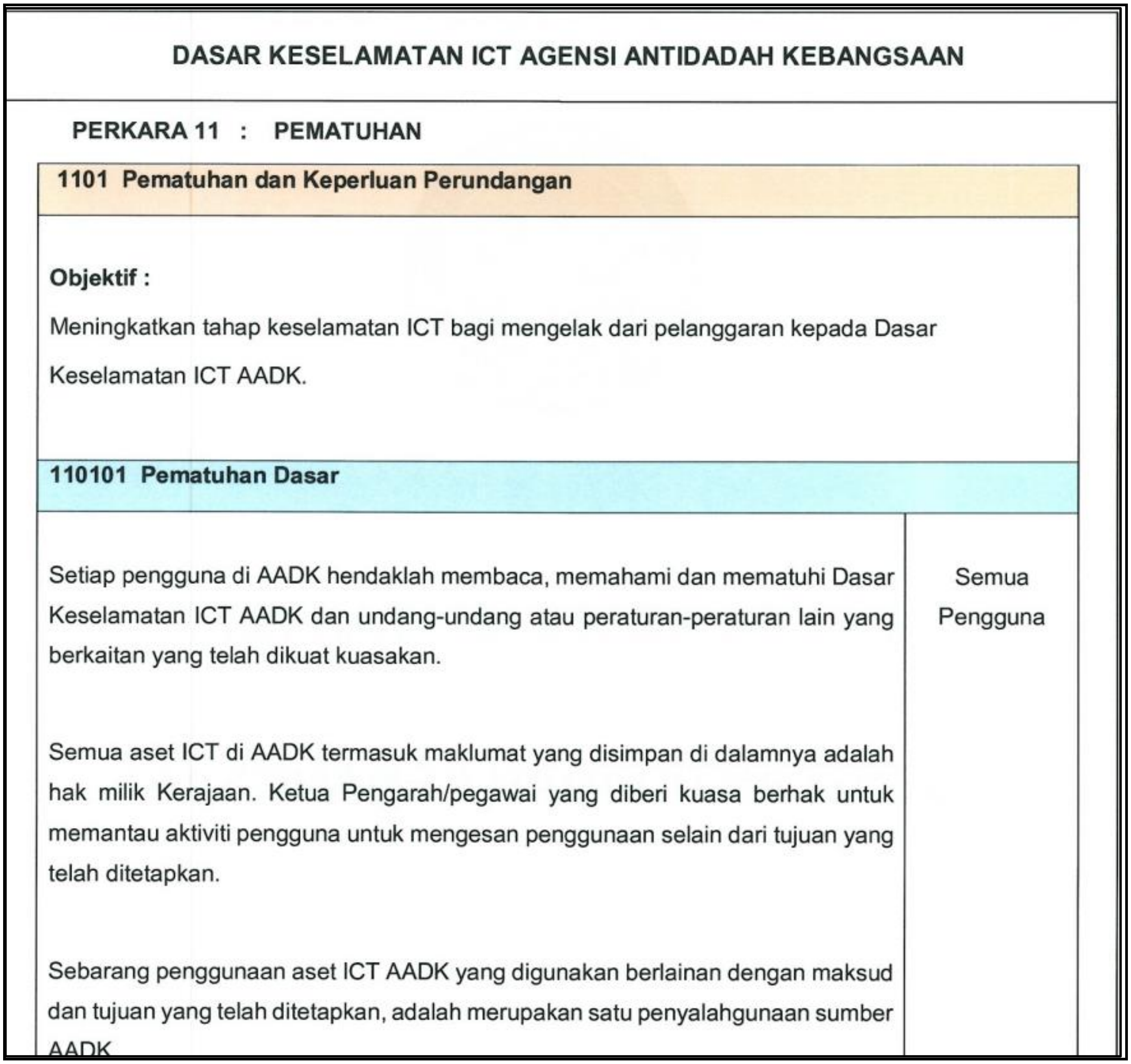
dataset

URL: https://www.data.gov.my/data/dataset/0a129238-6e0a-42c8-b2a0-f62914364dc2/resource/21907a15-f10d-4307-a41a-1d9147abe516/download/m-20210323032918…

This dataset shows the Number of Addicts By Type of Drugs, Malaysia, 2004 - 2019Source: National Anti-Drugs Agency

Figure 4.1 Data Collection On MAMPU's Website That Sourced From The AADK's website

The dataset also can be used in this project because it has obtained consent from AADK terms and conditions about the datasets given, which are for education purposes in Malaysia with the reference of Perkara 11: Pematuhan/110101 of Dasar Keselamatan ICT Versi 5 AADK. It is stated that the user must follow all the rules for accessing and using the data that is related to AADK. If the user happens to breach the rules, the AADK has the right to monitor and take action based on the best way to handle the situation.



DASAR KESELAMATAN ICT AGENSI ANTIDADAH KEBANGSAAN

PERKARA 11 : PEMATUHAN

1101 Pematuhan dan Keperluan Perundangan

Objektif :

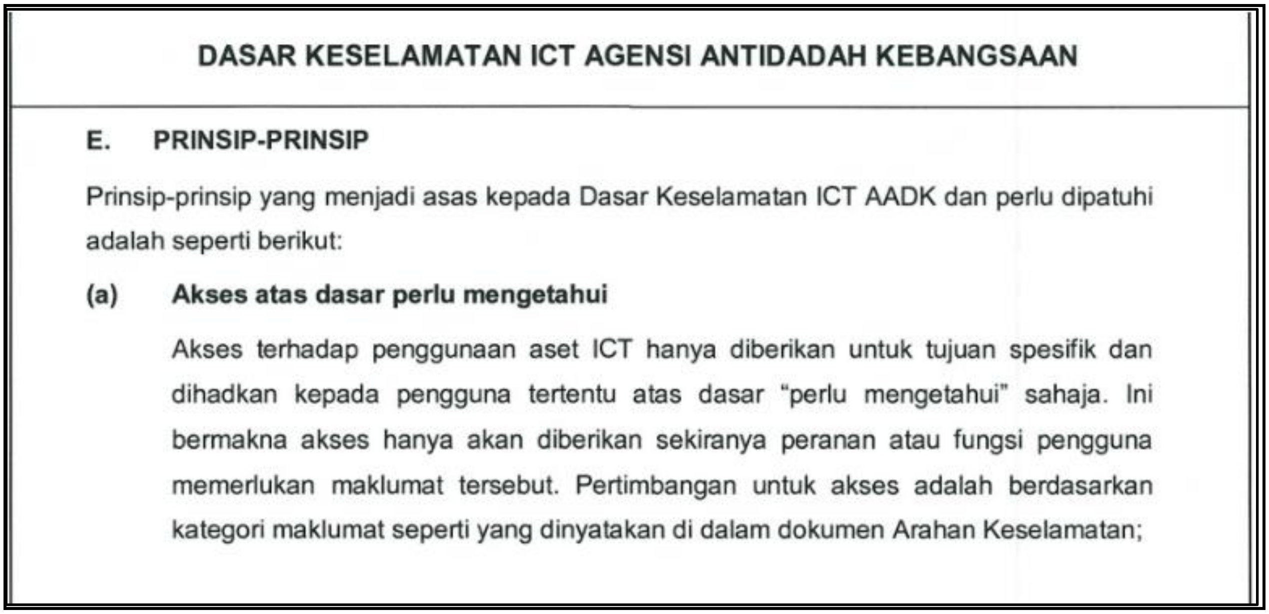
Meningkatkan tahap keselamatan ICT bagi mengelak dari pelanggaran kepada Dasar Keselamatan ICT AADK.

110101 Pematuhan Dasar

Setiap pengguna di AADK hendaklah membaca, memahami dan mematuhi Dasar Keselamatan ICT AADK dan undang-undang atau peraturan-peraturan lain yang berkaitan yang telah dikuat kuasakan. Semua aset ICT di AADK termasuk maklumat yang disimpan di dalamnya adalah hak milik Kerajaan. Ketua Pengarah/pegawai yang diberi kuasa berhak untuk memantau aktiviti pengguna untuk mengesan penggunaan selain dari tujuan yang telah ditetapkan. Sebarang penggunaan aset ICT AADK yang digunakan berlainan dengan maksud dan tujuan yang telah ditetapkan, adalah merupakan satu penyalahgunaan sumber AADK	Semua Pengguna

Figure 4.1.2 The Regulation of The Data In AADK

According to one of the principles of the policies stated in the statement “(a)" and “(b)" that says, the access to the dataset is for the purposes of "need to know" only. This means that any other intentions of using the dataset are forbidden. Any other means of using the data may require the user to get approval from the authority. So, this research is within the rules and can proceed with this dataset as the data for this project.



DASAR KESELAMATAN ICT AGENSI ANTIDADAH KEBANGSAAN

E. PRINSIP-PRINSIP

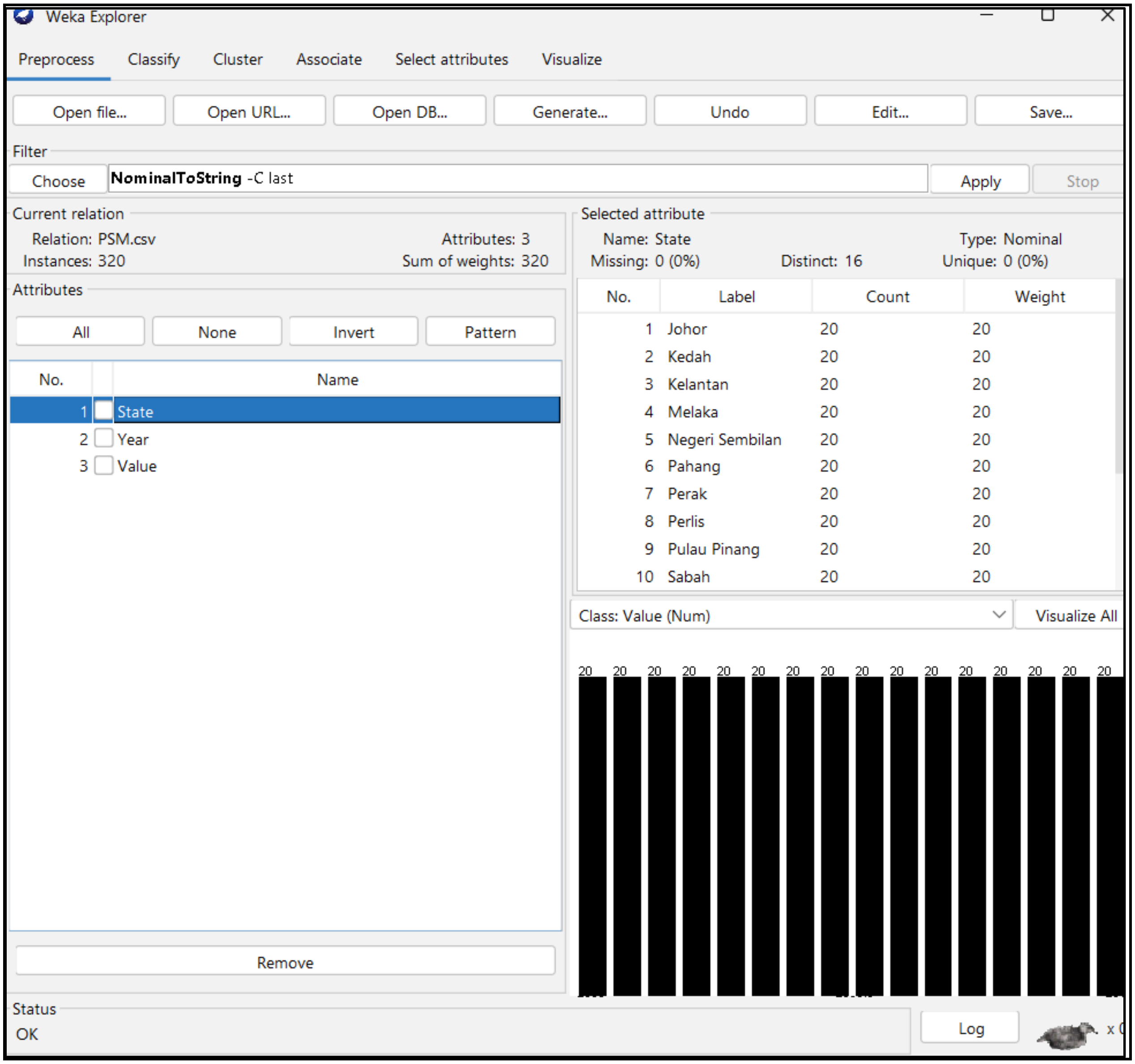
Prinsip-prinsip yang menjadi asas kepada Dasar Keselamatan ICT AADK dan perlu dipatuhi adalah seperti berikut:

(a) Akses atas dasar perlu mengetahui

Akses terhadap penggunaan aset ICT hanya diberikan untuk tujuan spesifik dan dihadkan kepada pengguna tertentu atas dasar "perlu mengetahui" sahaja. Ini bermakna akses hanya akan diberikan sekiranya peranan atau fungsi pengguna memerlukan maklumat tersebut. Pertimbangan untuk akses adalah berdasarkan kategori maklumat seperti yang dinyatakan di dalam dokumen Arahan Keselamatan;

Figure 4.1.3 The ICT' Security Rules of The Data In AADK

The name of the dataset is Number Of Drug Addicts By State, Malaysia, that is created on 25 March 2021. It is quite a recent dataset by Nur Atikah Binti Zabir that is under the MAMPU. This data is under the category of narcotics. The amount of instances in this dataset is 164. The Weka application on the current relation tab shows its total attributes, instances, and the sum of weight (which is the attribute contained). There are three attributes used in this dataset, which are the year, type of drug and value. Each attribute is categorical as a numeric and nominal base on the instances in it. Year and Value represent the numeric datatype, while the State represents the nominal datatype. The range of years is from 2000 until 2019 for each state. There are 16 instances of State, which are Perak, Perlis, Pulau Pinang, Sabah, Sarawak, Selangor, Terengganu, Wilayah Persekutuan Kuala Lumpur, Wilayah Persekutuan Labuan, Wilayah Persekutuan Putrajaya, Johor, Kedah, Kelantan, Melaka, Negeri Sembilan and Pahang. The minimum value is 0, while the highest maximum is 6685, with the average mean being 1528.756 with a standard variation being 1397.463. Figure 4.1.4, 4.1.5 and 4.1.6 shows the instances for each attribute with its descriptive statistics.



Filter

NominalToString -C last

Current relation

Relation: PSM.csv Attributes: 3

Instances: 320 Sum of weights: 320

Attributes

No.		Name
1	□	State
2	□	Year
3	□	Value

Selected attribute

Name: State Type: Nominal

Missing: 0 (0%) Distinct: 16 Unique: 0 (0%)

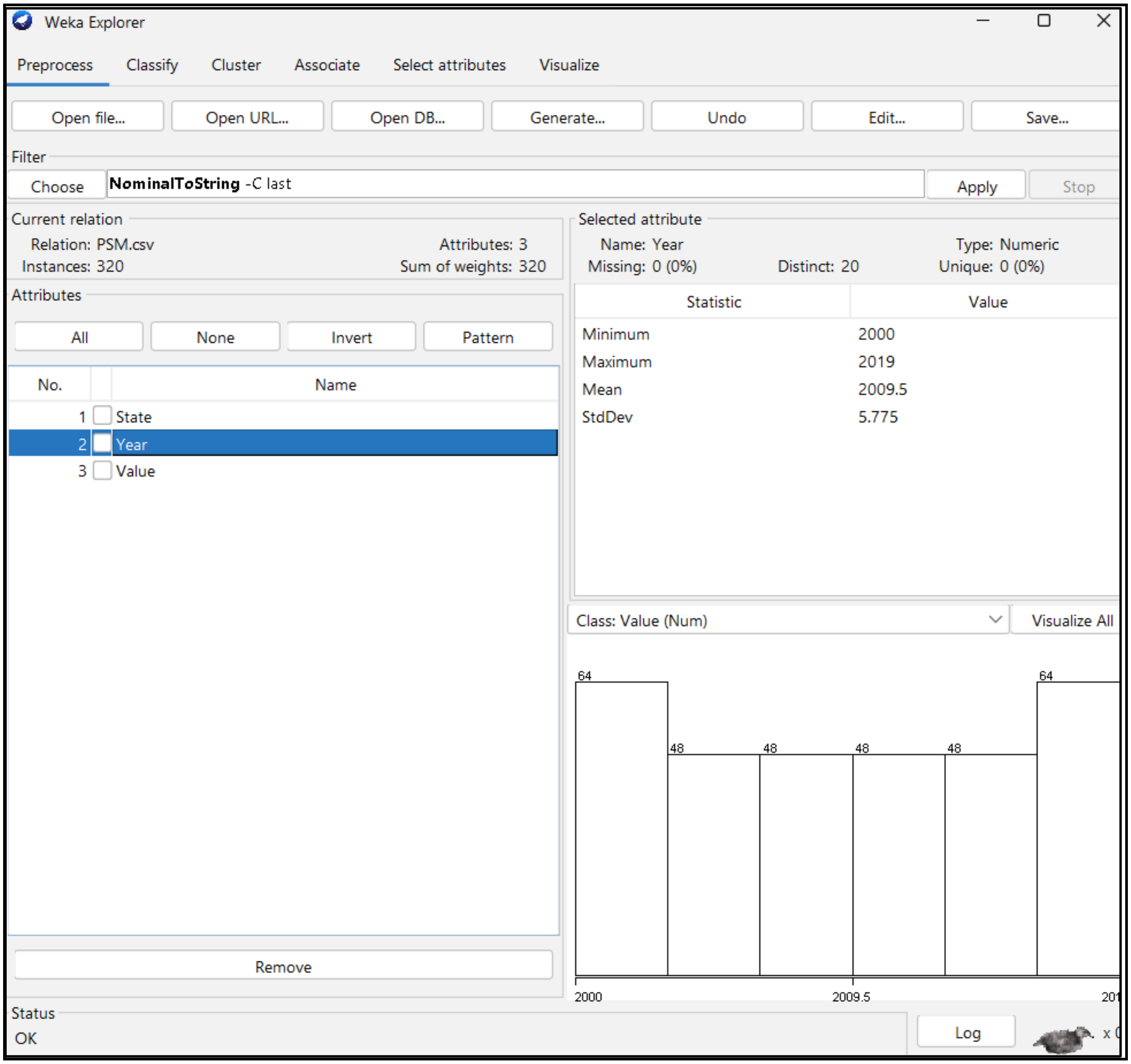
No.	Label	Count	Weight
1	Johor	20	20
2	Kedah	20	20
3	Kelantan	20	20
4	Melaka	20	20
5	Negeri Sembilan	20	20
6	Pahang	20	20
7	Perak	20	20
8	Perlis	20	20
9	Pulau Pinang	20	20
10	Sabah	20	20

