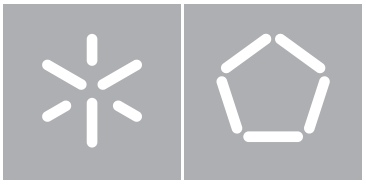
Universidade do Minho
Escola de Engenharia

Celso Filipe Nogueira Coutinho

A Data Mining approach towards effective Dengue outbreak prediction in Seremban, Malaysia

Outubro de 2015



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Celso Filipe Nogueira Coutinho

A Data Mining approach towards effective Dengue outbreak prediction in Seremban, Malaysia

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de
Professor Professor Manuel de Oliveira Orlando Belo
Professor Zalizah Awang Long

Outubro de 2015

Acknowledgements

This study was made possible only due to my family. The unconditional support of my parents, grandparents, brother and sister has always been what kept me going and to them I am especially grateful.

I would like to express my gratitude to my supervisors, Professor Orlando Belo for his support and guidance throughout my work and for always showing me the next step, and Professor Zalizah Awang Long for providing me with such an interesting subject.

I would also wish to thank everyone with whom I have worked as a volunteer in ESN Minho, for being the motivation I needed to do one step backwards in order to be able to do two steps forward. Without having been a member of ESN Minho I would not have the courage to get back to my studies and I would not have even started this dissertation.

I want to thank the UniKL international office, the international relations office in University of Minho and Professor Paulo Azevedo for all the help provided before, during and after my exchange in the University Kuala Lumpur, where I have started this research.

A thorough revision has been made to this document, and for that I must acknowledge the great contribution of Claire Rees. Not only she revised the whole document, but also provided me with good advice and a bit of motivation throughout most of the writing, and for that I am truly thankful.

Last, but not least, a word of appreciation to Frank Grieshaber for being the voice of reason in the early days of my dissertation, for his invaluable advice, and for everything that makes him one of the best friends someone can have.

Resumo

Uma abordagem de mineração de dados para a previsão eficaz de surtos de dengue

Na Malásia, a taxa de incidência de febre de dengue e febre hemorrágica de dengue atingiu o nível de epidemia, continuando os seus números a crescer. Nos últimos anos, um grande esforço tem sido empreendido no desenvolvimento de métodos para prever surtos de dengue, mas o caminho para realizar de forma eficaz essas previsões, e, portanto, salvar vidas humanas, é ainda muito longo. Este trabalho de dissertação incidiu no uso de técnicas de mineração de dados para descobrir padrões escondidos nos dados obtidos através do cruzamento de informação acerca de pacientes infectados com dengue na Malásia e dados meteorológicos relativos às áreas geográficas onde os pacientes foram infectados.

Palavras-Chave: Árvores de Decisão, Descoberta de Conhecimento em Bases de Dados, Febre de Dengue, Mineração de dados, Processo Padrão Inter-Indústrias para Mineração de Dados, Segmentação.

Abstract

A data mining approach towards effective dengue outbreak prediction

In Malaysia, the incidence rate of Dengue Fever and Dengue Haemorrhagic Fever has reached the level of epidemic, and its numbers keep growing. In the last few years, a big effort has been put into developing methods for predicting dengue outbreaks. However, the path for undertaking effectively those predictions, and therefore save Human lives, is still a very long one. This dissertation work focused on the use of Data Mining techniques, for discovering hidden patterns on data obtained by crossing information related to patients infected with dengue in Malaysia and meteorological data coming from the areas where those patients got infected.

Keywords: Clustering, Cross-Industry Standard Process for Data Mining, Data Mining, Decision Trees, Dengue Fever, Knowledge Discovery in Databases

Index

Index of Figures

Index of Tables

List of Symbols and Acronyms

AI	Artificial Intelligence
AIS	Artificial Immune System
ANN	Artificial Neural Network
AR	Association Rule
ARFF	Attribute-Relation File Format
CC	Correctly Classified
CP	Complexity Parameter
CRISP-DM	CRoss-Industry Standard Process for Data Mining
DB	Database
DF	Dengue Fever
DHF	Dengue Haemorrhagic Fever
DM	Data Mining
DSS	Dengue Shock Syndrome
DT	Decision Tree
FN	False Negative

FP	False Positive
GA	Genetic Algorithm
IBM	International Business Machines
ID	Identifier
KDD	Knowledge Discovery in Databases
MAV	Multiple Attribute Value
minConf	Minimum confidence
minSup	Minimum support
NCR	National Cash Register
NSA	Negative Selection Algorithm
PAM	Partitioning Around Medoids
PD	Port Dickson
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
RS	Rough Set
SE	SouthEast
SPSS	Statistical Package for the Social Sciences
TN	True Negative
TP	True Positive
UI	User Interface
UKM	*Universiti Kebangsaan Malaysia*

WHO	World Health Organisation

Chapter 1

1 Introduction

1.1 Contextualisation

Dengue Fever (DF) is a mosquito-borne viral infection caused by the dengue virus that occurs when a mosquito that is infected with the virus bites a person. The infection causes an illness similar to influenza (commonly known as flu), and can develop itself into a potentially lethal complication called *Dengue Haemorrhagic Fever* (DHF), which is characterised by leaky blood vessels. It can also evolve to other diseases, namely dengue shock syndrome (DSS), encephalitis and hepatitis. "The clinical features of classical DF include fever, headache, retro-orbital pain, myalgias and arthralgias, nausea, vomiting, and often a rash. Some patients develop haemorrhagic manifestations, such as haematuria, bleeding gums, epistaxis, haematemesis, melaena, and ecchymosis. DHF patients develop thrombocytopenia and haemoconcentration. Some may progress into DSS, leading to profound shock and death if not properly treated" (Ghazali, et al., 2012). Since both DHF and DSS correspond to late stages of infection with dengue virus and no distinction between them will be made in the remaining of this document, from this moment onwards only references to DHF will be made, for the sake of simplicity.

Aedes aegypti, a highly resilient mosquito species, is the main vector of dengue virus. Although dengue outbreaks have also been attributed to *aedes albopictus*, *aedes polynesiensis* and several

species of the *aedes scutellaris* complex (World Health Organization, 2009), this introductory section will only present data related with the *aedes aegypti species*. This mosquito is a tropical and subtropical species widely distributed around the world, mostly between latitudes 35º N and 35º S, but has already been found as far north as 45º N. Also, because of lower temperatures, *aedes aegypti* is relatively uncommon above 1000 metres (World Health Organization, 2009). It has 4 serotypes[1], DENV-1, DENV-2, DENV-3 and DENV-4, whereas the existence of a fifth has been discussed (Campbell, et al., 2013; Normile, 2013). Biological and immunological criteria are used to distinguish between the different serotypes. The details about the characteristics of the dengue serotypes go beyond the scope of this work - further readings about this topic can be consulted in the following works (Balmaseda, et al., 2006; Caribbean Epidemiology Center, Pan American Health Organization and World Health Organization, 2001; Kalayanarooj and Nimmannitya, 2000).

Aedes aegypti, prefers to lay its eggs in artificial containers commonly found in and around homes, for example, flower vases, old automobile tires, buckets that collect rainwater, and trash in general. Containers used for water storage, such as 55-gallon drums, cement cisterns, and even septic tanks, are also used by *aedes aegypti* to lay its eggs, and large numbers of adult mosquitoes in close proximity to human dwellings are produced within them. *Aedes aegypti* eggs acquire resistance to drying very rapidly, only 15h after having been laid. From then on, they can withstand long periods of drought - up to 450 days, according to WHO studies. This resistance is a major advantage for the mosquito since it allows the eggs to survive for many months in a dry place, until the next rainy and warm period, which is conducive to the outbreak.