Class: Value (Num)

Status

OK

Figure 4.1.4 The Instances of State's Attribute



Filter

NominalToString -C last

Current relation

Relation: PSM.csv
Attributes: 3
Instances: 320
Sum of weights: 320

Attributes

No.		Name
1	□	State
2	□	Year
3	□	Value

Selected attribute

Name: Year
Type: Numeric
Missing: 0 (0%)
Distinct: 20
Unique: 0 (0%)

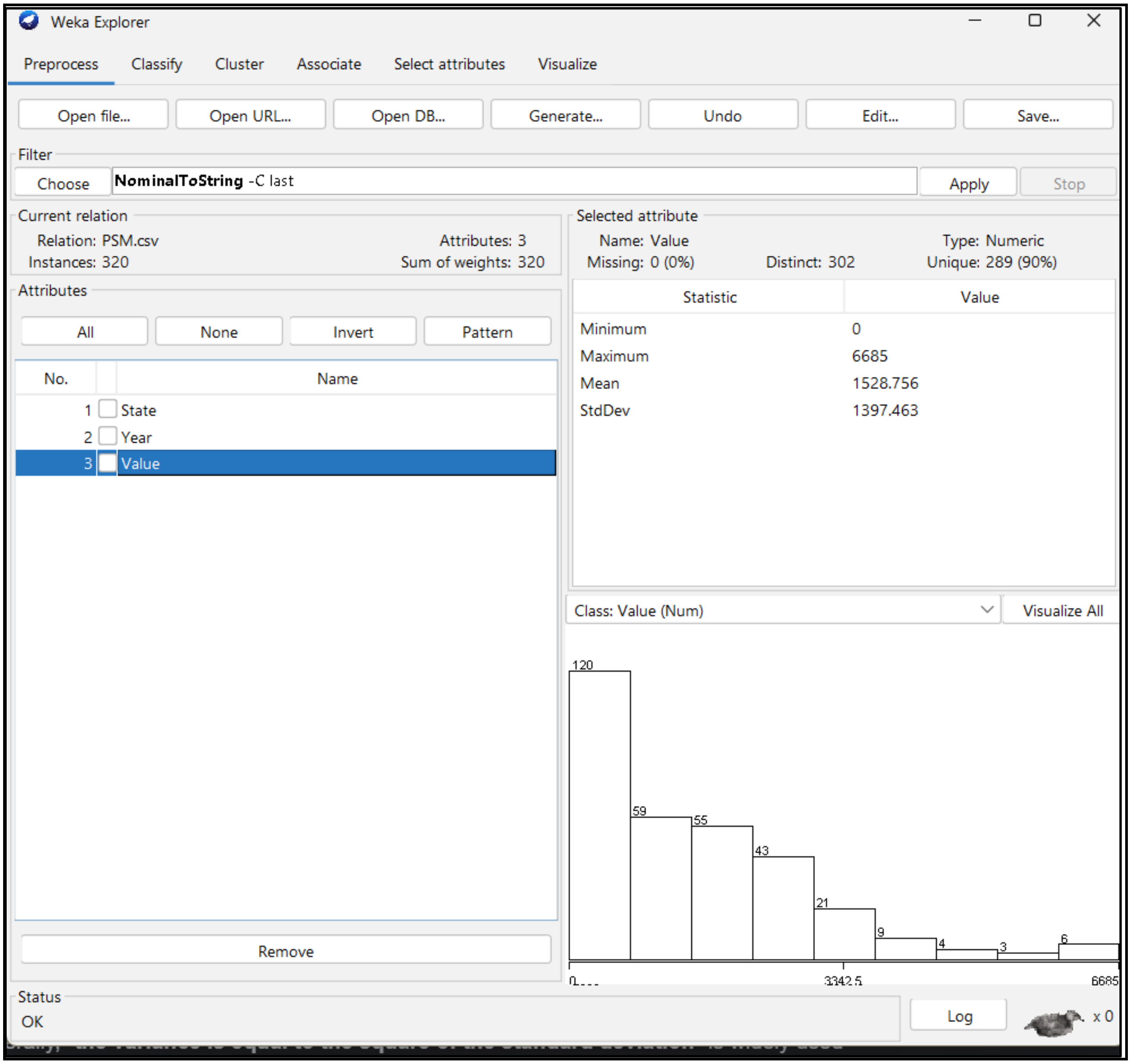
Statistic	Value
Minimum	2000
Maximum	2019
Mean	2009.5
StdDev	5.775

Class: Value (Num)

Status

OK

Figure 4.1.5 The Instances of Year's Attribute



NominalToString -C last

Current relation

Relation: PSM.csv — Attributes: 3
Instances: 320 — Sum of weights: 320

Attributes

No.		Name
1	□	State
2	□	Year
3	□	Value

Selected attribute

Name: Value — Type: Numeric
Missing: 0 (0%) — Distinct: 302 — Unique: 289 (90%)

Statistic	Value
Minimum	0
Maximum	6685
Mean	1528.756
StdDev	1397.463

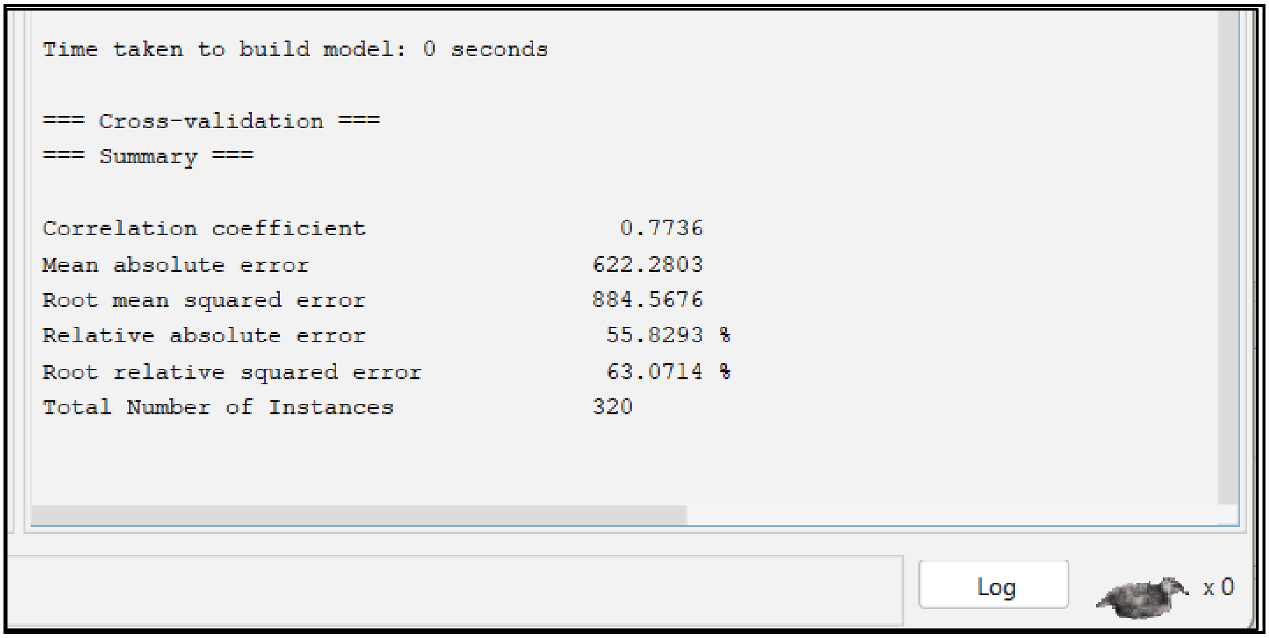
Class: Value (Num)

Status
OK

Figure 4.1.6 The Instances of Value’s Attribute

4.2 Data Preprocessing

After satisfying choosing the dataset, it needs to be preprocessed and cleaned. There are a lot of techniques and steps involved in this part that need to be done by selecting the proper method in a certain attribute, such as the nominal attribute has its own way of processing to clean it, which later to be used in modelling. In data mining, the most time spent in the process is this part, and because of that, this preprocessing is a very crucial step before moving on to the next process. This is because if the dataset is carefully preprocessed with the appropriate technique, later on, the result of using the model will be disrupted and may lead to wasting time by going back to this part and repeating it again until the desired and satisfying results reach after implementing the model based on the task given. Based on Figure 4.2 below shows that the correlation coefficient of the model used, Linear Regression, on the insufficient preprocessing on the certain dataset leads to low correlation between the attributes with a bigger error between each of the attributes. Preprocessing task is to reduce the error rate, thus increasing the correlation coefficient of the modelling.



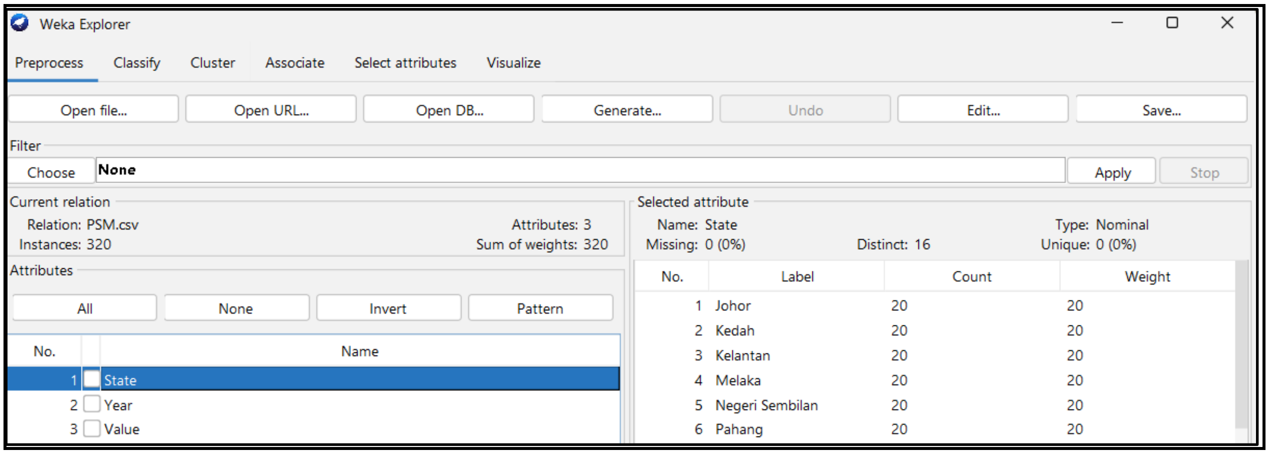
```
Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                   0.7736
Mean absolute error                     622.2803
Root mean squared error                 884.5676
Relative absolute error                  55.8293 %
Root relative squared error              63.0714 %
Total Number of Instances               320
```

Figure 4.2 The Results of Data Mining

4.2.1 Data Cleaning

Data cleaning in preprocessing aims to improve the data quality by removing any errors or redundant available in the data (Ridzuan & Wan Zainon, 2019). The dataset needs to be cleansed so that in further process, the evaluation of the model will produce a higher quality result that can have a good impact on the thesis decision. Even though this dataset is obtained from the MAMPU, it is not wise to just use it and proceed to the next stage of the process. Data cleansing is a process done on the original datasets to remove any outliers and abnormal values to ensure that the data is accurate and reliable to use. This process can be done either manually or by computer. To be honest, the use of a computer in this process is much more reliable and accurate because of the less time consumed and much safer. The instances of this dataset are 320 and doing the data cleaning for the entire instances is prone to errors for a human.

With the help of Weka, this can be done in a second. The first things need to do is find the missing data and search for the noisy data. Missing data here means that the value is filled with null of nan value in certain instances in a line of 320 instances available in the dataset. This dataset shows that there is no missing value to be found in each of the attributes. If there is a missing value, even it is one, it can interrupt and affect the evaluation of the model in the next phase. To avoid this, the missing value can be solved in two ways, which are by fill it with another mean value or just ignoring it. In this case, ignoring the missing tuple is a huge mistake, as this instance is not quite big enough for data mining not to be affected by it. So, the missing value needs to be fill by the mean value because the mean value is not affected by its outliers. The other one is noisy data, that is means the data cannot be translated by the computer. To handle the noisy data, there are three methods available which are binning, regression and clustering method.



Filter: None

Current relation

Relation: PSM.csv
Instances: 320
Attributes: 3
Sum of weights: 320

Attributes

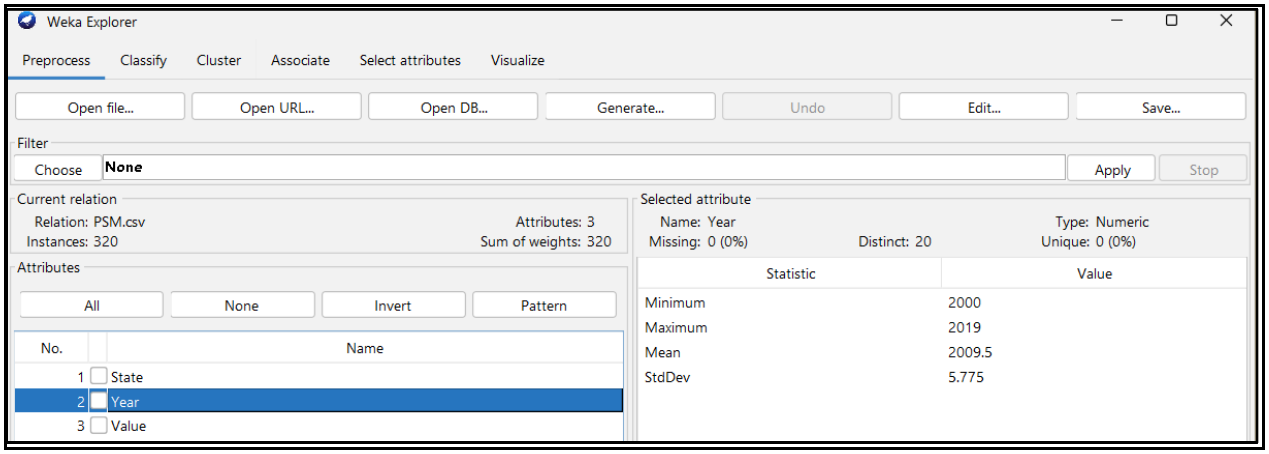
No.		Name
1	□	State
2	□	Year
3	□	Value

Selected attribute

Name: State
Missing: 0 (0%)
Distinct: 16
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Johor	20	20
2	Kedah	20	20
3	Kelantan	20	20
4	Melaka	20	20
5	Negeri Sembilan	20	20
6	Pahang	20	20

Figure 4.2.1 The Missing Value of State’s Attribute



Filter: None

Current relation

Relation: PSM.csv
Instances: 320
Attributes: 3
Sum of weights: 320

Attributes

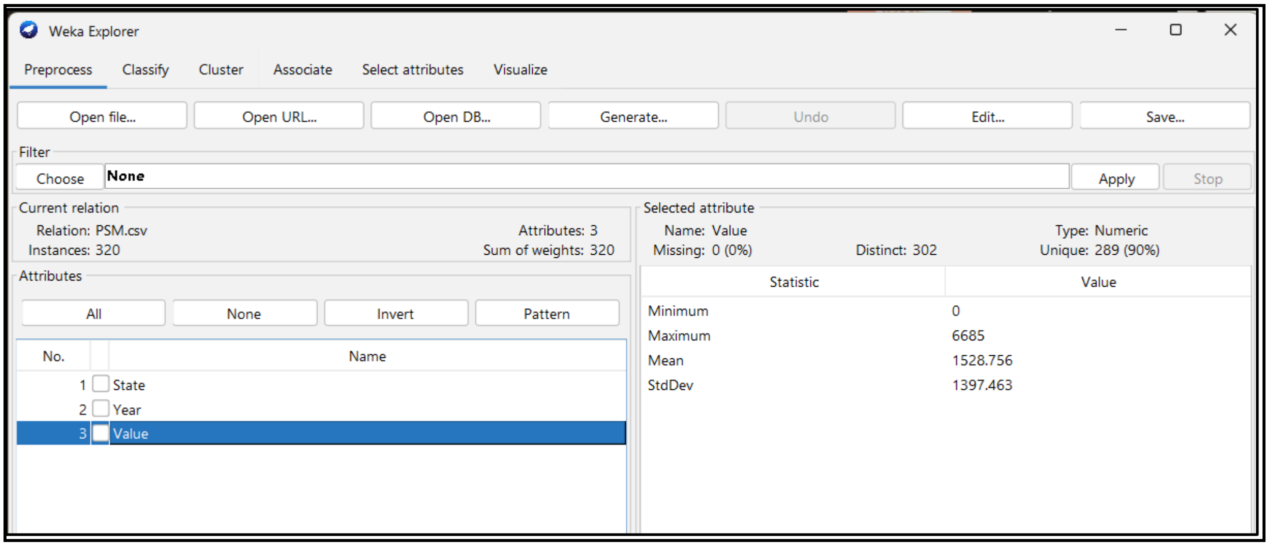
No.		Name
1	□	State
2	□	Year
3	□	Value

Selected attribute

Name: Year
Missing: 0 (0%)
Distinct: 20
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	2000
Maximum	2019
Mean	2009.5
StdDev	5.775

Figure 4.2.1.2 The Missing Value of Year’s Attribute



Filter: None

Current relation
Relation: PSM.csv
Attributes: 3
Instances: 320
Sum of weights: 320

Attributes

No.	Name
1	□ State
2	□ Year
3	□ Value

Selected attribute
Name: Value
Type: Numeric
Missing: 0 (0%)
Distinct: 302
Unique: 289 (90%)

Statistic	Value
Minimum	0
Maximum	6685
Mean	1528.756
StdDev	1397.463

Figure 4.2.1.3 The Missing Value of Value's Attribute

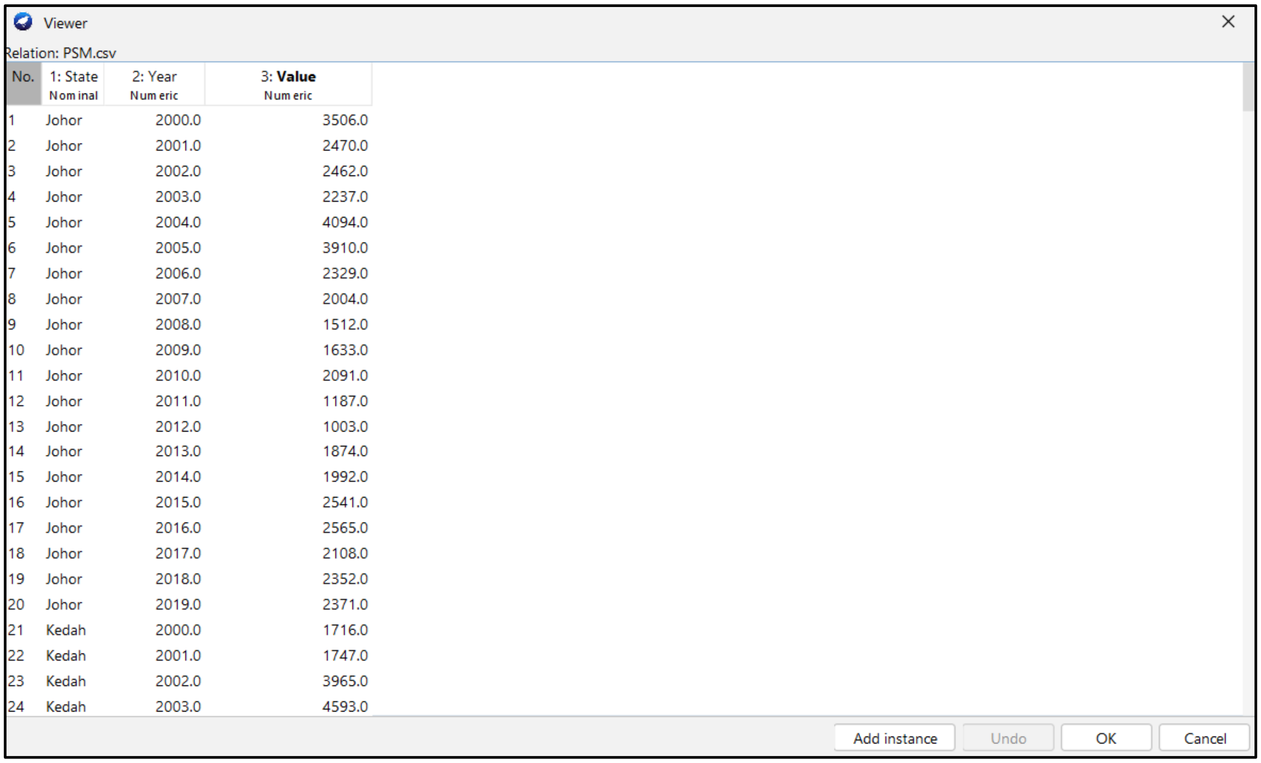
In Figures 4.2.1, 4.2.1.2 and 4.2.1.3, the Weka Explorer shows the missing value and its percentage from the entire selected attribute. It shows that all three attributes have zero missing values with zero percentages. So both the methods of handling the missing and noisy data do not need to be applied in this dataset because the data is in good condition that all can be read by the application.

4.2.2 Data Transformation

Data transformation is a process where the clean and appropriate data will be changed and transformed into a more effective format to ease the data mining process. There are a lot of techniques that can be used, such as data reduction, attribute selection, concept hierarchy generation and much more, but for this dataset, the suitable technique that can be used is the normalization and discretization. Other than that is not really suitable to be used as it can have a negative impact on the accuracy of the modelling phase later on because this dataset consists of one numerical data type of a high value that needs to use the normalization and discretization techniques. Later on, it will be combined with the unstructured data with structured data to be analyzed later.

Normalization is a must-do when the user wants to manage the attributes that have a high differentiation scale because the high differentiation scale can lead the data to interrupt the quality of the model used later (GeeksforGeeks, 2019b). So to reduce the huge gap between attributes, normalization is used. If the multiple attributes have a high differentiation in value, the model result later in the Weka will produce high error values, especially in the Mean Absolute Error and the Root Mean Squared Error. So these values need to be normalized to bring the attributes on the same scale.

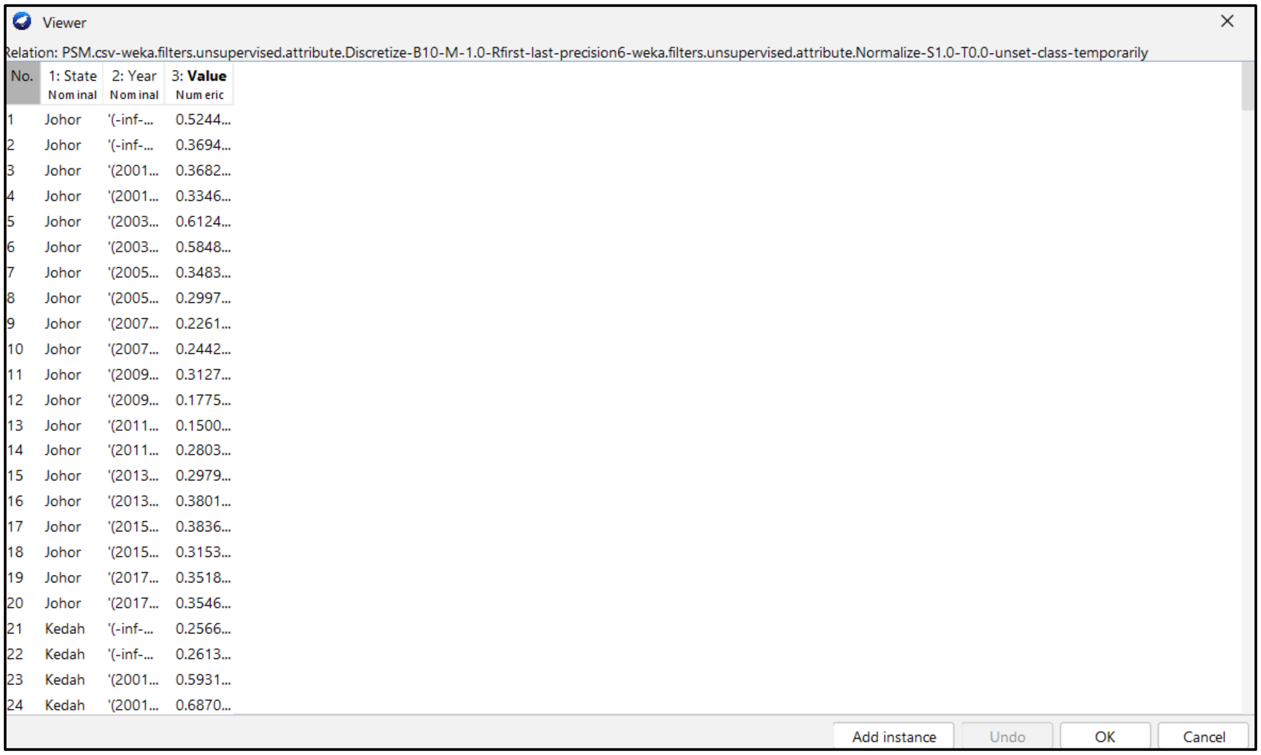
This can be seen on the value attribute as it is weighted so much value because of the range value from 0 until 6685. Figure 4.2.2.1 shows the actual instances of Value's attribute, while Figure 4.2.2.2 Shows the instances of Value's attribute after normalization. The normalization process has made the instances of Value's attribute to be changed from 0-6685 to 0-1.0 scaling with many floating numbers.



Relation: PSM.csv

No.	1: State Nominal	2: Year Numeric	3: **Value** Numeric
1	Johor	2000.0	3506.0
2	Johor	2001.0	2470.0
3	Johor	2002.0	2462.0
4	Johor	2003.0	2237.0
5	Johor	2004.0	4094.0
6	Johor	2005.0	3910.0
7	Johor	2006.0	2329.0
8	Johor	2007.0	2004.0
9	Johor	2008.0	1512.0
10	Johor	2009.0	1633.0
11	Johor	2010.0	2091.0
12	Johor	2011.0	1187.0
13	Johor	2012.0	1003.0
14	Johor	2013.0	1874.0
15	Johor	2014.0	1992.0
16	Johor	2015.0	2541.0
17	Johor	2016.0	2565.0
18	Johor	2017.0	2108.0
19	Johor	2018.0	2352.0
20	Johor	2019.0	2371.0
21	Kedah	2000.0	1716.0
22	Kedah	2001.0	1747.0
23	Kedah	2002.0	3965.0
24	Kedah	2003.0	4593.0

Figure 4.2.2.1 The Initial Instances of Value’s Attribute



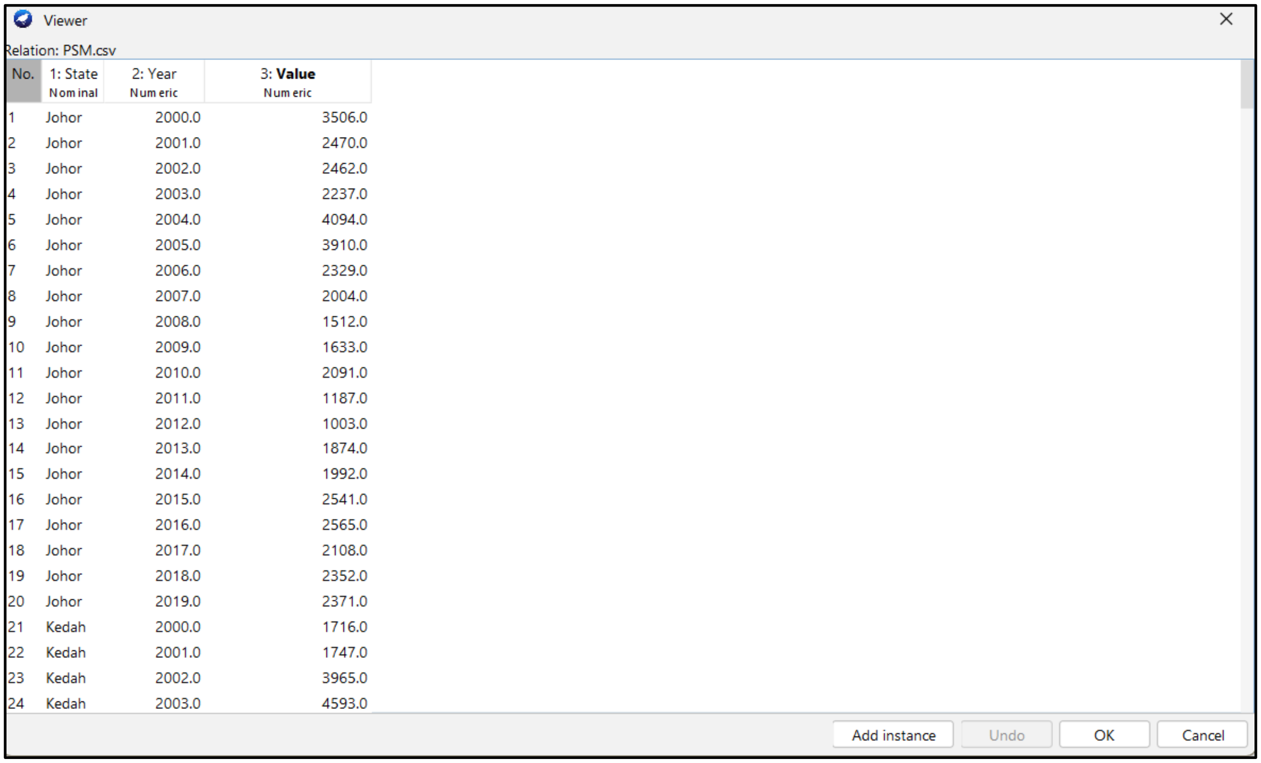
Relation: PSM.csv-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-unset-class-temporarily

No.	1: State Nominal	2: Year Nominal	3: **Value** Numeric
1	Johor	'(-inf-...	0.5244...
2	Johor	'(-inf-...	0.3694...
3	Johor	'(2001...	0.3682...
4	Johor	'(2001...	0.3346...
5	Johor	'(2003...	0.6124...
6	Johor	'(2003...	0.5848...
7	Johor	'(2005...	0.3483...
8	Johor	'(2005...	0.2997...
9	Johor	'(2007...	0.2261...
10	Johor	'(2007...	0.2442...
11	Johor	'(2009...	0.3127...
12	Johor	'(2009...	0.1775...
13	Johor	'(2011...	0.1500...
14	Johor	'(2011...	0.2803...
15	Johor	'(2013...	0.2979...
16	Johor	'(2013...	0.3801...
17	Johor	'(2015...	0.3836...
18	Johor	'(2015...	0.3153...
19	Johor	'(2017...	0.3518...
20	Johor	'(2017...	0.3546...
21	Kedah	'(-inf-...	0.2566...
22	Kedah	'(-inf-...	0.2613...
23	Kedah	'(2001...	0.5931...
24	Kedah	'(2001...	0.6870...

Figure 4.2.2.2 The Instances of Value’s Attribute After Being Normalize

Discretization is a method to transform a big number of data into a smaller one by grouping it. This is to ease the evaluation and management of data. Discretization is converting the continuous numerical data into a set of group intervals with minimum data loss. There are two types of data discretization, which are Supervised and Unsupervised. Supervised discretization means that it will discretize the data in the class data, while Unsupervised discretization means it discretizes the data according to the operation's flow either by top-down splitting strategy or the bottom-up merging strategy (Discretization in Data Mining - Javatpoint, n.d.).

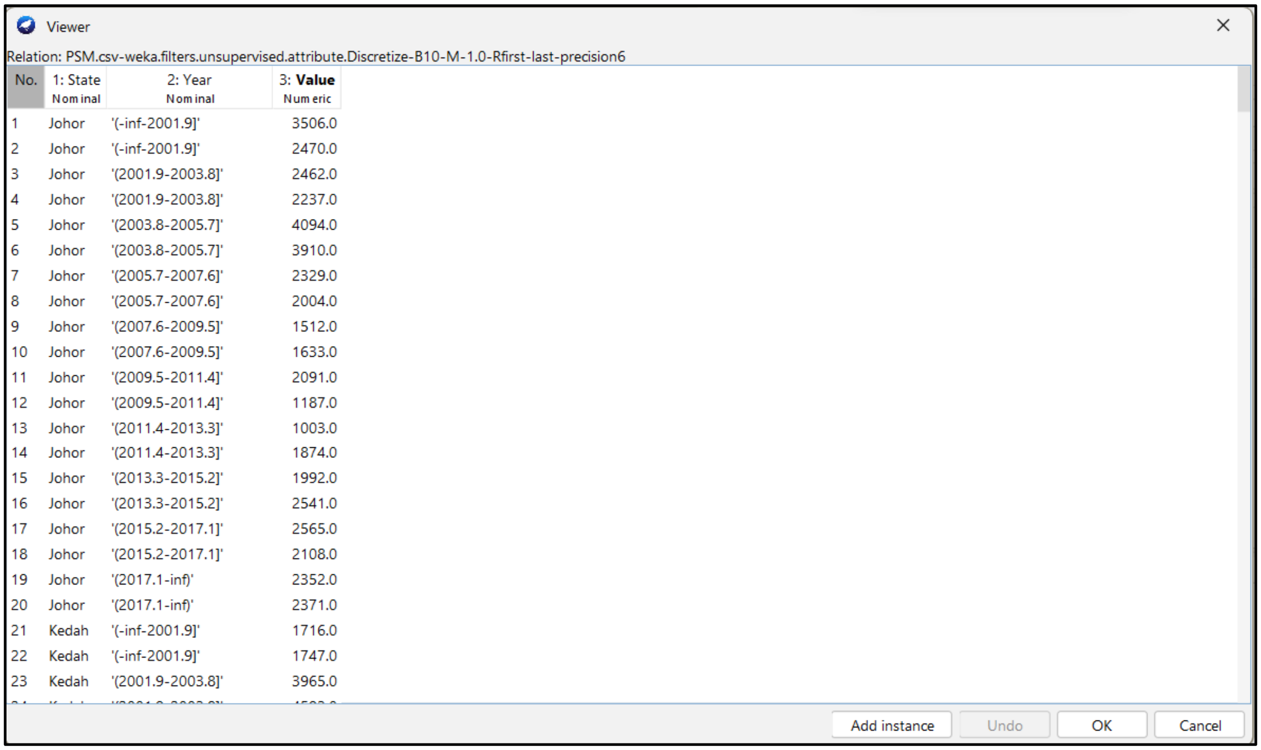
The discretization occurs in the Weka by dividing the normalized value into a hierarchy which is later on converted into a nominal data type. By doing this, the data will look more neat and fit to be input into the modelling phase later on rather than being discretised without normalising it that has a big instances gap value on the Year's attribute. For this thesis, the discretization method that will be used the Unsupervised discretization for a better result of the model.