The entire lifecycle of the species, which takes between 8 and 10 days, is depicted in **Figure 1**. Its growth starts from an egg, transforming itself in larvae afterwards. The next stage is the pupae, and then it finally becomes an adult mosquito. An adult mosquito can live from 2 to 3 weeks, and only the females will bite to feed on blood, which is necessary for the production of eggs. On average, each female produces 3 to 4 batches of eggs in her lifetime, and about 70 to 80 eggs per batch. According to the WHO, an interesting fact about the dynamics of the disease is that people, rather than mosquitoes, rapidly move the virus within and between communities and places. That happens because most female *aedes aegypti* may spend their lifetime in or around the houses where they emerge as adults and they usually fly an average of 400 metres.

[1] Group of closely related organisms, microorganisms, or cells distinguished by a characteristic set of antigens.

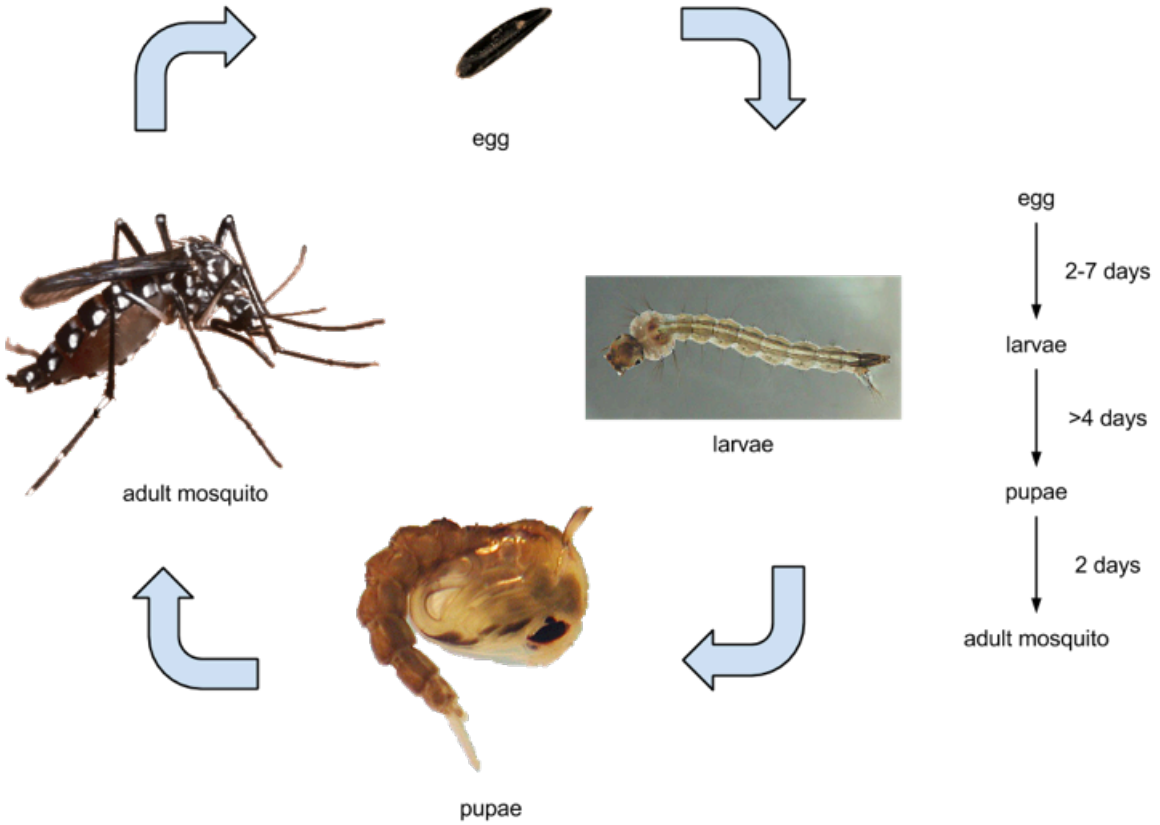


Figure 1 – Life cycle of *aedes aegypti* - based on (M., 2010)

Adult mosquitoes prefer to rest indoors. They are unobtrusive, and prefer to feed on humans during daylight hours. There are two peaks of biting activity, namely early morning for 2 to 3 hours after daybreak and in the afternoon for several hours before dark. However, these mosquitoes will feed all day indoors and on overcast days. The female mosquitoes are very nervous feeders, disrupting the feeding process at the slightest movement, only to return to the same or a different person to continue feeding moments later. Because of this behaviour, *aedes aegypti* females will often feed on several people during a single blood meal. When infective, they may transmit dengue virus to multiple people in a short time period, even if they only probe[2] without taking blood. It is not uncommon to see several members of the same household become ill with dengue

[2] With less than 5% of the skin being blood vessels, the mosquito searches for them by probing the skin.

within a 24- to 36-h time frame, suggesting that a single infective mosquito infected all of them. It is this behaviour that makes *aedes aegypti* such an efficient epidemic vector (Gubler, 1998; World Health Organization, 2009).

After an infective mosquito bites a person, the virus undergoes an (intrinsic) incubation period of 3 to 14 days (average, 4 to 7 days). After that, the person may experience acute onset of fever accompanied by a variety of nonspecific signs and symptoms. This acute febrile period may be as short as 2 days and as long as 10 days. If other a*edes aegypti* mosquitoes bite the ill person during this febrile viremic stage, those mosquitoes may become infected and subsequently transmit the virus to other uninfected people, after an extrinsic incubation period[3] of 8 to 12 days. The extrinsic incubation period is influenced in part by environmental conditions, especially ambient temperature. Thereafter, the mosquito remains infective for the rest of its life (Gubler, 1998; World Health Organization, 2009). Mosquitoes can also inherit the disease. If a female mosquito is infected with dengue, there is the possibility that its descendant larvae will already be born with the virus, which is called vertical transmission.

During the 19th century, dengue was an intermittent disease that caused epidemics[4] at long intervals, a reflection of the slow pace of transportation and limited travel at that time. Moreover, the disease pattern associated with dengue-like illnesses from 1780 to 1940 was characterised by relatively infrequent but often large epidemics. DF and DHF incidence escalated dramatically due to many factors, namely the rapid and unplanned urbanisation, increased human movement, and ineffective mosquito control. The combination of these factors in developing countries contributed to the expansion of DF and DHF in the urban centres, especially because of the increase of inadequate water storage and disposable containers, which are the ideal breeding sites for *aedes aegypti* mosquitoes. Nevertheless, it was especially due to the ecological disruption in the Southeast (SE) Asia and Pacific theatres during and following World War II that ideal conditions for increased transmission of mosquito-borne diseases were created, and it was in this setting that a global pandemic[5] of DF began, which led to the rise of DHF. The first known epidemic of DHF occurred in Manila, Philippines, between 1953 and 1954. But within 20 years the malady had spread throughout SE Asia. By the mid-1970s, DHF had become a leading cause of hospitalisation

[3] The extrinsic incubation period is the time the disease takes to develop in the *aedes aegypti* mosquito.
[4] An epidemic of a certain disease occurs in areas where its transmission reaches high levels and is then interrupted. For example, in a country where there is a dry season and no mosquitoes breed in a given time of year.
[5] Global epidemics.

and death among children in the region (Gubler, 1998; Wearing and Rohani, 2006; World Health Organization, 1991).

Today, according to *World Health Organisation* (WHO), dengue ranks as the most important arthropod-borne[6] viral disease in the world, and it been like that since 1983 (Focks, et al., 1993; Rosen, 1982). In the last 50 years incidence has increased 30-fold. An estimated 2.5 billion people live in over 100 endemic[7] countries and areas where dengue viruses can be transmitted. Up to 50 million infections occur annually with 500 thousand cases of DHF and 22 thousand deaths, mainly among children. Prior to 1970, only 9 countries had experienced cases of DHF. Since then the number has increased more than 4-fold and continues to rise (World Health Organization, 2009). Although these numbers are already frightening enough, some authors argue they fall short of the real picture. In (Bhatt, et al., 2013; Campbell, et al., 2013; Normile, 2013) the authors claim the yearly number of infections reaches 390 million, of which 96 million are manifested, far exceeding the WHO estimates in 2009. In (Normile, 2013) the author goes even further, stating that 10% of the DF cases evolve to DHF, contrary to the 1% (500 thousand of 50 million) presented by WHO.

In Malaysia, notwithstanding the government's effort towards sensitising the population on how to deal with DF and DHF, through several awareness campaigns and properly educating the community, the malady has reached the epidemic level. According to the BBC news, the number of people who have died from one of the dengue variants in Malaysia has more than tripled in the first 3 trimesters of 2014, compared to the same period in 2013, with 250 cases being reported daily. The following pictures show the growth of the malady in Malaysia throughout the years.

[6] Invertebrate animal that has an exoskeleton, a segmented body, and jointed limbs. This Includes insects, arachnids, myriapods and crustaceans.
[7] A disease is endemic when its transmission rates are somewhat stable over the year. In the case of dengue, it occurs in hot countries, with high humidity values, wherein mosquitoes breed throughout the whole year.

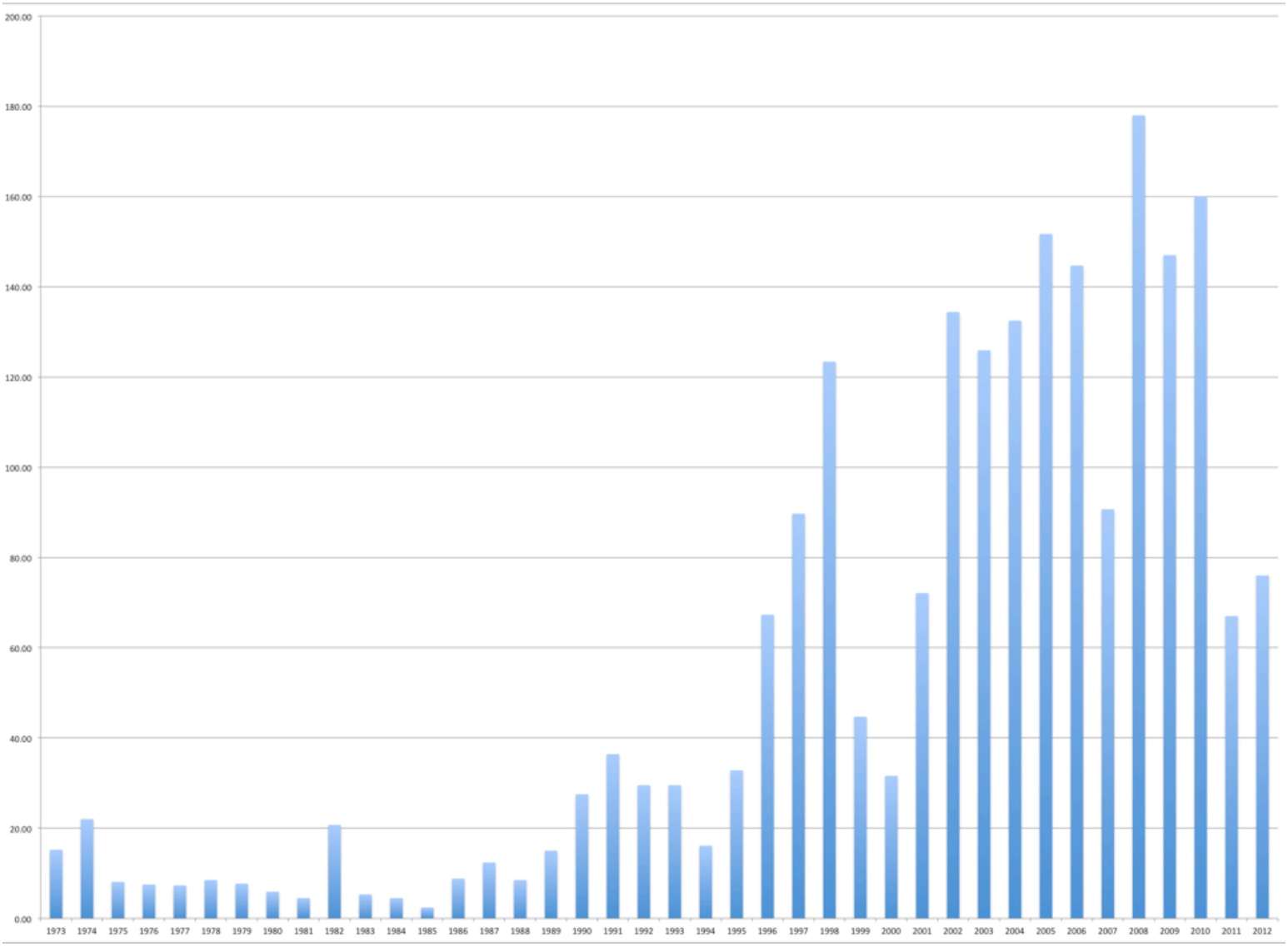


Figure 2 - Incidence rate of dengue in Malaysia per year per 100.000 inhabitants. Information extracted from (Ministry of health Malaysia, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012)

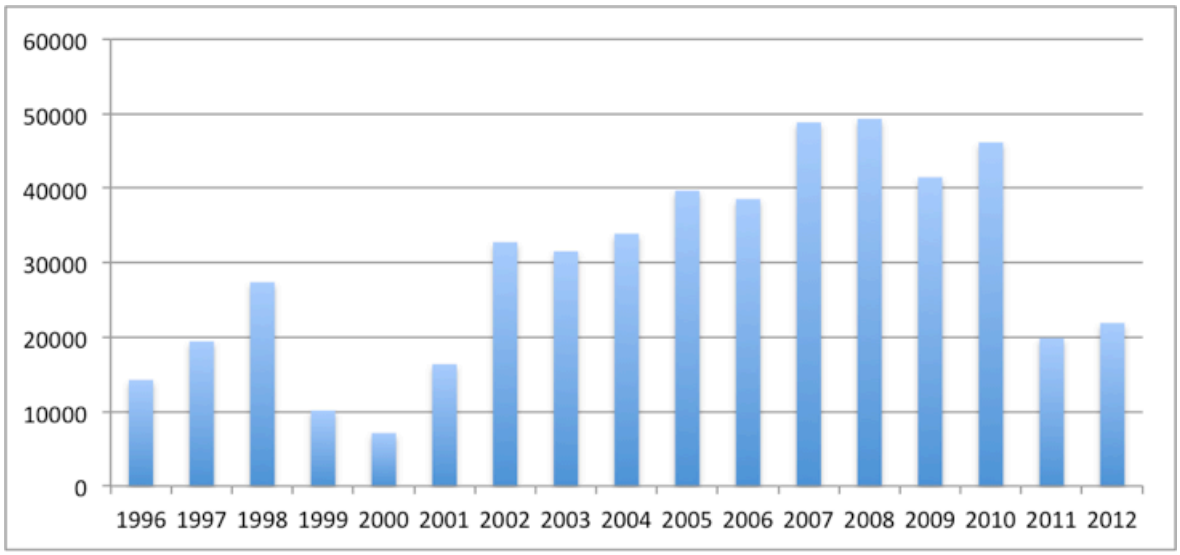


Figure 3 - Cases of dengue in Malaysia per year. Information extracted from (Ministry of health Malaysia, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012)