Relation: PSM.csv

No.	1: State Nominal	2: Year Numeric	3: **Value** Numeric
1	Johor	2000.0	3506.0
2	Johor	2001.0	2470.0
3	Johor	2002.0	2462.0
4	Johor	2003.0	2237.0
5	Johor	2004.0	4094.0
6	Johor	2005.0	3910.0
7	Johor	2006.0	2329.0
8	Johor	2007.0	2004.0
9	Johor	2008.0	1512.0
10	Johor	2009.0	1633.0
11	Johor	2010.0	2091.0
12	Johor	2011.0	1187.0
13	Johor	2012.0	1003.0
14	Johor	2013.0	1874.0
15	Johor	2014.0	1992.0
16	Johor	2015.0	2541.0
17	Johor	2016.0	2565.0
18	Johor	2017.0	2108.0
19	Johor	2018.0	2352.0
20	Johor	2019.0	2371.0
21	Kedah	2000.0	1716.0
22	Kedah	2001.0	1747.0
23	Kedah	2002.0	3965.0
24	Kedah	2003.0	4593.0

Figure 4.2.2.3 The Instances of Year’s Attribute Before Being Discretization



Relation: PSM.csv-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6

No.	1: State Nominal	2: Year Nominal	3: Value Numeric
1	Johor	'(-inf-2001.9]'	3506.0
2	Johor	'(-inf-2001.9]'	2470.0
3	Johor	'(2001.9-2003.8]'	2462.0
4	Johor	'(2001.9-2003.8]'	2237.0
5	Johor	'(2003.8-2005.7]'	4094.0
6	Johor	'(2003.8-2005.7]'	3910.0
7	Johor	'(2005.7-2007.6]'	2329.0
8	Johor	'(2005.7-2007.6]'	2004.0
9	Johor	'(2007.6-2009.5]'	1512.0
10	Johor	'(2007.6-2009.5]'	1633.0
11	Johor	'(2009.5-2011.4]'	2091.0
12	Johor	'(2009.5-2011.4]'	1187.0
13	Johor	'(2011.4-2013.3]'	1003.0
14	Johor	'(2011.4-2013.3]'	1874.0
15	Johor	'(2013.3-2015.2]'	1992.0
16	Johor	'(2013.3-2015.2]'	2541.0
17	Johor	'(2015.2-2017.1]'	2565.0
18	Johor	'(2015.2-2017.1]'	2108.0
19	Johor	'(2017.1-inf)'	2352.0
20	Johor	'(2017.1-inf)'	2371.0
21	Kedah	'(-inf-2001.9]'	1716.0
22	Kedah	'(-inf-2001.9]'	1747.0
23	Kedah	'(2001.9-2003.8]'	3965.0

Figure 4.2.2.4 The Instances of Year's Attribute After Being Discretization

4.3 Selected Models

To do the model to be implemented into the preprocessing datasets, the user must select the model that has information on how it will work. Each of the processes in the model must be fully understood in order to do the machine learning. For this thesis, the model that will be used is the Linear Regression and the Random Tree models. These two models are best for finding the prediction value between variables. The regression problem also can be solved by using these two models as these models are easier to understand and also produce a good result for evaluation later. In Weka, these models are available for the user to be playing and implement with normal settings has been set to ease the beginner user.

4.3.1 Linear Regression

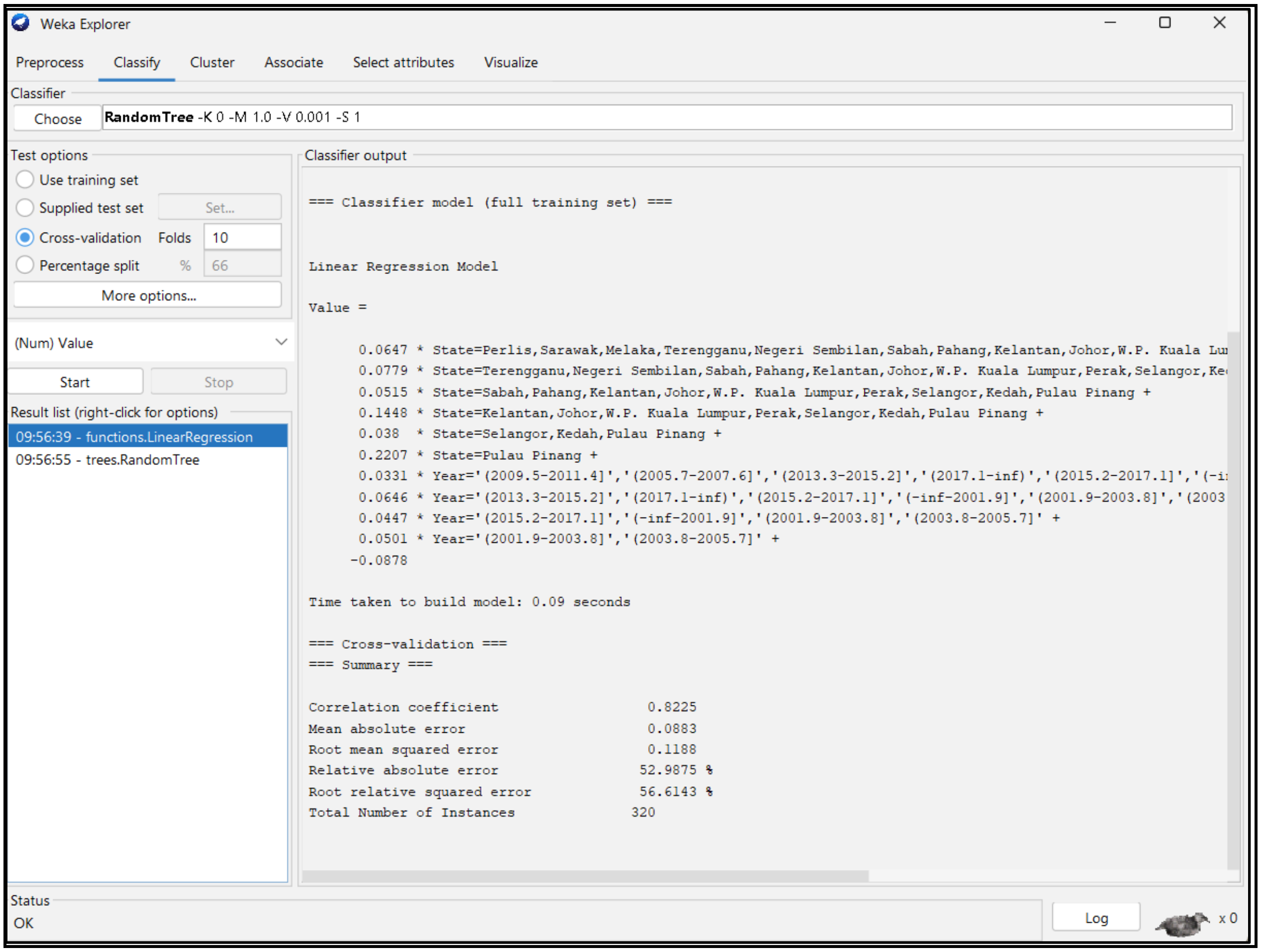
The most popular of all statistical approaches, linear regression analysis, is the study of linear, additive connections between variables. If Y represents the "dependent" variable whose values that is intended to forecast, then let X1,..., Xk represent the "independent" variables from which it wishes to predict it, with Xit representing the value of variable Xi in period t (or in row t of the data set). Thus, the following equation is used to get the expected value of Yt:

$$\boxed{Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + \ldots\ldots. + b_kX_{kt}}$$

Figure 4.3.1.1 The Formula For Linear Regression Model

This formula states that the prediction for the value of Y is a straight-line function that depends on the X-variables, containing the others' value as a constant one, and the affection of different X variables to the result models is additive. The slope of the Linear Regression model is related to the Y values of the constants b1, b2, …, bk, which is named as the correlation of the coefficient of the variables. Linear Regression is best used when the user wants to make the prediction on the value of a variable based on the value of the other variable in the datasets, while Multiple Linear Regression is used when there is more than one input variable used to make the comparison on the other variable.

This can be done in Weka under the Classify tabs with the loaded dataset under four test options which are using a training set, supplying it with the test set, cross-validation based on how many folds and lastly, the percentage split of how many it will train and how many it will reserve for the validation set. Only one training set is used for this thesis which is the cross-validation testing option for both algorithms. Figure 4.3.1.2 shows the results of the Linear Regression algorithm using the cross-validation testing option with an iteration of ten times.



Classifier: RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) Value

Result list (right-click for options)
09:56:39 - functions.LinearRegression
09:56:55 - trees.RandomTree

Classifier output

```
=== Classifier model (full training set) ===


Linear Regression Model

Value =

      0.0647 * State=Perlis,Sarawak,Melaka,Terengganu,Negeri Sembilan,Sabah,Pahang,Kelantan,Johor,W.P. Kuala Lu
      0.0779 * State=Terengganu,Negeri Sembilan,Sabah,Pahang,Kelantan,Johor,W.P. Kuala Lumpur,Perak,Selangor,Ke
      0.0515 * State=Sabah,Pahang,Kelantan,Johor,W.P. Kuala Lumpur,Perak,Selangor,Kedah,Pulau Pinang +
      0.1448 * State=Kelantan,Johor,W.P. Kuala Lumpur,Perak,Selangor,Kedah,Pulau Pinang +
      0.038  * State=Selangor,Kedah,Pulau Pinang +
      0.2207 * State=Pulau Pinang +
      0.0331 * Year='(2009.5-2011.4]','(2005.7-2007.6]','(2013.3-2015.2]','(2017.1-inf)','(2015.2-2017.1]','(-i
      0.0646 * Year='(2013.3-2015.2]','(2017.1-inf)','(2015.2-2017.1]','(-inf-2001.9]','(2001.9-2003.8]','(2003
      0.0447 * Year='(2015.2-2017.1]','(-inf-2001.9]','(2001.9-2003.8]','(2003.8-2005.7]' +
      0.0501 * Year='(2001.9-2003.8]','(2003.8-2005.7]' +
     -0.0878

Time taken to build model: 0.09 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8225
Mean absolute error                      0.0883
Root mean squared error                  0.1188
Relative absolute error                 52.9875 %
Root relative squared error             56.6143 %
Total Number of Instances              320
```

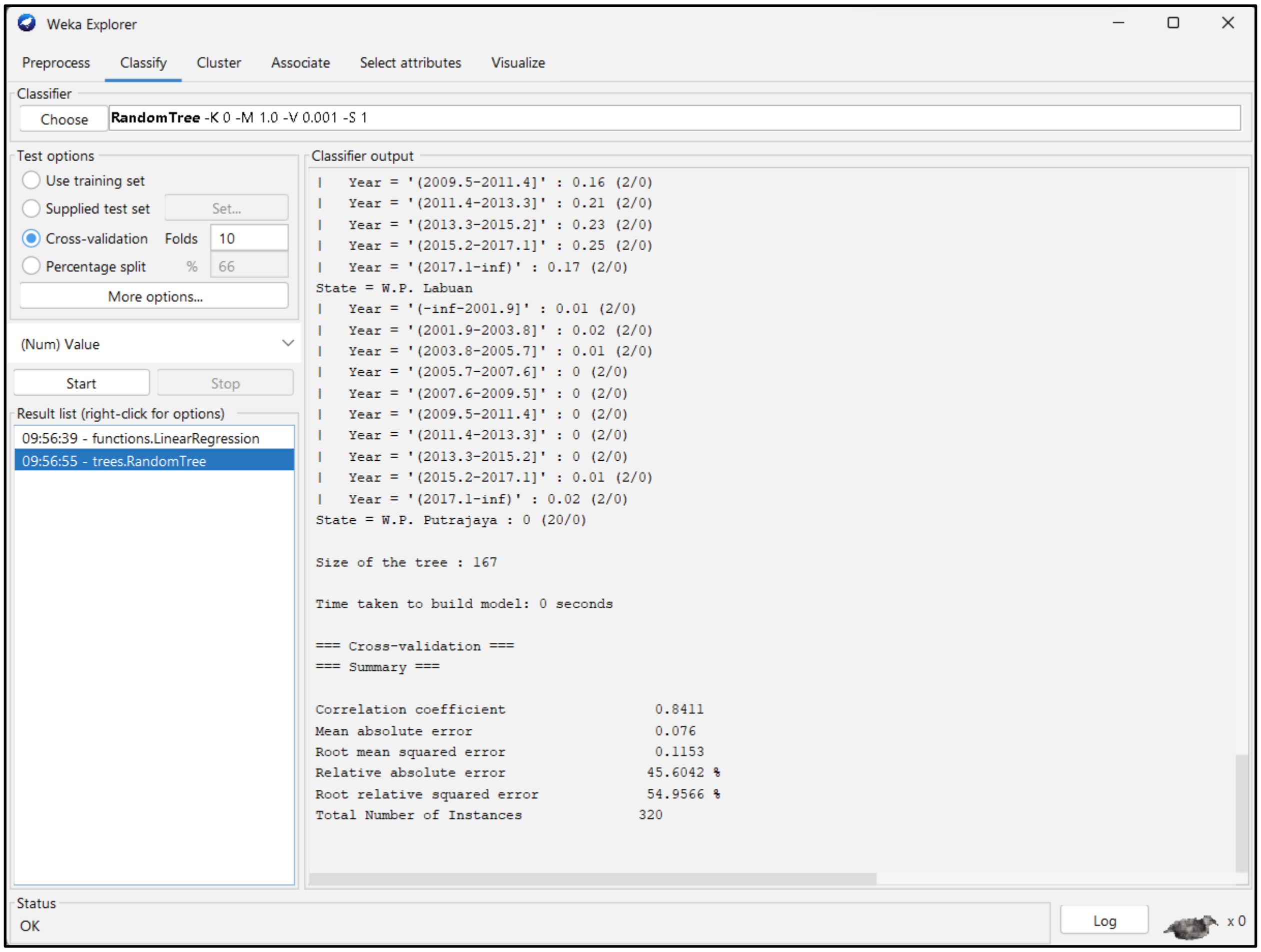
Status
OK

Figure 4.3.1.2 The Results of Linear Regression Algorithm Using the Cross-Validation

4.3.2 Random Tree

Random Tree is a model that combines the two algorithms in Weka, which is the single model tree combined with the Random Forest ideas. The tree model is a concept of a tree that has many branches; with every branch, there is a leaf that contains the linear model that has been corrected to be used for the local subspace described by the leaf. Random Forest is proven to increase the result of a single Random Tree by considering the tree diversity created by using the two ways of randomization. The training data will be a sample in order to replace every single Tree, as shown in the Bagging. The Tree keeps growing when the model computes the best possible split for each of the leaves by considering random instances of the overall attributes and choosing the best split for that occasion. For the Tree that has a Random model tree for first-time use, it will combine the model tree and the random forests. To simplify the optimization during the procedure, the Random Tree will create the split selection and bring about the reasonably balanced Tree by applying only one global setting for the ridge value to be implemented in all of the leaves in that Tree.

This also can be done in Weka under the Classify tabs with the loaded dataset and the same with the previous Linear Regression algorithm. This thesis will use two testing options which are the percentage split and the cross-validation. Figure 4.3.2.1 shows the Random Tree algorithm using the cross-validation testing option with an iteration of ten times



Classifier

RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Num) Value

Result list (right-click for options)

09:56:39 - functions.LinearRegression

09:56:55 - trees.RandomTree

Classifier output

```
|   Year = '(2009.5-2011.4]' : 0.16 (2/0)
|   Year = '(2011.4-2013.3]' : 0.21 (2/0)
|   Year = '(2013.3-2015.2]' : 0.23 (2/0)
|   Year = '(2015.2-2017.1]' : 0.25 (2/0)
|   Year = '(2017.1-inf)' : 0.17 (2/0)
State = W.P. Labuan
|   Year = '(-inf-2001.9]' : 0.01 (2/0)
|   Year = '(2001.9-2003.8]' : 0.02 (2/0)
|   Year = '(2003.8-2005.7]' : 0.01 (2/0)
|   Year = '(2005.7-2007.6]' : 0 (2/0)
|   Year = '(2007.6-2009.5]' : 0 (2/0)
|   Year = '(2009.5-2011.4]' : 0 (2/0)
|   Year = '(2011.4-2013.3]' : 0 (2/0)
|   Year = '(2013.3-2015.2]' : 0 (2/0)
|   Year = '(2015.2-2017.1]' : 0.01 (2/0)
|   Year = '(2017.1-inf)' : 0.02 (2/0)
State = W.P. Putrajaya : 0 (20/0)

Size of the tree : 167

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8411
Mean absolute error                      0.076
Root mean squared error                  0.1153
Relative absolute error                 45.6042 %
Root relative squared error             54.9566 %
Total Number of Instances              320
```

Status

OK

Figure 4.3.2.1 The Random Tree Algorithm Using the Cross-Validation

4.4 Result and Discussions

The analysis of data is carried out after the modelling method, and its technique on each level is applied to determine the distribution of value dataset that is required by this task which is the prediction of drug abuse. The evaluation is carried out to evaluate the performance of each algorithm classifier in order to compare the accuracy of the results obtained. Table 4.1 shows the results of each algorithm.

Cross Validation = 10 Folds		
	Linear Regression	**Random Tree**
Correlation Coefficient	0.8255	0.8411
Mean Absolute Error	0.0883	0.0760
Root Mean Squared Error	0.1188	0.1153
Relative Absolute Error	52.9875 %	45.6042 %
Root Relative Squared Error	56.6143 %	54.9566 %

Table 4.1 The Results of Linear Regression and Random Tree Classifiers Algorithms

A correlation coefficient is an overall proportion for determining the correct correlation between the attributes. It is a simple way to measure the effectiveness of the model used to make the regression prediction. The high the correlation coefficient of the model, the more accurate the prediction value it will show (Fürnkranz et al., 2011). For the mean absolute error on the test option of ten foldings cross-validation, the Linear Regression model seems to have a higher overall error compared to the Random Tree method. This can be seen in the first error measure, which is the mean absolute error. The Linear Regression model stated a record of 0.0883 compared to the other one stated only 0.0760. On the root mean squared error, the difference is not much bigger as the difference is only 0.0035. For the relative absolute squared error, the Linear Regression model also stated a higher percentage of error rate, which is about 52.9875%, compared to the Decision Tree model, which is only 45.6042%. Lastly, the comparison between both models based on the root relative squared error is the least difference, only about 1.6577%. In conclusion, the cross-validation testing is suitable for the Random Tree algorithm because it is a better fit with the best correlation between the attributes because of the lowest overall error rate such as mean absolute error, root means squared error, relative absolute squared error and root relative squared error.

4.5 Visualizing with Power BI

The presentation of results can be more interactive by using the Power BI software from Microsoft. This software automatically reads the data from the file that the user has already converted from the ARFF.file (Attribute-Relation File Format) to the CSV.file (comma-separated values) and any other file formats. This thesis focuses entirely on three format files which are the ARFF, CSV, and excel type of file to store the datasets, either it is being preprocessed, or is already implemented a model to it to do the regression problem. On the main page of the dashboard, there are two clustered column charts. On the top page of the dashboard, there are eight buttons available which text which is Data Overview, Decision Tree, Decision Tree 2, Decision Tree 3, Decision Tree 4, Linear Regression (Percentage Splits), Linear Regression (Cross-Validations), and lastly Comparison. Each button has an action that will redirect the user to the respective page. On the page Decision Tree, Decision Tree 2, Decision Tree 3, and Decision Tree 4, there are two button arrows which are the right arrow and the left arrow.

The first one is the value of drug abusers sorted by the year. This shows all the drug abusers from all 16 states in Malaysia on the Y-axis and its value on the X-axis. The raw data sets show that the total number of drug abusers keep increasing from the year 2000 until the year 2004. The value then decreased drastically from that year to the year 2008. From the year 2008 until the year 2010, the trend seems to climb again and reach a downfall again until the year 2012. Afterwards, it levelled up drastically back on until the year 2016. From the year 2016 until 2019, the trend seems to be changing only a little bit. The next chart is the value of drug abusers sorted by 16 states that including all the years from 2000 until 2020. On the Y-axis, it shows the value of drug abusers from all the years' range, while on the X-axis, it shows the list of 16 states containing drug abusers sorted by descending order. The state with the highest number of drug abusers all year from 2000 until 2019 is Pulau Pinang, Kedah listed in second place, while Selangor on third place, followed by Perak, Wilayah Persekutuan Kuala Lumpur, Johor, Kelantan, Pahang, Sabah, Negeri Sembilan, Terengganu, Melaka, Sarawak, Perlis, Wilayah Persekutuan Labuan and lastly with the least drug abusers is the Wilayah Persekutuan Putrajaya. Pulau Pinang recorded a huge sum of numbers as the total of drug abusers reached 81000. The value of drug abusers from Kedah, Selangor, Perak, Wilayah Persekutuan Kuala Lumpur, Johor, and Kelantan recorded steadily from a range of 52000 to 45000. Pahang, Sabah, Negeri Sembilan, and Terengganu recorded values in the range of 20000, while Melaka, Sarawak, Perlis, Wilayah Persekutuan Labuan, and Wilayah Persekutuan recorded the value below 12000.

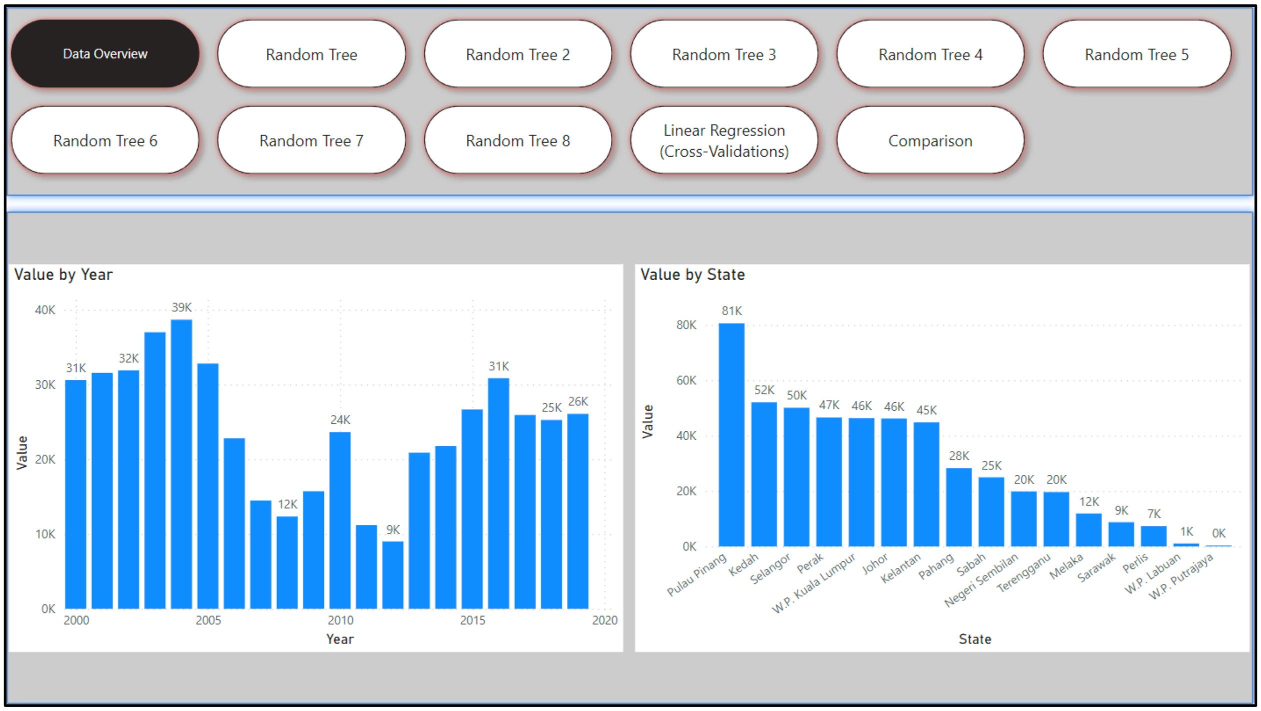
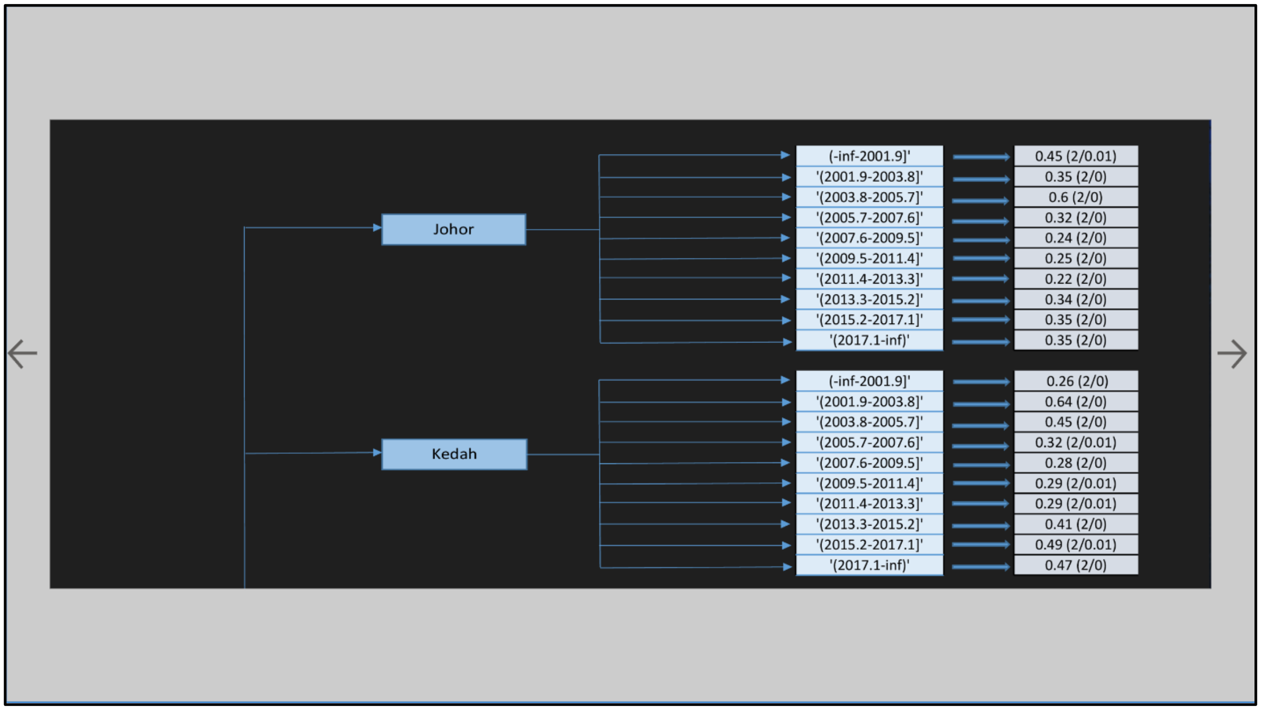


Figure 4.5.1 The Overview of Overall Datasets

Moving on, the first model used is the Random Tree. The result of the model displayed as a visual of a tree that has its own branches and its own leaf on a branch. Each branch represents the state and the year. The year is being discretized into ten parts, which are (-inf-2001.9), (2001.9-2003.8), (2003.8-2005.7), (2005.7-2007.6), (2007.6-2009.5), (2009.5-2011.4), (2011.4-2013.3), (2013.3-2015.2), (2015.2-2017.1), and lastly (2017.1-inf) years. The first branch is the Johor, with the year respectively.

Johor branch has the most inconsistent value throughout the years from (-inf-2001.9) until (2005.7-2007.6), then the value began to stable around 0.32 until 0.35. The second state is Kedah, which also consists of inconsistent value starting from (-inf-2001.9) until (2007.6-2009.5). In the year (2007.6-2009.5), the value seems to be steadily increasing slowly until the last year, which is (2017.1-inf) year, recorded on 0.47 predicted value. The third branch is the Kelantan state. The Kelantan state started with an upside-down value from the year (-inf-2001.9) until the year (2017.1-inf), which is the value recorded of a minimum of 0.07 in the year (2011.4-2013.3), and the maximum value in the year (2017.1-inf). The next branch is the Melaka state. The Melaka state has so low value predicted throughout the entire year, in which the highest recorded number is only 0.13 the year (2013.3-2015.2), and its lowest in the year (2009.5-2011.4), which is 0.04.



Johor	(-inf-2001.9]'	0.45 (2/0.01)
	'(2001.9-2003.8]'	0.35 (2/0)
	'(2003.8-2005.7]'	0.6 (2/0)
	'(2005.7-2007.6]'	0.32 (2/0)
	'(2007.6-2009.5]'	0.24 (2/0)
	'(2009.5-2011.4]'	0.25 (2/0)
	'(2011.4-2013.3]'	0.22 (2/0)
	'(2013.3-2015.2]'	0.34 (2/0)
	'(2015.2-2017.1]'	0.35 (2/0)
	'(2017.1-inf)'	0.35 (2/0)
Kedah	(-inf-2001.9]'	0.26 (2/0)
	'(2001.9-2003.8]'	0.64 (2/0)
	'(2003.8-2005.7]'	0.45 (2/0)
	'(2005.7-2007.6]'	0.32 (2/0.01)
	'(2007.6-2009.5]'	0.28 (2/0)
	'(2009.5-2011.4]'	0.29 (2/0.01)
	'(2011.4-2013.3]'	0.29 (2/0.01)
	'(2013.3-2015.2]'	0.41 (2/0)
	'(2015.2-2017.1]'	0.49 (2/0.01)
	'(2017.1-inf)'	0.47 (2/0)

Figure 4.5.2 The Decision Tree Predicted Value of Johor and Kedah’s State

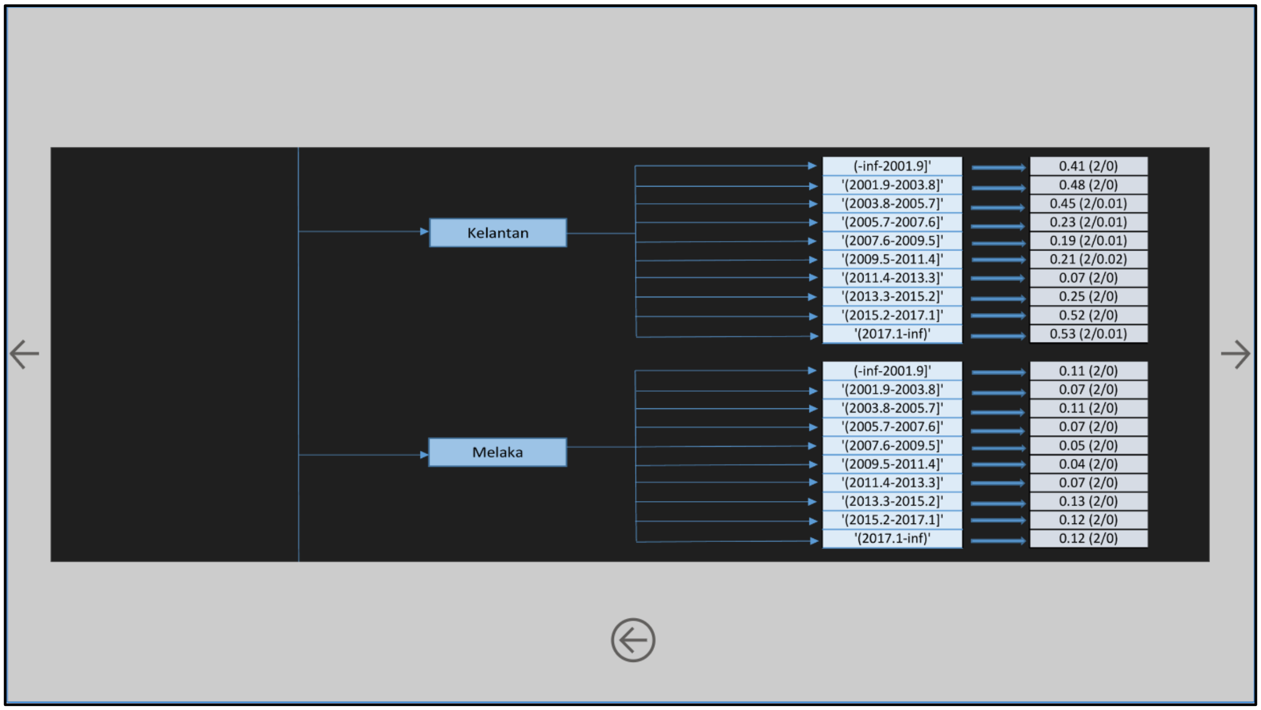
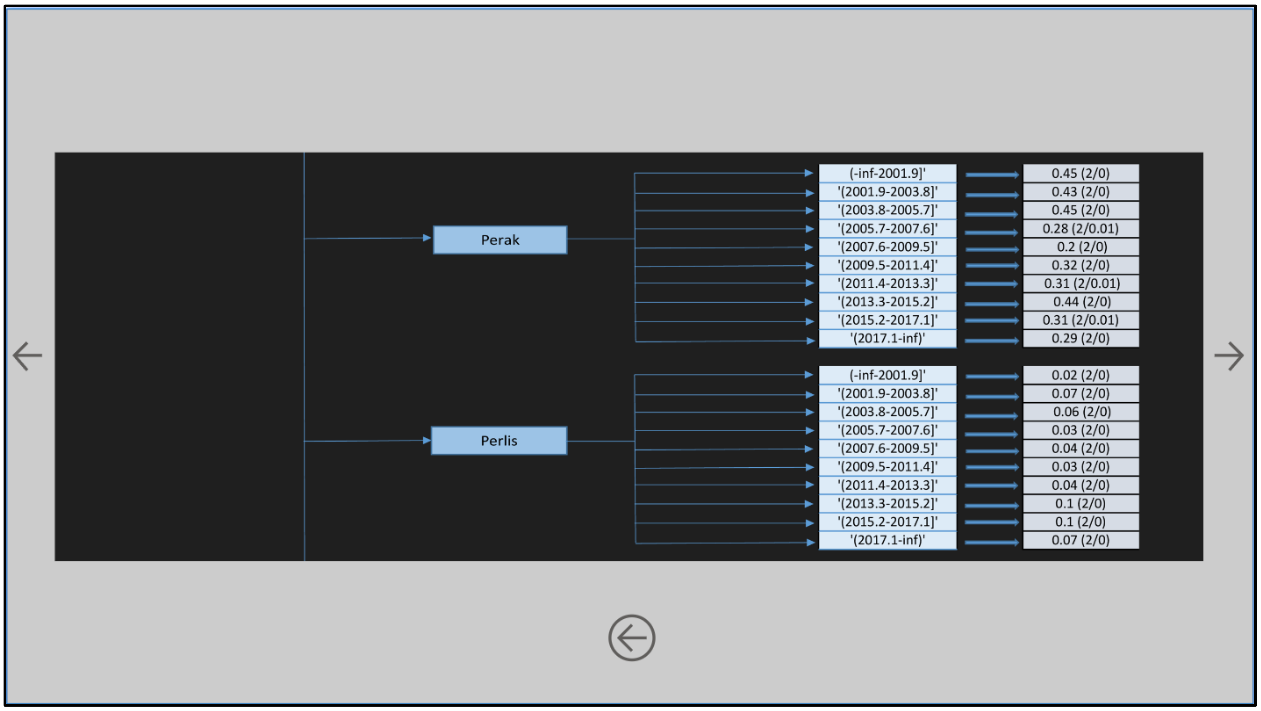


Figure 4.5.3 The Decision Tree Predicted Value of Kelantan and Melaka's State

Moving on to the next branch, which is the Negeri Sembilan state. This state has a gradual increase and decrease number of predicted values from 0.23 in the year (-inf-2001.9) to 0.11 on its lowest in three groups of the year which are from (2005.7-2007.6), (2007.6-2009.5), (2009.5-2011.4). The Pahang state is the sixth branch of the Random Tree modelling. The Pahang state also has a small gradual value predicted differences throughout the year. The smallest value recorded on this branch is in the year (2005.7-2007.6), resulting in a 0.09 value. 0.34 is the highest value in Pahang's state, which is 0.34 for the year (2015.2-2017.1). Perak is the seventh branch in the Random Tree modelling. The Perak state gets the predicted value range from 0.45 to 0.29. The value is steadily decreasing from the year (2003.8-2005.7) until the year (2007.6-2009.5). After that, the predicted value of the Perak state rise and fell within the value of 0.32 and 0.44 at its highest in the year

(2013.3-2015.2). The eighth branch is the Perlis state. Perlis state record steadily low predicted value throughout the entire discretized year. The maximum predicted value is only 0.07 in the years (2001.9-2003.8) and (2017.1-inf). On the first discretized year, which is (-inf-2001.9), the Perlis state only recorded the predicted value at 0.02..