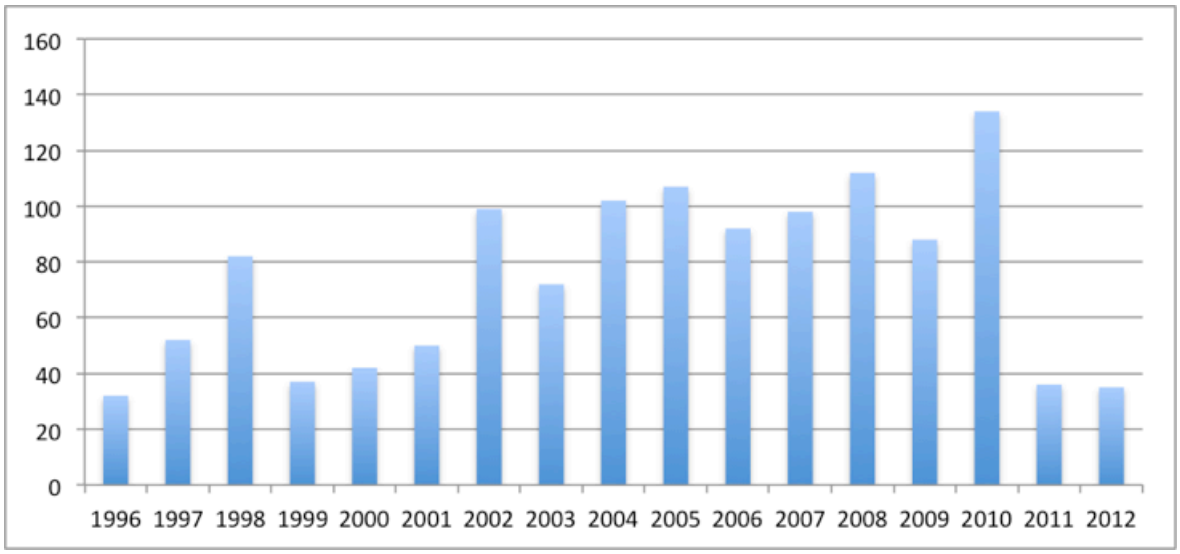


Figure 4 - Number of deaths due to dengue infection in Malaysia per year. Information extracted from (Ministry of health Malaysia, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012)

In **Figure 2** it can be seen the incidence rate of the disease in Malaysia per each 100.000 inhabitants, from 1973 until 2012. Although there has been a decrease in the amount of people

infected with the disease in 2011 and 2012, the tendency since 1973 has been an increase in the incidence rate. This tendency is confirmed in **Figure 3**, which shows the number of cases of dengue in Malaysia since 1996 up until 2012. The number of deaths in Malaysia between 1996 and 2012 is shown in **Figure 4**. A closer look to the graphs depicted in **Figure 3** and **Figure 4** suggests a correlation between the number of cases and the number of deaths.

There is no specific treatment for the disease and only recently a vaccine to DF/DHF has been approved.

During December 2015, a vaccine named Dengvaxia has been approved for use in three countries. Mexico and the Philippines approved the vaccine in early December, while in late December the drug has received the green light in Brazil, which had more than 1.4 million cases of the disease in 2015. The clinical development of the product took over 20 years of study and more than 40000 participants, including children, adolescents and adults in 15 countries. The French pharmaceutical company Sanofi developed this vaccine. Exactly when the inoculations will be deployed, and at what price, remains unclear as terms of the vaccine are being negotiated between the company and the countries (clicRBS, 2015; Maron, 2015). Other dengue vaccines are also in development (Bocchini, 2015) but none received approval (Maron, 2015).

Nonetheless, Dengvaxia is not a perfect vaccine. In clinical trials it only reduced the chances of developing the disease by 66%. Also, it is only approved for use in people 9 to 45 years old who live in dengue-endemic areas, not young children or the elderly, or tourists coming from non-endemic areas. "Patients who do chemotherapy treatment, pregnant women, patients who have, for example, AIDS, with very weakened immunity system can not make use of this vaccine", said the Emilio Ribas Institute's virologist, Jean Gorinchteyn. The vaccine seems to be least effective in children younger than nine years old, particularly among kids under six, whose immune systems are especially vulnerable and are therefore the ones who need the vaccines the most. There are also unanswered questions regarding vaccinated individuals who could potentially have more severe cases of the disease if they contract it later in life (Globo, 2015; Maron, 2015).

The approval of Dengvaxia in the three aforementioned countries is remarkable and the next few months will be very interesting and important to assess about how good the vaccine really is, and discuss about its implementation in the vaccination system of endemic countries. But there is still a

long way to go. In Brazil, it will take at least three months for the vaccine to start to be sold, and in April 2016 advisers from WHO will examine the vaccine and provide recommendations for its use (clicRBS, 2015; Globo, 2015; Maron, 2015). Therefore, it is important to devise the means to increase the effectiveness of dengue prevention, in order to avoid that it becomes pervasive (Bee, Lye and Yean, 2009; Tarmizi, et al., 2013b; Yusof and Mustaffa, 2011).

For a disease that is complex in its manifestations, management is relatively simple, inexpensive and very effective in saving lives, as long as correct and timely interventions are performed. It is critical for the public to seek treatment immediately after the appearance of the first symptoms of dengue, and if there has been a previous warning concerning an outbreak, people showing symptoms could look for proper care earlier, that way avoiding any complications that could lead to death. Furthermore, the existence of a surveillance system for dengue might suggest in advance that an outbreak was about to take place in a resource-limited and endemic location, allowing the government to take appropriate measures in due time (Long, 2014; Tarmizi, et al., 2013b; World Health Organization, 2009).

Surveillance can be defined as a reporting system that consists of collection, analysis and dissemination of information to authorities, facilitating timely decision-making and further actions. Through appropriate surveillance systems, agencies dealing with public health might be able to trim down the consequences of outbreaks by predicting potential outbreaks or detecting its occurrence as early as possible. This can lead to a reduction in the mortality, morbidity, and even in the economic effect of an outbreak (Long, 2014; Racloz, et al., 2012; Tarmizi, et al., 2013b).

It is in the context presented above that this study will be carried out. The focus will be in the development of descriptive and predictive *Data Mining* (DM) models that might allow to estimate the probability of dengue outbreaks and its final size, as well as recommending possible ways of action, by relating climate changes and, possibly, other factors[8], with the incidence rate of DF/DHF. The meteorological data comes from the *Fakulti Sains Kesihatan* (Faculty of Health Sciences) in *Universiti Kebangsaan Malaysia* (UKM), while the demographic dengue dataset comes from a study conducted by the *Unit Kawalan Vektor* (Vector Control Unit) of the *Pusat Kesihatan* (Health Centre)

[8] Although the literature suggests that climate changes may be the most important cause behind DF/DHF outbreaks, other factors might also need to be considered (Bee, Lye and Yean, 2009; Reiter, 2001).

from the Hulu Langat district, in the state of Negeri Sembilan, 1 of the 16 states and federal territories in Malaysia. The dengue-related data is private, and thus shall not be disclosed to public.

Work previously done on this subject is very recent, as before the Big Data era it was not possible to extract valuable information (and therefore knowledge) from data contained in *Databases* (DBs) due to computational limitations. But with the Big Data, the fast development, creation and maturity of technologies to store, manipulate and analyse this data in new and efficient ways allowed companies, governments and non-profit organisations to do so in a very effective way (Dean, 2014).

1.2 Motivation and objectives

Mosquitoes can be found throughout the world except in places that are permanently frozen, and constitute the source of many diseases, especially in the tropics. These mosquito-borne diseases have emerged as a major human health concern worldwide, particularly in Malaysia where dengue occurrence has increased recently. It is widely accepted that climate variables, such as high temperatures, humidity and precipitation lead to an increase in mosquito population, which in turn magnifies the incidence and geographic range of dengue. Moreover, the anticipated global warming will surely aggravate this situation. Therefore, strategies to control mosquito population need to be developed and implemented, as well as early warning strategies for dengue. The motivation for carrying out this dissertation work comes from the dimension the disease has reached, especially in Malaysia, but also from the opportunity to work with real data coming from the Malaysian government. By analysing such data, valuable knowledge can be learnt, and given that this knowledge could have enormous applications in human health, any kind of discovery would make this work a success and a step forward in making easier to predict dengue outbreaks in the future, thus saving many lives (Bee, Lye and Yean, 2009; Reiter, 2001). As has been formerly mentioned, factors other than the weather related ones might contribute to the appearance of an outbreak. Since it is commonly accepted that there is a correlation between high temperature, humidity and rainfall with the onset of dengue outbreaks, the study of these factors is the main motivation behind this work.