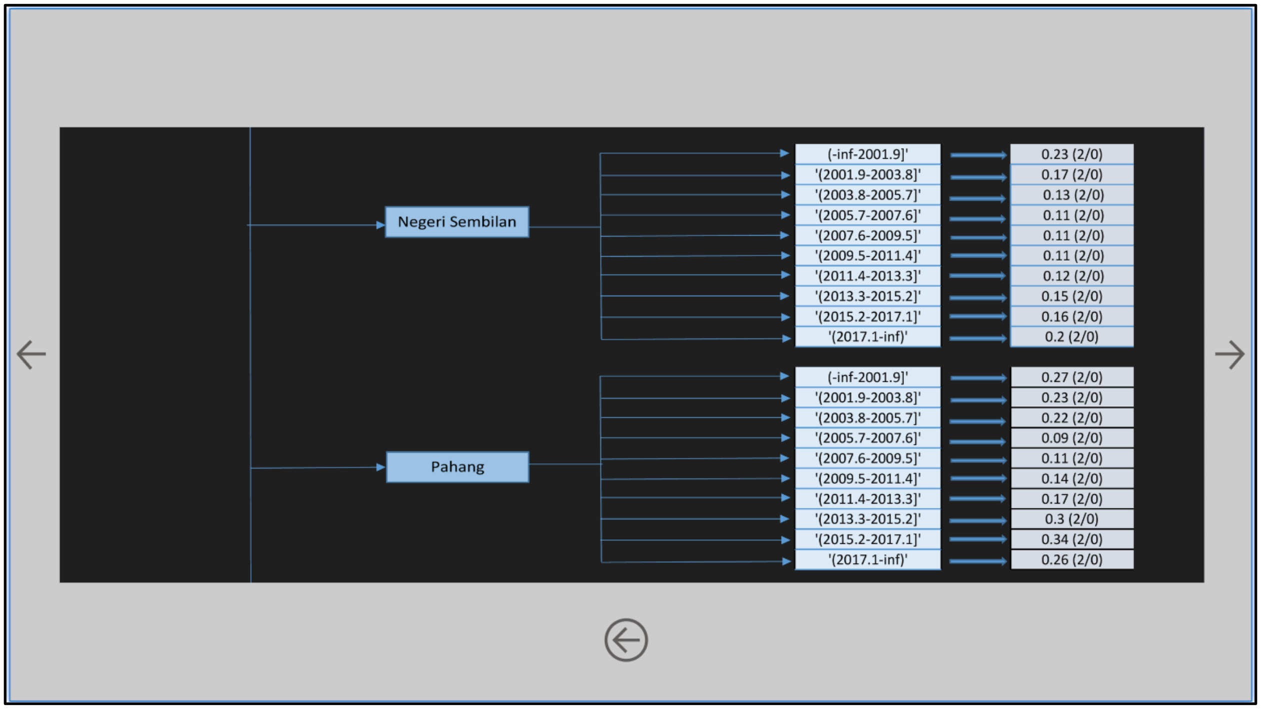
Perak

(-inf-2001.9]'	0.45 (2/0)
'(2001.9-2003.8]'	0.43 (2/0)
'(2003.8-2005.7]'	0.45 (2/0)
'(2005.7-2007.6]'	0.28 (2/0.01)
'(2007.6-2009.5]'	0.2 (2/0)
'(2009.5-2011.4]'	0.32 (2/0)
'(2011.4-2013.3]'	0.31 (2/0.01)
'(2013.3-2015.2]'	0.44 (2/0)
'(2015.2-2017.1]'	0.31 (2/0.01)
'(2017.1-inf)'	0.29 (2/0)

Perlis

(-inf-2001.9]'	0.02 (2/0)
'(2001.9-2003.8]'	0.07 (2/0)
'(2003.8-2005.7]'	0.06 (2/0)
'(2005.7-2007.6]'	0.03 (2/0)
'(2007.6-2009.5]'	0.04 (2/0)
'(2009.5-2011.4]'	0.03 (2/0)
'(2011.4-2013.3]'	0.04 (2/0)
'(2013.3-2015.2]'	0.1 (2/0)
'(2015.2-2017.1]'	0.1 (2/0)
'(2017.1-inf)'	0.07 (2/0)

Figure 4.5.4 The Decision Tree Predicted Value of Perak and Perlis's State



Negeri Sembilan

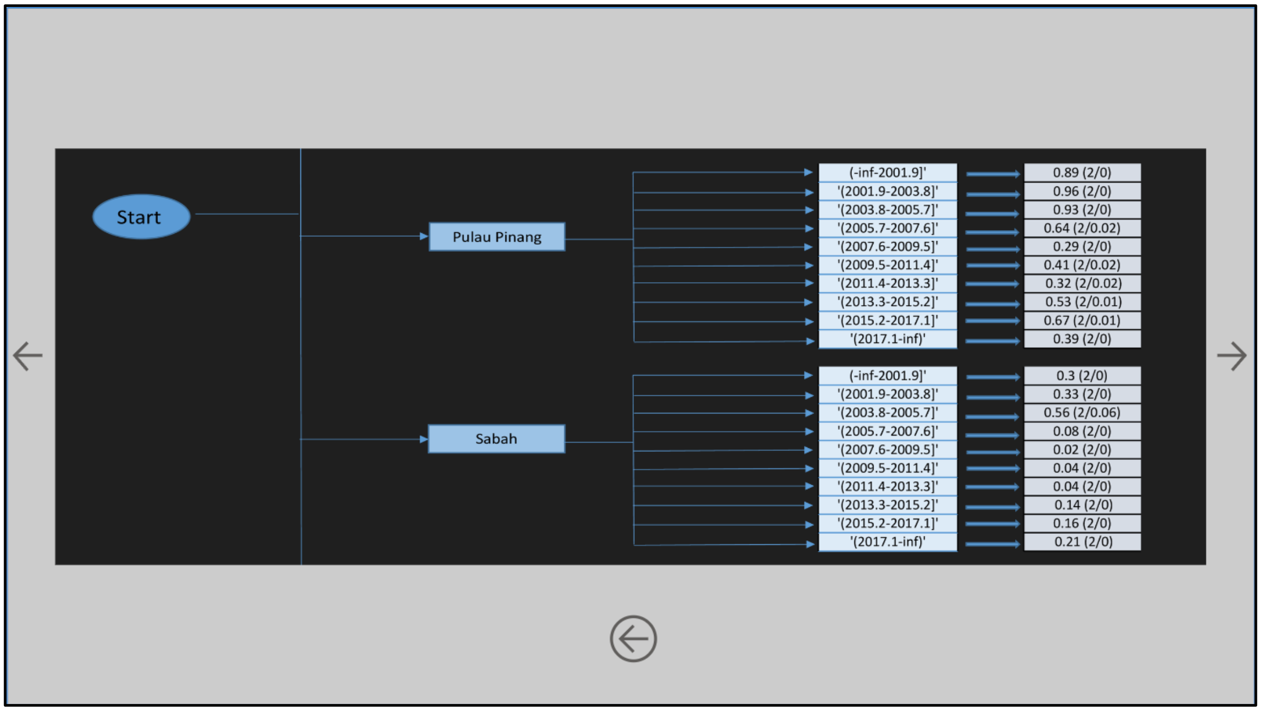
(-inf-2001.9]'	0.23 (2/0)
'(2001.9-2003.8]'	0.17 (2/0)
'(2003.8-2005.7]'	0.13 (2/0)
'(2005.7-2007.6]'	0.11 (2/0)
'(2007.6-2009.5]'	0.11 (2/0)
'(2009.5-2011.4]'	0.11 (2/0)
'(2011.4-2013.3]'	0.12 (2/0)
'(2013.3-2015.2]'	0.15 (2/0)
'(2015.2-2017.1]'	0.16 (2/0)
'(2017.1-inf)'	0.2 (2/0)

Pahang

(-inf-2001.9]'	0.27 (2/0)
'(2001.9-2003.8]'	0.23 (2/0)
'(2003.8-2005.7]'	0.22 (2/0)
'(2005.7-2007.6]'	0.09 (2/0)
'(2007.6-2009.5]'	0.11 (2/0)
'(2009.5-2011.4]'	0.14 (2/0)
'(2011.4-2013.3]'	0.17 (2/0)
'(2013.3-2015.2]'	0.3 (2/0)
'(2015.2-2017.1]'	0.34 (2/0)
'(2017.1-inf)'	0.26 (2/0)

Figure 4.5.5 The Decision Tree Predicted Value of Negeri Sembilan and Pahang's State

The Decision Tree 3 pages show the ninth, tenth, eleventh and twelfth branches in the Random Tree model, which represent the state of Pulau Pinang, Sabah, Sarawak and Selangor. For the ninth branch, the value predicted is quite interesting. This began in the year (-inf-2001.9) until the year (2003.8-2005.7), Pulau Pinang reached a very high predicted drug abusers value from 0.89 to 0.96, then down to 0.93. After the high predicted value was reached, it fell very much to a value of 0.29 in the year (2007.6-2009.5). The predicted value of Random Tree afterwards on Pulau Pinang scored between 0.41 and 0.67 except for the last year (2017.1-inf), which is recorded as quite low, which is 0.39. Sabah's state is the tenth branch. Sabah also has an unstable predicted value. The highest recorded value is for the years (2003.8-2005.7), which is 0.56, while the lowest hit for the year (2007.6-2009.5) gets a value of 0.02. The lower value stayed on until the years (2011.4-2013.3), then it began to steadily increase until the end with the value stated as 0.21. This is different from its neighbourhood state, Sarawak, which is on the eleventh branch. Sarawak got a steadily low predicted value throughout the year, from 0.04 to 0.12, at its highest in the last year, which is (2017.1-inf). The twelfth branch is the Selangor state. Selangor, which is also stated as the third highest in the datasets before the preprocessing, got below the average predicted value. The highest is recorded in the year (-inf-2001.9), which is 0.55. The next year is also on the medium value, which is 0.54. The value then decreased steadily until the year (2009.5-2011.4), rose up to 0.37, and then declined steadily again until the year (2015.2-2017.1), stating the predicted value at 0.42, but this value is also below the medium value, which is 0.50.



Start

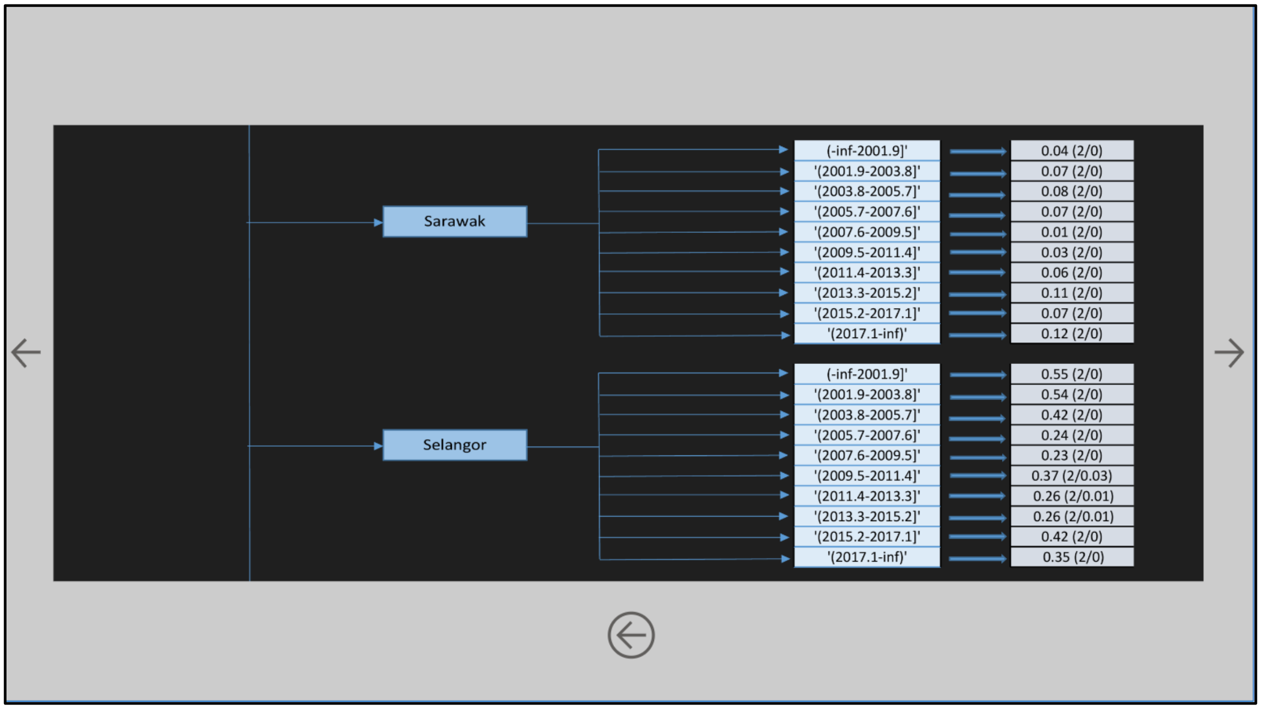
Pulau Pinang

(-inf-2001.9]'	0.89 (2/0)
'(2001.9-2003.8]'	0.96 (2/0)
'(2003.8-2005.7]'	0.93 (2/0)
'(2005.7-2007.6]'	0.64 (2/0.02)
'(2007.6-2009.5]'	0.29 (2/0)
'(2009.5-2011.4]'	0.41 (2/0.02)
'(2011.4-2013.3]'	0.32 (2/0.02)
'(2013.3-2015.2]'	0.53 (2/0.01)
'(2015.2-2017.1]'	0.67 (2/0.01)
'(2017.1-inf)'	0.39 (2/0)

Sabah

(-inf-2001.9]'	0.3 (2/0)
'(2001.9-2003.8]'	0.33 (2/0)
'(2003.8-2005.7]'	0.56 (2/0.06)
'(2005.7-2007.6]'	0.08 (2/0)
'(2007.6-2009.5]'	0.02 (2/0)
'(2009.5-2011.4]'	0.04 (2/0)
'(2011.4-2013.3]'	0.04 (2/0)
'(2013.3-2015.2]'	0.14 (2/0)
'(2015.2-2017.1]'	0.16 (2/0)
'(2017.1-inf)'	0.21 (2/0)

Figure 4.5.6 The Decision Tree Predicted Value of Pulau Pinang and Sabah's State



Sarawak

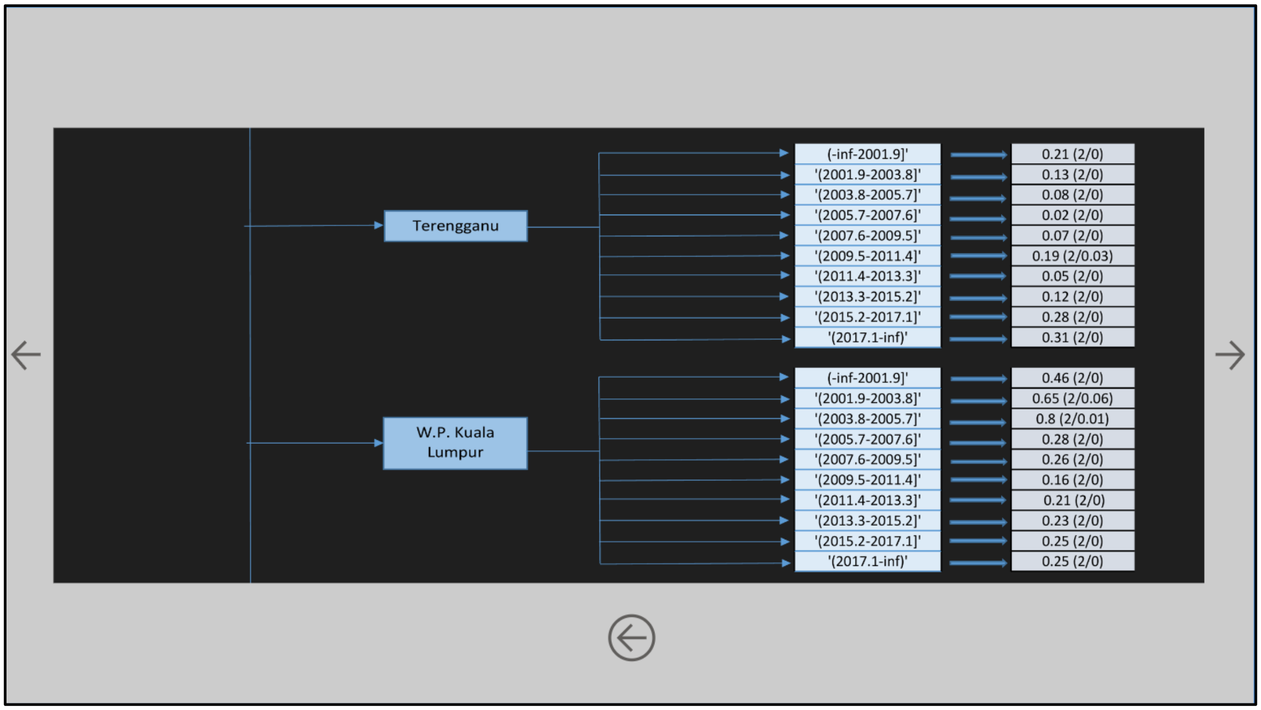
(-inf-2001.9]'	0.04 (2/0)
'(2001.9-2003.8]'	0.07 (2/0)
'(2003.8-2005.7]'	0.08 (2/0)
'(2005.7-2007.6]'	0.07 (2/0)
'(2007.6-2009.5]'	0.01 (2/0)
'(2009.5-2011.4]'	0.03 (2/0)
'(2011.4-2013.3]'	0.06 (2/0)
'(2013.3-2015.2]'	0.11 (2/0)
'(2015.2-2017.1]'	0.07 (2/0)
'(2017.1-inf)'	0.12 (2/0)

Selangor

(-inf-2001.9]'	0.55 (2/0)
'(2001.9-2003.8]'	0.54 (2/0)
'(2003.8-2005.7]'	0.42 (2/0)
'(2005.7-2007.6]'	0.24 (2/0)
'(2007.6-2009.5]'	0.23 (2/0)
'(2009.5-2011.4]'	0.37 (2/0.03)
'(2011.4-2013.3]'	0.26 (2/0.01)
'(2013.3-2015.2]'	0.26 (2/0.01)
'(2015.2-2017.1]'	0.42 (2/0)
'(2017.1-inf)'	0.35 (2/0)

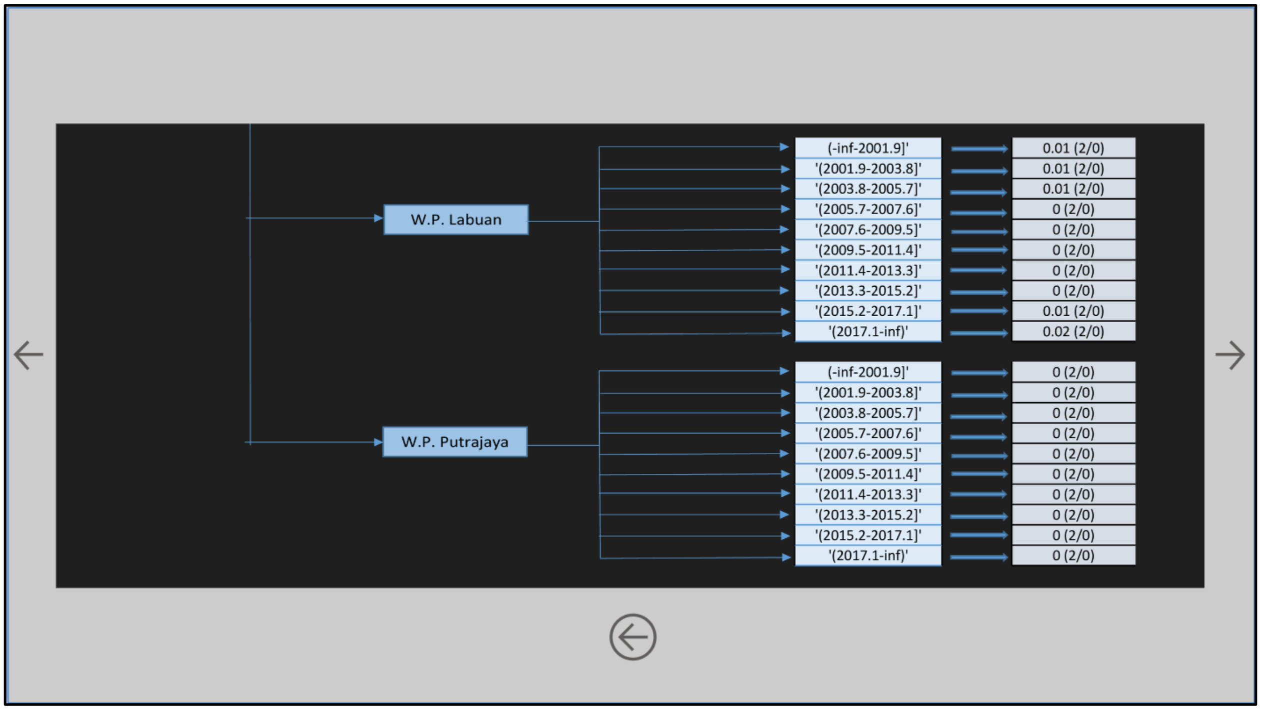
Figure 4.5.7 The Decision Tree Predicted Value of Sarawak and Selangor's State

The last four branches are the thirteenth branch, Terengganu, Wilayah Persekutuan Kuala Lumpur, the fourteenth's branch, Wilayah Persekutuan Labuan is the fifteenth's branch, and Wilayah Persekutuan Putrajaya is the sixteenth's branch. The highest predicted value recorded in Terengganu is in the year (2017.1-inf), which is 0.31, while the lowest is in the year (2005.7-2007.6). The Terengganu state's trend decreased steadily from the year (-inf-2001.9) until the year (2007.6-2009.5), then rose quite a bit high in the next year, predicted at 0.19. The next four discretized years are full of down and growing up steadily. Wilayah Persekutuan Kuala Lumpur stated high value of the predicted number of drug abusers n the first three years, which are 0.46, 0.65 and 0.80 on year (-inf-2001.9), (2001.9-2003.8) and (2003.8-2005.7) respectively. Then it has begun to decline drastically until 0.28 in the next year which is (2005.7-2007.6). From there, the Random Tree predicted the value would be steadily to be around 0.26 until 0.25. The lowest stated is on year (2009.5-2011.4), which is 0.16. Next is the Wilayah Persekutuan Labuan, which recorded the second lowest predicted values among the sixteens branches in the Random Tree Model. Wilayah Persekutuan Labuan only scored between the value of 0.00 and 0.02 at its highest. The predicted value started at 0.01 in the first year and stayed steady until the third year, which was (2003.8-2005.7). The value reached 0.00 in the next year, and it stayed until the year (2013.3-2015.2). Years (2015.2-2017.1) and (2017.1-inf) only have an increment of only 0.01, respectively, on both years. Lastly, in the Wilayah Persekutuan Putrajaya's branch, the Random Tree made shocking values of 0.00 predicted throughout the entire discretized years. This is because the Random Tree model only predicts the value on two decimal places value. A value lower than two decimal places will be considered as 0.



Terengganu	(-inf-2001.9]'	0.21 (2/0)
	'(2001.9-2003.8]'	0.13 (2/0)
	'(2003.8-2005.7]'	0.08 (2/0)
	'(2005.7-2007.6]'	0.02 (2/0)
	'(2007.6-2009.5]'	0.07 (2/0)
	'(2009.5-2011.4]'	0.19 (2/0.03)
	'(2011.4-2013.3]'	0.05 (2/0)
	'(2013.3-2015.2]'	0.12 (2/0)
	'(2015.2-2017.1]'	0.28 (2/0)
	'(2017.1-inf)'	0.31 (2/0)
W.P. Kuala Lumpur	(-inf-2001.9]'	0.46 (2/0)
	'(2001.9-2003.8]'	0.65 (2/0.06)
	'(2003.8-2005.7]'	0.8 (2/0.01)
	'(2005.7-2007.6]'	0.28 (2/0)
	'(2007.6-2009.5]'	0.26 (2/0)
	'(2009.5-2011.4]'	0.16 (2/0)
	'(2011.4-2013.3]'	0.21 (2/0)
	'(2013.3-2015.2]'	0.23 (2/0)
	'(2015.2-2017.1]'	0.25 (2/0)
	'(2017.1-inf)'	0.25 (2/0)

Figure 4.5.8 The Decision Tree Predicted Value of Terengganu and Wilayah Persekutuan Kuala Lumpur's State



W.P. Labuan	(-inf-2001.9]'	0.01 (2/0)
	'(2001.9-2003.8]'	0.01 (2/0)
	'(2003.8-2005.7]'	0.01 (2/0)
	'(2005.7-2007.6]'	0 (2/0)
	'(2007.6-2009.5]'	0 (2/0)
	'(2009.5-2011.4]'	0 (2/0)
	'(2011.4-2013.3]'	0 (2/0)
	'(2013.3-2015.2]'	0 (2/0)
	'(2015.2-2017.1]'	0.01 (2/0)
	'(2017.1-inf)'	0.02 (2/0)
W.P. Putrajaya	(-inf-2001.9]'	0 (2/0)
	'(2001.9-2003.8]'	0 (2/0)
	'(2003.8-2005.7]'	0 (2/0)
	'(2005.7-2007.6]'	0 (2/0)
	'(2007.6-2009.5]'	0 (2/0)
	'(2009.5-2011.4]'	0 (2/0)
	'(2011.4-2013.3]'	0 (2/0)
	'(2013.3-2015.2]'	0 (2/0)
	'(2015.2-2017.1]'	0 (2/0)
	'(2017.1-inf)'	0 (2/0)

Figure 4.5.9 The Decision Tree Predicted Value of Wilayah Persekutuan Labuan and Wilayah Persekutuan Putrajaya's State

In the Linear Regression model, the prediction result is displayed on a stacked bar chart for the ease of the user to see the result. This is because the stacked bar chart can stack the results on the X-axis, making it easier to make the comparison of each of the states in Malaysia based on the Y-axis label, which is the discretized year. Below the stacked bar chart is the clustered column chart. This chart is used in relationship with the below-stacked bar chart, based on the predicted value of the state that the user clicked on. For example, if the user wishes to see the predicted value of Johor, then the clustered column chart will be displayed based on the total value predicted on all states with each of the discretized years. On this, the user can be seen how much the value of predicted on the total value of predicted for that state in that certain discretized year.

The state sorted is based on the ascending order of the alphabet, starting with Johor, Kedah, Kelantan, Melaka, Negeri Sembilan, Pahang, Perak, Perlis, Pulau Pinang, Sabah, Sarawak, Selangor, Terengganu, Wilayah Persekutuan Kuala Lumpur, Wilayah Persekutuan Labuan and lastly the Wilayah Persekutuan Putrajaya. The result of the predicted value on the stacked bar chart on the Y-axis is sorted based on the highest total value predicted in all states. In the year (2003.8-2005.7), the predicted value is 10.6, followed by the year (2001.9-2003.8), which gets a value of 10.4, while the year (-inf-2001.9) is 9.0. The fourth highest predicted value is in the year (2015.2-2017.1), which is 8.7, followed by the year (2017.1-inf), which is 7.6 and the year (2013.3-2015.2), which is 7.5. Starting from the year (2005.7-2007.6) and year (2009.5-2011.4), the value seems to fall on the value between 5.3 and 5.1. Lastly, the two with the least predicted values are the year (2007.6-2009.5) and (2011.4-2013.3), which recorded values between 4.7 and 4.6, which is only a differences value of 0.1.

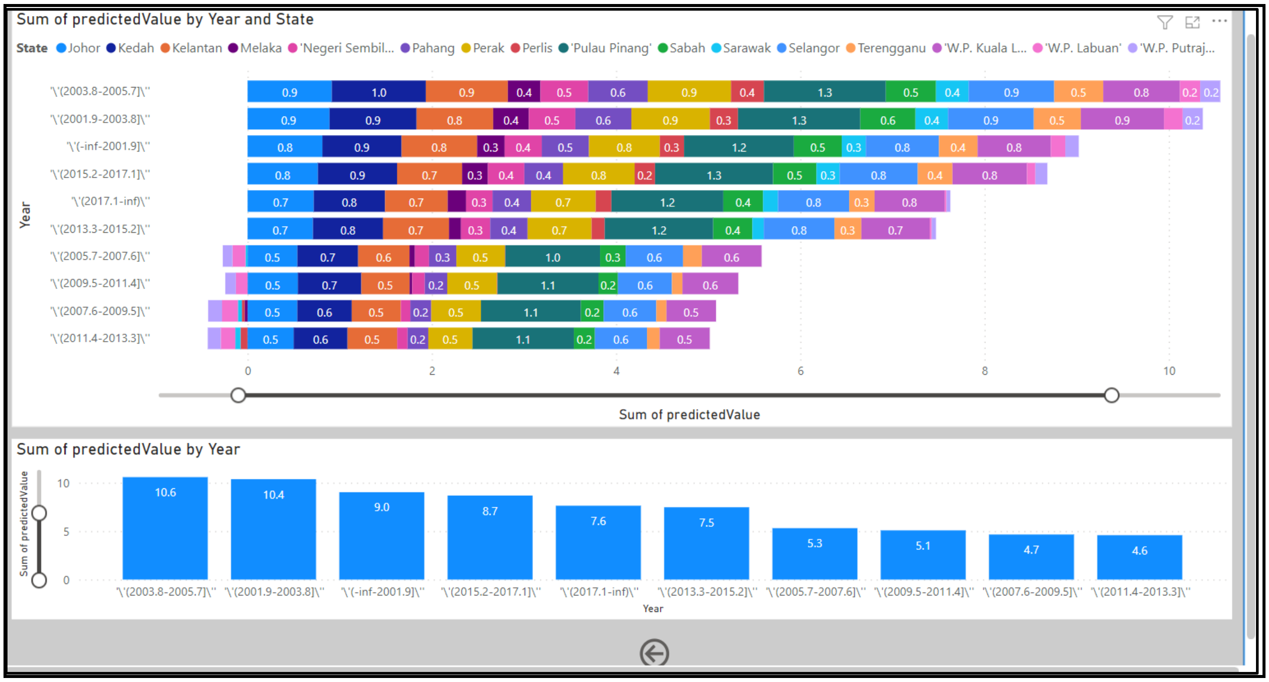


Figure 4.5.10 The Predicted Value of Linear Regression Model on Overall States

The state with the highest value predicted for drug abusers is Pulau Pinang, stating an average value predicted of 1.2 on throughout the year. Pulau Pinang recorded the highest value in the year (2003.8-2005.7) and (2001.9-2003.8), which is 1.3 valued for both years, the highest recorded in the overall states. The total value predicted for this state is 11.83 throughout the year. Pulau Pinang dominates the highest predicted value for all states in Malaysia, which is 1.33 in the year (2001.9-2003.8). The second state that reached the value of 7.74 total predicted throughout the year is Kedah. The highest predicted value for Kedah is in the year (2003.8-2005.7), which is 1.02, while the lowest predicted value is in the year (2011.4-2013.3), a 0.58 predicted value for the record. Moved on to the next state Selangor. Selangor stated 7.39 for the total predicted value throughout the year. The highest value recorded which is 0.93 in both years (2001.9-2003.8) and (2003.8-2005.7), while the lowest recorded value in the year (2007.6-2009.5) and (2011.4-2013.3), is 0.57. Wilayah Persekutuan Kuala Lumpur is next to the Selangor in predicted value. This state collected a score of 7.18 predicted value, which is its highest in the year (2001.9-2003.8) at 0.90, while the lowest is in the year (2007.6-2009.5) and (2011.4-2013.3) stated at 0.54. Next is the Johor state. Johor's state has a total value of 6.93 predicted value. The highest is in the year

(2003.8-2005.7), which is 0.91, while the lowest is in the year (2011.4-2013.3), recorded for a predicted value of 0.50. After Johor is the Kelantan state. The Kelantan state has a total predicted value of 6.80. Kelantan recorded its highest value in the year (2003.8-2005.7) at the value of 0.89, while in the year (2009.5-2011.4), it is in the lowest predicted value, which was 0.52. Perak is in seventh place in the table of highest drug abusers predicted in the Linear Regression model. The model predicts a Perak state of 6.78, which is only a 0.02 difference from Kelantan. In a year (2003.8-2005.7), Perak had a predicted value of 0.90, which is the highest throughout the year for Perak's state, while the lowest recorded is in the year (2011.4-2013.3), that get a predicted value of 0.48. In eighth place is the Pahang states. The Pahang has a total predicted value of 4.01 in all discretized years. The Pahang began its value by a small amount of increment from the year (-inf-2001.9) until (2003.8-2005.7). Then after that, its value continuing decreases until the year (2011.4-2013.3). During that period, the value continues to linger around 0.30 to 0.22 at its lowest. This contradicts Sabah's state. The Linear Regression model predicts Sabah's state at around the predicted value of 0.54 to 0.23. The highest predicted value is in the year (2003.8-2005.7), while the lowest is in the year (2009.5-2011.4). Starting from Terengganu, the total predicted value has begun to value below 3.00. Terengganu's total predicted value throughout the year is 2.96, with only 0.03 differences from Negeri Sembilan, which is 2.93. Terengganu's highest predicted value is 0.54, while the Negeri Sembilan is 0.52, both stated in the year (2003.8-2005.7). The lowest predicted value in Terengganu is in two years, which are the year (2007.6-2009.5) until (2009.5-2011.4), stated the value at 0.11, while in Negeri Sembilan, it is stated in the year (2007.6-2009.5) for the record of 0.10. There are three states that have the total predicted value throughout the year between 1.80 and 1.30, which are Melaka, Perlis and Sarawak. These three states have the value of 1.71, 1.38 and 1.37, respectively. These three states also have begun to get a negative value in the Linear Regression model, which is considered faulty. This is because there is no negative value in the original dataset from the government. The negative value in Melaka is in the year (2007.6-2009.5), which is at -0.03. In Perlis, the negative value recorded in the year (2007.6-2009.5) and (2011.4-2013.3), which are -0.03 and -0.07, while in Sarawak, the negative value began in years (2005.7-2007.6) and (2007.6-2009.5) at the value of -0.02 and -0.04 but then the predicted value seems to be positive again until the year (2011.4-2013.3), that stated the lowest negative value at -0.06. Lastly, for the Wilayah Persekutuan Putrajaya and Wilayah Persekutuan Labuan, these two states recorded the total predicted value throughout the year are only 0.32 and 0.11, respectively. These two states also have much more negative values than Sarawak, Perlis and Melaka, which are in four discretized years. Both states the predicted negative values in the year (2005.7-2007.6), (2007.6-2009.5), (2009.5-2011.4), and (2011.4-2013.3). Wilayah Persekutuan Putrajaya stated a total of -0.51, while the other one is much lower, which is -0.60.