DM can be divided in 3 distinct classes: descriptive, predictive and prescriptive. The most common class, the descriptive, is used solely to summarise Big Data into smaller chunks of information, from which you can easily obtain valuable knowledge. Predictive analytics is one step further in data reduction. By the means of a more complex set of tools and techniques, it allows analysts to make assumptions about the future. Prescriptive analytics goes a little bit further than the 2 aforementioned DM classes, in a way that it can recommend several courses of action, as well as show the possible outcome of each of them (Bertolucci, 2014). The focus of this study will be on descriptive (like clustering and Association Rules - ARs) and predictive DM (Decision Trees – DTs – are examples of such techniques). Due to decisions that were made while writing this dissertation, clustering and DTs were substantially more used than ARs, and in the next chapters the author will go on details about these techniques. Hence it has been decided to introduce ARs in the present chapter and section.

ARs represent knowledge embedded in data sets as probabilistic hypotheses, and interesting patterns are expected to be found through the use of AR mining, and from them possible dengue preventive measures will be derived. But this process is everything but trivial. When analysing data sets having n attributes there will be $2^n - 1$ possible item sets, and $3^n - 2^{n+1} + 1$ possible ARs. Even for modest values of n the number of possible nontrivial ARs will be very large. For $n = 14$, which is the number of attributes used in the beginning of this research, there exists more than 16 thousand possible item sets and almost 5 million non-trivial ARs. A brute-force approach to the problem will compute the support and confidence for every AR possible, hence having to scan the whole item set for every AR. But as the support of each rule depends solely on the support of the corresponding item set, decomposing the AR mining into two major subtasks, beginning by seeking frequent item sets followed by the generation of strong ARs from the frequent item sets, is an initial step towards a huge boost in the performance. This step is already done by most of the AR mining algorithms nowadays (Dou, et al., 2008; Simovici; Tan, Steinbach and Kumar, 2005).

Apriori (Agrawal and Srikant, 1994) is the most famous algorithm for mining frequent item sets. It starts from candidate item sets with two items, and keeps increasing the size of the item sets. During each phase, it scans the DB for obtaining the support count of the candidate item sets, extracting all the sets satisfying the minimum support (minSup) requirement. Apriori was a pioneer algorithm, to the extent that it was the first to use a support-based pruning, which allowed it to

control the exponential growth of candidate item sets. This happens due to the downward-closure property of support (Tan, Steinbach and Kumar, 2005).

Downward-closure property of support

- If an item set I is frequent, so is every subset of I.
- If an item set I is not frequent, no superset of I is.

Despite the improved performance that can be obtained through the support-based pruning, finding frequent item sets with a prescribed support can still offer formidable computational challenges, and is a much more studied problem than generating strong ARs from the frequent item sets. In fact, after the computation of all the frequent item sets, calculating the confidence of an AR does not require additional scans of the transactional data set, which makes the second and last step in AR mining straightforward when compared with the first (Tan, Steinbach and Kumar, 2005).

Rule quality and rule quantity are, therefore, the problems to be addressed in AR mining. Singular items that could be of interest to decision makers may not be found if the minSup is set too high, and setting minSup low, however, may cause combinatorial explosion[9]. Hence it is necessary to handle this problem and the aforementioned complexity in the most appropriate way, and that is why the use of GAs will bring added value to the generation of ARs (Kumar and Iyakutti, 2011).

GAs are search heuristics that try to reproduce the process of natural selection, therefore incorporating Darwinian evolutionary theory with sexual reproduction. They were invented by John Holland in 1970, and later developed by him and his colleagues and students, becoming popular in 1975 after (Holland, 1975) was published. The main motivation for its use in the discovery of high-level prediction rules is that they perform a global search, allowing to find the maximal frequent item sets in due time. Also, by using tools implementing GAs it will be possible to perform variations and modifications in the GA parameters before each of these runs, allowing to tune and direct the results to what is more desirable (Dou, et al., 2008; Kumar and Iyakutti, 2011).

[9] The exponential growth rate of many functions representing search problems, as a result of combinatorial considerations.

The evolutionary process is an extremely simplified simulation of its biological version, starting from a randomly generated population of individuals called genome, and evolving this population in steps called generations, towards an objective function or fitness function, that indicates how good a solution is. Each individual in the population is called chromosome, and is a DNA string consisting of genes. In each generation, several chromosomes are selected for reproduction (parents), producing offsprings through random transformations over the parents, namely crossover between pairs of chromosomes and mutations of genes on the selected chromosomes. If the newly generated chromosomes are better (have a better fitness), they are then added to a new population, and the older ones are discarded. By doing this over successive generations, better solutions will thrive, while the least fit solutions will die out. This process is iterative, ending when some previously specified criteria is met (Indira and Kanmani, 2012). A comprehensive description of GAs can be found in (Goldberg, 1989).

1.3 Research methodology

Throughout the literature, the terms process model and methodology have been used to refer to the same thing, which inevitably led to some confusion. Therefore, prior to going into details concerning the employed methodology, the difference between the concepts of process model and methodology must be clarified (Marbán, Mariscal and Segovia, 2009).

A process model can be defined as the set of tasks to be executed to develop a particular element, as well as the elements that are produced in each task (outputs) and the elements that are necessary to perform a task (inputs). Methodology is the instance of a process model that lists tasks, inputs and outputs and stipulates how to do the tasks (Pressman, 2005).

This study was organised in several stages of development. In the first one, some related information about the theme approached was collected and analysed. This phase was carried out as thoroughly as possible, for the purpose of defining the objectives and specific requirements of this project in its entirety. It is necessary to develop an understanding of the application domain before applying the *Knowledge Discovery in Databases* (KDD) process model, as well as identifying its goal from the customer's point of view. The application of the KDD process, depicted in **Figure**

5, has three major tasks: preprocessing, DM process and postprocessing, which in turn can be further subdivided. The actual first step consists in cleaning and integrating the data sources. The occurrence of inconsistencies and duplicate entries is handled in this stage, and mistakes coming from manual inputs, and from other sources, must be properly addressed. In the specific case of this dissertation work, the meteorological data must be matched to the dengue-related data through the time period. Since not all data is relevant to the objectives to be achieved, the second phase consists in selecting task related data from the integrated resources, transforming it afterwards into a format that is ready to be mined. In this stage, the dataset is divided in training and testing datasets in order to derive classification rules, attributes might have to be rescaled for the clustering algorithms to work and the non-nominal variables might be converted to nominal ones, to make it easier to use the association algorithms in next phases of the process (Zhao and Bhowmick, 2003).

The modelling techniques that are more consistent with the main purpose of this dissertation are identified and evaluated, and the most appropriate parameters for these algorithms shall be searched upon through several variations and modifications in those parameters. Subsequently, DM techniques are applied to the derived rules and it is expected that meaningful information shall be learnt. These steps form the core stage of this work. Following there is a period of analysis and evaluation, both of the developed system as well as of the quality of the information derived. The results may not satisfy the requirements, or they might even contradict the application domain. This is why the process is iterative, and it is expected that each step of the KDD has to be reviewed and improved until relevant knowledge is found, or even that the requirements might have to be changed (Zhao and Bhowmick, 2003). The last moment in the development of the research work is the organisation and presentation of acquired knowledge in an intuitive way, so that the final user can easily handle it. This stage is called deployment, and although it is depicted in **Figure 5**, it is a post KDD task (Azevedo and Santos, 2008; Cios, et al., 2007; Zhao and Bhowmick, 2003).

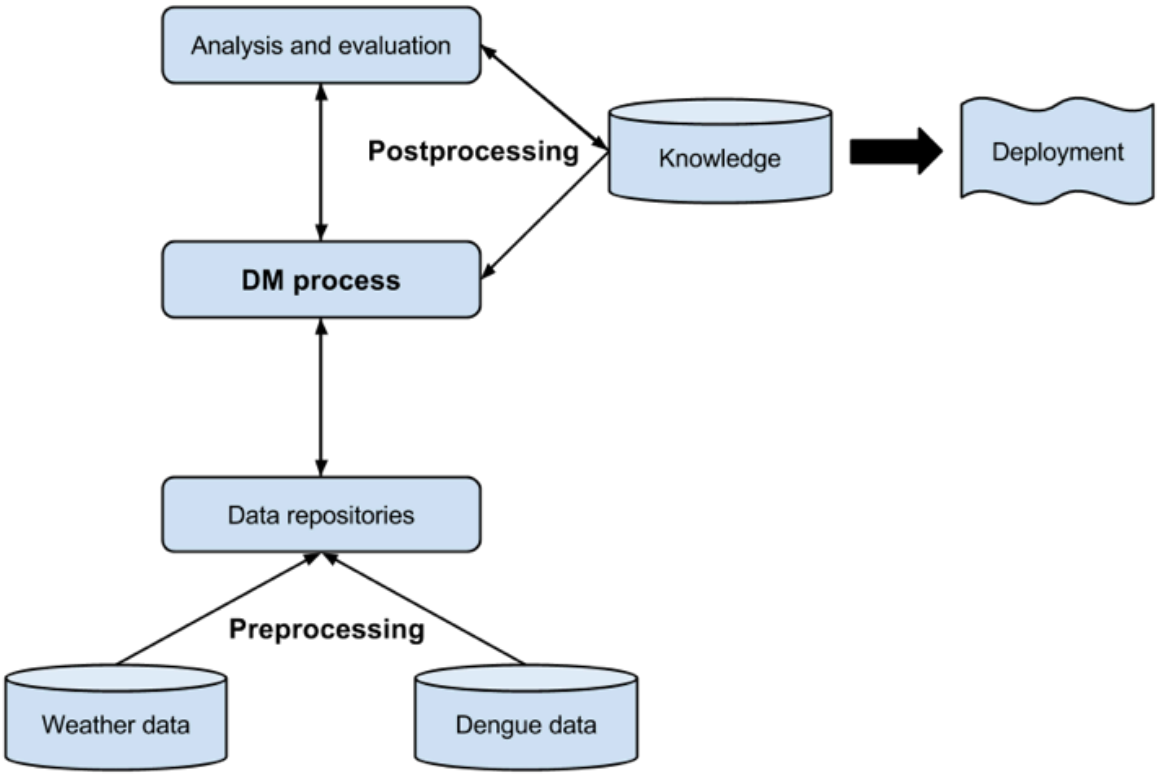


Figure 5 – The KDD process for the dissertation work

The year 2000 marked the most important milestone in the field of KDD process models: the publishing of the CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology in (Chapman, et al., 2000). CRISP-DM can be viewed as an implementation of the KDD process model, and is the most used methodology for developing KDD projects. In fact, it is nowadays a *de facto* standard, thus naturally it was the methodology chosen to carry out this dissertation work. The steps previously defined in this section, including the business understanding that precedes the KDD process model and the models' deployment that comes after the application of the KDD process model, briefly explain its implementation in the context of the present work (Azevedo and Santos, 2008; Marbán, Mariscal and Segovia, 2009).

CRISP-DM methodology consists of 6 stages (Azevedo and Santos, 2008; Chapman, 1999; Chapman, et al., 2000; Cios, et al., 2007; IBM Corporation, 2011; Rupnik and Jaklič, 2009) (**Figure 6**), namely:

1. Business understanding. The first step consists of trying to gain as much insight as possible into the project objectives and requirements from a business perspective, and outline a preliminary plan designed to accomplish those objectives.
2. Data understanding. Starting with a data collection and proceeding afterwards to a familiarisation with the collected data, this phase enables data quality problems to be identified, to discover first insights about the data or to detect interesting data subsets.
3. Data preparation. At this stage, all activities needed to construct the final dataset from the initial raw data will be covered. The designed dataset aggregates the data that will be fed into DM tool(s) in the next moment of the CRIPS-DM methodology.
4. Modelling. At this point, various modelling techniques are selected, created and then applied, and their parameters are calibrated to optimal values. This step finishes with the assessment of models, but since some DM methods may require a specific format for input data, stepping back to data preparation phase is often necessary.
5. Evaluation. The models that have been built having high quality from a data analysis perspective are now evaluated from a business objective perspective, and the steps executed during the model construction are also reviewed. The purpose is to determine whether any important business objective has not been sufficiently scrutinised. A decision about the use of the DM results should be reached at the end of this stage.
6. Deployment. Creating the model might not be the end of the project. Usually, it ends only after organising and presenting the knowledge gained in a way that the customer could use it within decision-making.

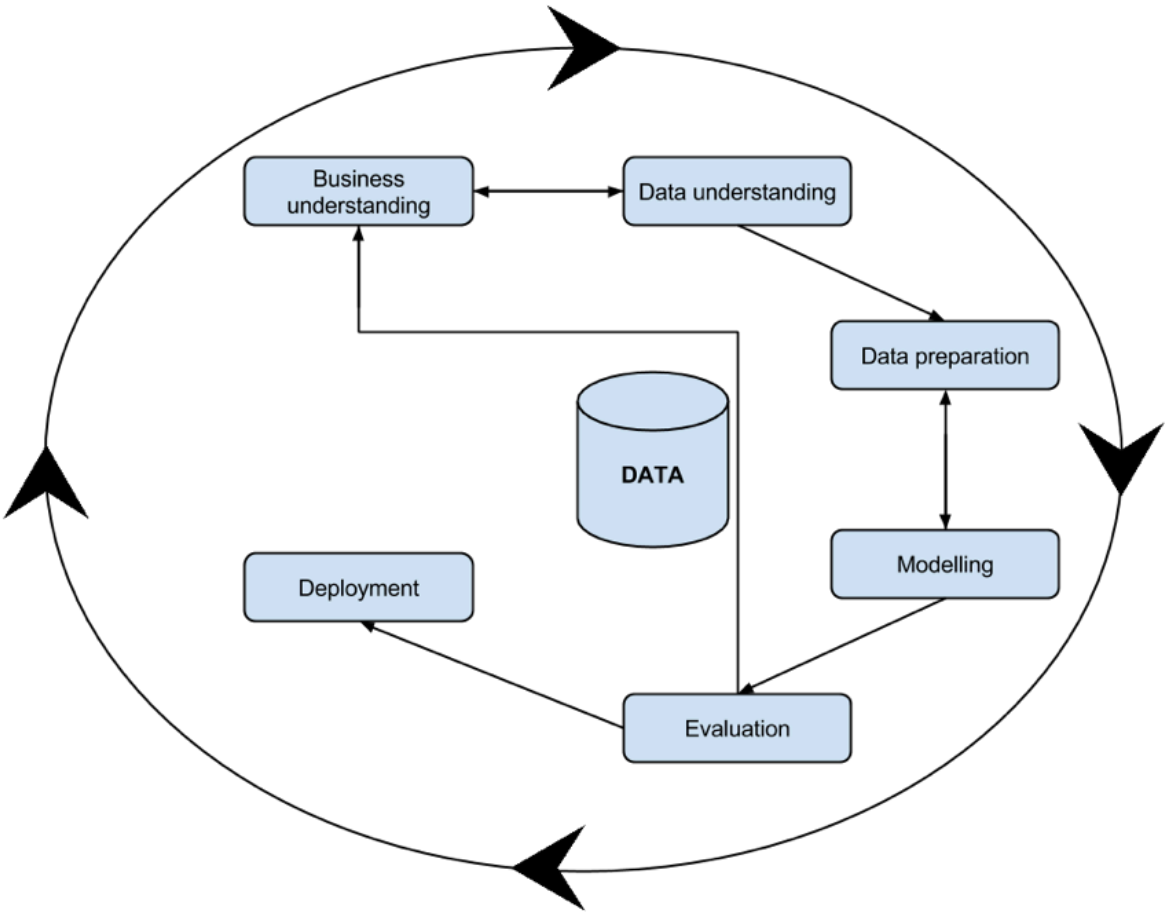


Figure 6 – CRISP-DM methodology - based on (Cios, et al., 2007)

Although the CRISP-DM methodology has only been published in 2000, it was first conceived in 1996, and further developed by a consortium of companies, namely SPSS (Statistical Package for the Social Sciences) Inc., Teradata Corporation, Daimler AG, National Cash Register (NCR) Corporation and OHRA. The first version would be first presented in 1999 (Chapman, 1999). After the publication of the methodology in (Chapman, et al., 2000), there have been discussions throughout the years about updating it, but it was only after the SPSS Inc. acquisition by International Business Machines (IBM) Corporation in 2009 that a new version (IBM Corporation, 2011) would be published.