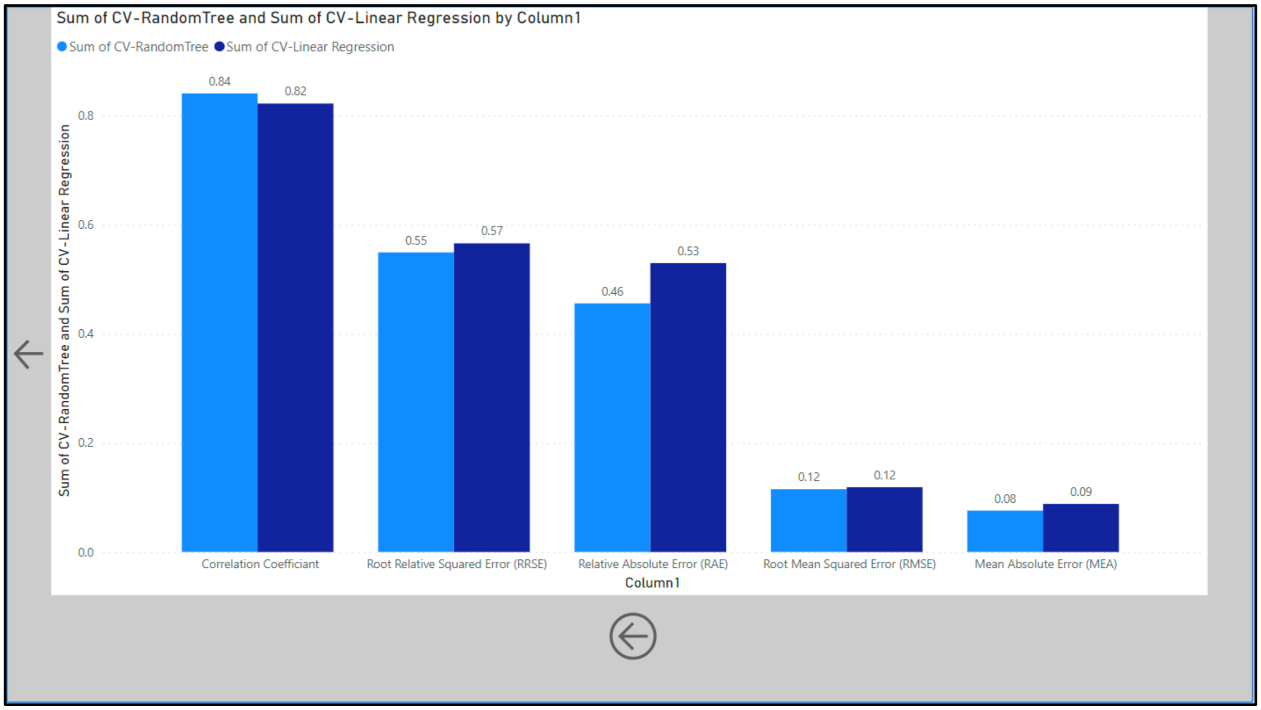


Figure 4.5.11 The Comparison Between Two Models

The last page for the Power BI dashboard is the comparison of the value between the Linear Regression and the Random Tree models. There are four comparisons being made, which are the Correlation Coefficient, Root Relative Squared Error (RRSP), Relative Absolute Error (RAE), Root Mean Squared Error (RMSE), and lastly, the Mean Absolute Error (MAE). The value is displayed in two decimal points for all the comparisons. In the previous, the RAE and RRSE were being compared in the percentage value, but in this dashboard, these values were being compared in two decimal places values because of the Power BI integration system. All these values also are being rounded up. This can be seen in the Weka; the Correlation Coefficient for both models are

0.8225 for Linear Regression and 0.8441 for Random Tree, but in Power BI, the value changed to 0.84 for Random Tree and 0.82 for Linear Regression. The RRSE also changed from 56.6143 % for Linear Regression and 54.9566% for Random Tree to the value of 0.55 for the Random Tree and 0.57 for Linear Regression. The RAE value also changed from 52.9875% to 0.53 in the Linear Regression model and from 45.6042% to 0.46 in the Random Tree model. Meanwhile, for the RMSE value, the value changed from 0.1188 to 0.12 in Linear Regression and from 0.1153 to 0.12 also in Random Tree. Lastly, for the MEA value, the changes in Linear Regression are from 0.0883 to 0.09, and for the Random Tree, the value is from 0.0760 to 0.08.

CHAPTER 5

CONCLUSION

5.1 Introduction

This chapter will discuss objective revisits, project limitations, and the future work of the Prediction of Drug Abused Before The Quarantine Of The Pandemic COVID-19 In Malaysia (2000-2019).

5.2 Objective Revisited

During the work process of the Drug Abused Before The Quarantine of The Pandemic COVID-19 In Malaysia (2000-2019), there three objectives are being focused on. The first objective is to study the trend of the drug abusers on Malaysian before the pandemic COVID-19 from year 2000 until 2019. From this study, there are multiple trends that have been identified and analyzed, and the result of the analysis is being used as a guide to implementing machine learning in this thesis.

The second objective is to predict the usage of the drug by Malaysian during the COVID-19 pandemic by detecting earlier the potential for a future pandemic. This study has given the opportunity to its reader to let know what an upcoming trend for drug abuse in Malaysia based on the prediction of machine learning by using the unsupervised algorithm to solve the regression problem.

The last objective is to visualize the yield of study and raise awareness of Malaysians on health issues before the pandemic of COVID-19. This is really important and acts as a stepping stone's throw away to let Malaysians know about the trend of drug abuse, whether it has been increasing or decreasing during the past twenty years. This trend can be used to make a prediction for upcoming future works.

5.3 Limitations

During the process of completing this thesis, there are a few limitations must be face, such as:

I. The limitation of resources and knowledge about the machine learning whole process from the start until its finished

II. The verified and correct use of datasets within the thesis’s objectives and requirements

III. The unused of advance machine learning is hard to learn and understand without guidance from the industry’s worker

5.4 Future Works

This thesis has the potential that can be used as a guide for the upcoming thesis studies about data mining and machine learning in the future. In the future, there will be more studies to be made based on machine learning, whether it is implemented in the basic or the advanced machine learning models. These theses can do this thesis to be more completed and accurate. This thesis also can be referred to the beginner that wants to explore the data mining and machine learning algorithm that is becoming a hot topic nowadays and might be much more in the future.

REFERENCES

Elengoe, A. (2020). COVID-19 Outbreak in Malaysia. Osong Public Health and Research Perspectives, 11(3), 93–100. https://doi.org/10.24171/j.phrp.2020.11.3.08

Abramson, A. (2021, March 1). Substance use during the pandemic. American Psychology Association. Retrieved October 20, 2021, from https://www.apa.org/monitor/2021/03/substance-use-pandemic

World Health Organization. (2020, April 28). Archived: WHO Timeline - COVID-19. Who. Retrieved October 21, 2021, from https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

© 2018 Agensi Antidadah Kebangsaan Malaysia. (2021, October 19). Drugs Statistics. National Anti-Drugs Agency. Retrieved October 19, 2021, from https://www.adk.gov.my/en/public/drugs-statistics/#tabs_v2-paneeluide4547833_1_9

Zaami, S., Marinelli, E., & Varì, M. R. (2020). New Trends of Substance Abuse During COVID-19 Pandemic: An International Perspective. Frontiers in Psychiatry, 11. https://doi.org/10.3389/fpsyt.2020.00700

Bach, B. (2017, December 20). New method of predicting drug abuse shows promise. Scope By Stanford Medicine. Retrieved October 22, 2021, from https://scopeblog.stanford.edu/2017/02/23/new-method-of-predicting-drug-abuse-shows-promise/

Idrus, P. G. (2021, July 3). Suicide rising in Malaysia due to hardships amid coronavirus pandemic. Anadolu Agency. Retrieved October 22, 2021, from https://www.aa.com.tr/en/asia-pacific/suicide-rising-in-malaysia-due-to-hardships-amid-coronavirus-pandemic/2293079

Panchal, N., Kamal, R., Cox, C., & Garfield, R. (2021, July 20). The Implications of COVID-19 for Mental Health and Substance Use. KFF. Retrieved October 22, 2021, from https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/#:%7E:text=During%20the%20COVID%2D19%20pandemic,largely%20stable%20since%20spring%202020.

Stedman, C., & Hughes, A. (2021, September 7). data mining. SearchBusinessAnalytics. Retrieved November 29, 2021, from https://searchbusinessanalytics.techtarget.com/definition/data-mining?utm_source=youtube&utm_medium=video&utm_campaign=032021dataminingutm_content%253Ddatamining%253FOffer%253DOTHR-youtube_OTHR-video_OTHR-datamining_2021March10_datamining

World Health Organization (WHO). (2020, April 28). Archived: WHO Timeline - COVID-19. Retrieved November 29, 2021, from https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

STAT. (2020, January 13). Woman in Thailand is first case with novel pneumonia virus outside China. Retrieved November 29, 2021, from https://www.statnews.com/2020/01/13/woman-with-novel-pneumonia-virus-hospitalized-in-thailand-the-first-case-outside-china/

Chelvan, V. P. (2021, July 25). Malaysia’s total COVID-19 caseload passes 1 million mark. CNA. Retrieved November 30, 2021, from https://www.channelnewsasia.com/asia/1-million-covid-19-cases-malaysia-jul-25-2065386

Shah, A. U. M., Safri, S. N. A., Thevadas, R., Noordin, N. K., Rahman, A. A., Sekawi, Z., Ideris, A., & Sultan, M. T. H. (2020). COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government. International Journal of Infectious Diseases, 97, 108–116. https://doi.org/10.1016/j.ijid.2020.05.093

Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. The Lancet, 395(10223), 470–473. https://doi.org/10.1016/s0140-6736(20)30185-9

Singhal, T. (2020). A Review of Coronavirus Disease-2019 (COVID-19). The Indian Journal of Pediatrics, 87(4), 281–286. https://doi.org/10.1007/s12098-020-03263-6

Ministry Of Health Malaysia. (n.d.). in Malaysia. COVIDNOW. Retrieved December 4, 2021, from https://covidnow.moh.gov.my/

Wikipedia contributors. (2021, December 19). Malaysian movement control order. Wikipedia. Retrieved December 10, 2021, from https://en.wikipedia.org/wiki/Malaysian_movement_control_order

Ying, T. P. (2021, November 7). Malaysia confirms first two cases of Covid-19 Delta Plus variant. NST Online. Retrieved December 10, 2021, from https://www.nst.com.my/news/nation/2021/11/743067/malaysia-confirms-first-two-cases-covid-19-delta-plus-variant

Coronavirus Disease 2019 (COVID-19). (2020, February 11). Centers for Disease Control and Prevention. Retrieved December 10, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/variants/about-variants.html

Data Science Process Alliance. (2021, August 24). CRISP-DM. Retrieved January 1, 2022, from https://www.datascience-pm.com/crisp-dm-2/

Smart Vision Europe. (2020, June 17). Crisp DM methodology. Retrieved January 1, 2022, from https://www.sv-europe.com/crisp-dm-methodology/

javatpoint. (n.d.). Classification Algorithm in Machine Learning - Javatpoint. Www.Javatpoint.Com. Retrieved January 3, 2022, from https://www.javatpoint.com/classification-algorithm-in-machine-learning

GeeksforGeeks. (2020, September 28). Advantages and Disadvantages of different Classification Models. Retrieved January 3, 2022, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-classification-models/

GeeksforGeeks. (2021, October 20). Supervised and Unsupervised learning. Retrieved January 3, 2022, from https://www.geeksforgeeks.org/supervised-unsupervised-learning/?ref=leftbar-rightbar

Jamil, J. M., & Shaharanee, I. N. M. (2016). An innovative forecasting and dashboard system for Malaysian dengue trends. AIP Conference Proceedings. https://doi.org/10.1063/1.4960910 **https://repo.uum.edu.my/id/eprint/20522/1/JTEC%208%2010%202016%209%2012.pdf**

Najafi-Ghobadi, S., Najafi-Ghobadi, K., Tapak, L., & Aghaei, A. (2019). Application of data mining techniques and logistic regression to model drug use transition to injection: a case study in drug use treatment centers in Kermanshah Province, Iran. Substance Abuse Treatment, Prevention, and Policy, 14(1). https://doi.org/10.1186/s13011-019-0242-1

University of Regina DBD. (n.d.). KDD Process/Overview. http://www2.cs.uregina.ca/%7Edbd/cs831/notes/kdd/1_kdd.html

Hazarhun, E. H. (2022, October 7). KDD vs CRISP-DM - EcemHazarhun. Medium. Retrieved December 5, 2022, from https://medium.com/@ecemhazarhun/kdd-vs-crisp-dm-f7b8ea99640

Tamboli, N. (2022, November 30). All You Need To Know About Different Types Of Missing Data Values And How To Handle It. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

Karan, R. K. (2021, December 6). Knowledge Discovery in Databases (KDD) in Data Mining. Naukri Learning. Retrieved December 20, 2022, from https://www.naukri.com/learning/articles/kdd-in-data-mining/

Tim, M. (2022, December 10). Discretization Methods (Data Mining). Microsoft Learn. Retrieved December 30, 2022, from https://learn.microsoft.com/en-us/analysis-services/data-mining/discretization-methods-data-mining?view=asallproducts-allversions

GeeksforGeeks. (2019, June 25). Data Normalization in Data Mining. https://www.geeksforgeeks.org/data-normalization-in-data-mining/?ref=lbp

Tolety, K. (2023, January 13). Data Normalization Techniques in Data Mining Simplified 101. Learn | Hevo. https://hevodata.com/learn/normalization-techniques-in-data-mining/

Jason Brownlee. (2019, August 22). How To Estimate The Performance of Machine Learning Algorithms in Weka. Machine Learning Mastery. Retrieved December 17, 2022, from https://machinelearningmastery.com/estimate-performance-machine-learning-algorithms-weka/

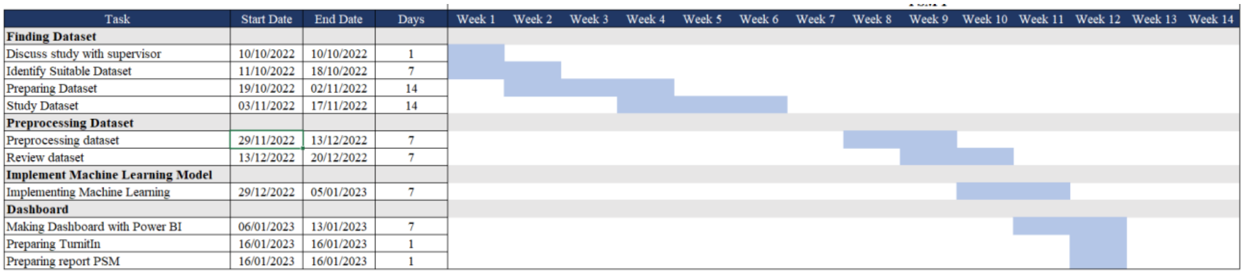
Ridzuan, F., & Wan Zainon, W. M. N. (2019). A Review on Data Cleansing Methods for Big Data. Procedia Computer Science, 161, 731-738. https://doi.org/10.1016/j.procs.2019.11.177

GeeksforGeeks. (2019b, June 25). Data Normalization in Data Mining. https://www.geeksforgeeks.org/data-normalization-in-data-mining/

Discretization in data mining - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/discretization-in-data-mining

Fürnkranz, J., Chan, P. K., Craw, S., Sammut, C., Uther, W., Ratnaparkhi, A., Jin, X., Han, J., Yang, Y., Morik, K., Dorigo, M., Birattari, M., Stützle, T., Brazdil, P., Vilalta, R., Giraud-Carrier, C., Soares, C., Rissanen, J., Baxter, R. A., . . . De Raedt, L. (2011). Mean Absolute Error. Encyclopedia of Machine Learning, 652–652. https://doi.org/10.1007/978-0-387-30164-8_525

APPENDIX



Task	Start Date	End Date	Days	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14
Finding Dataset																	
Discuss study with supervisor	10/10/2022	10/10/2022	1	▨													
Identify Suitable Dataset	11/10/2022	18/10/2022	7	▨	▨												
Preparing Dataset	19/10/2022	02/11/2022	14		▨	▨	▨										
Study Dataset	03/11/2022	17/11/2022	14				▨	▨	▨								
Preprocessing Dataset																	
Preprocessing dataset	29/11/2022	13/12/2022	7								▨	▨					
Review dataset	13/12/2022	20/12/2022	7									▨	▨				
Implement Machine Learning Model																	
Implementing Machine Learning	29/12/2022	05/01/2023	7										▨	▨			
Dashboard																	
Making Dashboard with Power BI	06/01/2023	13/01/2023	7											▨	▨		
Preparing TurnitIn	16/01/2023	16/01/2023	1												▨		
Preparing report PSM	16/01/2023	16/01/2023	1												▨		

Gant Chart for entire process of study