1.4 Document structure

After this chapter it may be found an overview of the work related with the studied topic in chapter 2 - Literature review. In this chapter, an analysis to the work done in the field of creating descriptive and predictive models for dengue outbreaks is shown. The approaches taken by other authors while devising these models are also presented, and an attempt is made to obtain an understanding on the most relevant models that have been developed.

The following chapter - Understanding and preparing the data - goes through the first three steps of the application of the CRISP-DM methodology to the topic of this research, namely Business understanding, Data understanding, and Data preparation, and the following chapter, Modelling the data presents a detailed description of the last three steps undertaken in this work: modelling, evaluation, and deployment. The last chapter of this dissertation - Conclusions and future work - exhibits the most important remarks captured during this research work, along with some notes concerning functionalities and details that can be improved in a near future. In this last chapter the author also talks about the biggest difficulties he had while implementing the CRIPS-DM methodology, and decisions that had to be taken throughout the developing process.

Chapter 2

2 Literature review

The literature review is very important in the scientific work context. It represents the development or knowledge level of some technique, methodology, science, etc., in a specific moment. It must contain the essence of the work that has been done in the approached domain, by referencing tools, technologies, works, etc., that are somehow related to that domain. This part of the work must be critique and weighted: it is not supposed to just list all the works related to the approached area. In fact, at this stage it is expected that the author reflects on the data collected, and relates that data with the subject developed in his work, referencing other works properly, whenever these are mentioned. By doing so, it is possible to achieve a better management of the research, as well as expedite the process of construction and innovation of it.

This chapter is divided in two sections. In the first, DM models that aim to describe and predict DF outbreaks are presented and the author talks briefly about them, and the second will dissert about what has been done in the field of GAs towards deriving (good) ARs. The former will thus address research works that attempt the ultimate goal of this dissertation, which is to devise a mathematical model (or several) able to predict future dengue outbreaks. The second chapter has been included in the literature review because although the focus of the dissertation has shifted from that in an early stage, a lot of effort has been put on it.

2.1 Descriptive and predictive Data Mining models

Several authors have already studied the literature concerning predicting DF outbreaks through DM models extensively (Focks, et al., 1995; Focks, et al., 1993; Racloz, et al., 2012), thus aiming to analyse different modelling methods and their outputs. Therefore, it can be concluded that there is quite an extensive literature regarding the creation of descriptive and predictive dengue models, as will be shown next. However, the work done on this field focus mostly on meteorological drivers, rather than demographic factors. There is also quite an extensive literature regarding vector dynamics.

Newton and Reiter developed the first dengue model (Newton and Reiter, 1992), a deterministic representation in which populations of vectors and hosts were divided into subpopulations representing disease status and the flow between subpopulations was described by differential equations. They have also shown that the use of Ultra Low Volume (ULV) insecticide sprays to control epidemics is not effective in controlling the *aedes aegypti*, and therefore controlling the epidemic. Several other deterministic models have been devised after that, considering different aspects of the malady. In (Esteva and Vargas, 1998), the authors created a model in which dengue is transmitted in constant human population and variable vector population, and in (Esteva and Vargas, 1999) the same authors modelled the transmission of dengue in a variable human population. They presented a third article (Esteva and Vargas, 2000), in which vertical[10] and mechanical[11] transmission of dengue in the vector population were taken into account, and created a non-linear system of differential equations that modelled the dynamics of transmission of dengue. Two years later, seasonally varying parameters and the presence of two dengue serotypes simultaneously, would be addressed in (Hartley, Donnelly and Garnett, 2002), and in the year after, age structure in the human population (Pongsumpun and Tang, 2003) and presence of two serotypes of dengue at separated intervals of time (Derouich, Boutayeb and Twizell, 2003) were covered in two separate works. In addition, Tran and Raffy demonstrated that remote-sensing data

[10] Vertical transmission occurs when the disease is transmitted from one generation to another, in contrast to horizontal transmission, when the spread of the infectious agent happens between members of the same species having no parent-child relationship.

[11] A transmission is mechanical when the disease agent does not replicate or develop in the vector, being simply transported from one animal to another (flies). Contrary to mechanical transmission, biological transmission takes place when the vector gets infected, usually through blood from an infected animal, and the agent replicates and/or develops inside the vector, and then regurgitates the pathogen onto or injects it into a susceptible animal. Fleas, ticks, and mosquitoes are common biological vectors of disease.

could be used for building a model of the spatial and temporal dynamics of dengue (Tran and Raffy, 2006).

Several other authors have chosen to take a different approach. In 1995 a pair of stochastic models describing the daily dynamics transmission of DF/DHF in the urban environment were published in (Focks, et al., 1995). This work was based on the simulation of a human population growing in response to country- and age-specific birth and death rates. Many years later, an inter-host three level cellular automata model was presented in (Santos, et al., 2009), describing the pertinent population groups in an urban environment: human, adult vector mosquito, and immature vector in the aquatic phase. Between 2006 and 2009 a series of minimalist stochastic models were developed in Buenos Aires. In (Otero, Solari and Schweigmann, 2006) the seasonal dynamics of *aedes aegypti* populations in a homogeneous environment are modelled, based on the life cycle of the mosquito. The second model describes the *aedes aegypti* dispersal driven by the availability of oviposition[12] sites in an urban environment (Otero, Schweigmann and Solari, 2008). The last model, presented in (Otero and Solari, 2010), considered the seasonal and spatial dynamics of the vectors and characterises the disease dynamics triggered by the arrival of viremic people in a city. These three works put a lot of emphasis on the stages of *aedes aegypti* mosquito life cycle, previously depicted in **Figure 1**, and shown again in **Figure 7**: eggs, larvae, pupae, and adults.

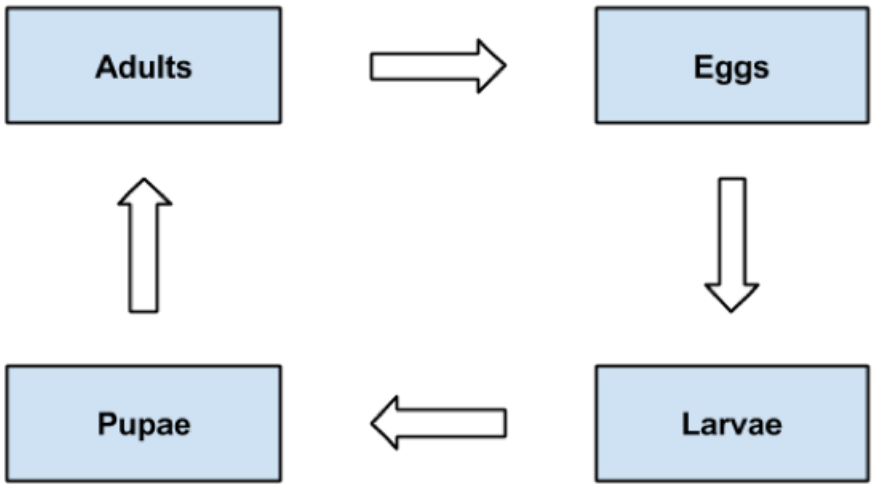


Figure 7 – Life cycle of *aedes aegypti*

[12] The act of depositing or laying eggs by a female oviparous female.

In (Yusof and Mustaffa, 2011) a prediction model incorporating least squares support vector machines in predicting future dengue outbreaks was proposed. The data used in the study was collected in 5 districts in the state of Selangor, and comprises of data on dengue cases and rainfall.

Several other DM techniques were reviewed in (Long, 2014), in which the focus was on frequent pattern mining and outlier mining to be applied on a generic outbreak detection model.

An outbreak detection model for dengue was presented in (Tarmizi, et al., 2013a; Tarmizi, et al., 2013b). In these articles, which are very similar, three classification methods were used: Artificial Neural Network (ANN), DT - more precisely J48 which is an implementation of the C4.5 algorithm in WEKA (Waikato Environment for Knowledge Analysis), and Rough Set (RS) theory. The evaluation process was carried out through the comparison of the models determined by using the 2 test options available in WEKA: 10-fold cross validation and percentage split. Each model created has been evaluated in terms of several criteria, namely the:

- Correctly Classified (CC), also known as the accuracy of the tested data sample. It indicates the percentage of correct tests of the entire data sample.
- Receiver Operating Characteristics (ROC) value, which has a value between 0 and 1 and it is used to determine whether a model is good to use for prediction. A predictive model is considered weak if the ROC is close to the value of 0.5, while a good predictive model will produce the ROC value close to the value of 1.0.
- The Root Mean Square Error (RMSE) measures the differences between values predicted by a model or an estimator and the values actually observed. Its value is also between 0 and 1 and the best RMSE value is the lowest value.
- F-measure. This is another measure of a test's accuracy. The traditional F-measure or balanced F-score is the harmonic mean of precision and recall. The recall is defined as the fraction of relevant instances that are retrieved, and precision is defined as the fraction of retrieved instances that are relevant. A high value of F-measure indicates that the classifiers have reasonably high precision and recall values.

All models devised in the study conducted in (Tarmizi, et al., 2013a; Tarmizi, et al., 2013b) performed well, especially when compared with the results obtained by previously developed

models (Bakar, et al., 2011; Long, et al., 2010; Mousavi, et al., 2013) that used similar datasets, with different number of attributes and representation schemes. Such comparison was also done in (Tarmizi, et al., 2013a; Tarmizi, et al., 2013b) and it is here shown in **Figure 8**, where the models were evaluated (only) in terms of accuracy (CC). The number in brackets refers to the number of parameters used in each approach. It has been concluded that the significant selection of attributes in the dataset contributed to the good results.

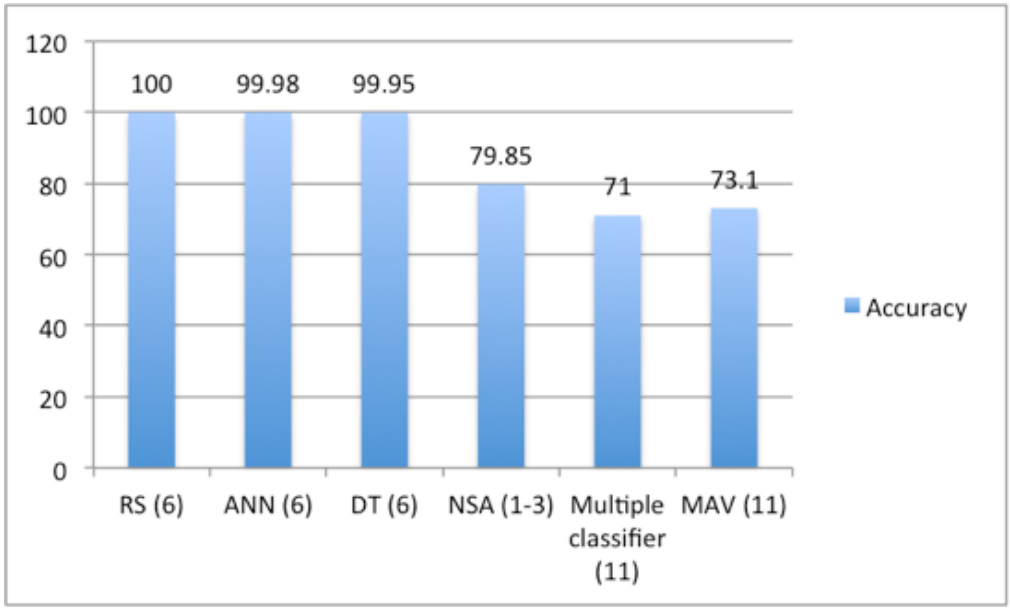


Figure 8 - Comparison of models using similar datasets – based on (Tarmizi, et al., 2013a; Tarmizi, et al., 2013b)

A Negative Selection Algorithm (NSA) is presented in (Mousavi, et al., 2013). In this article, the authors claim that over the past few years, a lot of effort has been placed on Artificial Immune Systems (AIS), one of the Artificial Intelligence (AI) approaches, in an attempt to successfully predict future dengue outbreaks. "AIS are adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving" (Castro and Timmis, 2002). The aforementioned paper (Mousavi, et al., 2013) focus on creating a new model for dengue outbreak detection, based in an algorithm from the field of AIS, namely the NSA. As shown in the **Figure 8**, the accuracy of the results was close to 80%.

NSA outperforms Multiple Attribute Value (MAV) method for dengue outbreak detection, presented in (Long, et al., 2010), in terms of successful detection rate and false alarm rate, which obtains a 73.1% value (at most) in terms of accuracy, as depicted in **Figure 8**. MAV focuses on the use of AR mining to identify a potential attribute or a combination of attributes, within the data, to be used as indicator of dengue outbreaks. The experiment is conducted using Apriori concept that finds a frequent item based on the MAV of a real data repository and generates ARs. This research started by identifying the possible number of items to be considered in determining outbreaks, based on detection rate, false positive rate and overall performance, and it has been concluded that using maximum item length a better performance is achieved in detecting outbreak and that the use of high volumes of records is not critical in determining the existence of a dengue outbreak.

The model depicted in **Figure 8** as multiple classifier was published in (Bakar, et al., 2011). This work presents a predictive model for epidemic detection using multiple rule-based classifiers. The classifiers used are the DT, rough set classifier, naive bayes, and associative classifier. Several models have been developed to study the performance of various rule-based classifiers individually, as well as the combination of the classifiers. The combination of classifiers obtained a better accuracy in comparison with the single classifiers, wherein the best-recorded accuracy was 71.25%.

The quantitative statistical relationship between dengue incidences and rainfall in the Western province of Sri Lanka between 2000 and 2004 has been studied in (Pathirana, Kawabata and Goonatilake, 2009), and a temporal and spatial correlation between post rainfall seasons and dengue disease was discovered. A regression model was obtained using statistical analyses, and it has accurately predicted disease outbreaks using the data. The correlation coefficient for this model has shown that there is a 3 to 4 week lag time between the rainfall and outbreaks. This gives a good hint on where to look for correlations between meteorological factors and DF outbreaks.

In (Flamand, et al., 2014) the objective was to identify local meteorological drivers of DF/DHF in French Guiana, and an original DM method was applied to the available epidemiological and climatic data. The contribution of the DM method to the understanding of factors associated with the propagation of infectious diseases and their spatiotemporal spread was assessed throughout

the research. Contextual sequential pattern extraction techniques were applied to epidemiological and meteorological data to identify the most significant climatic factors for dengue, and the relevance of the extracted patterns was investigated towards an early warning of dengue outbreaks in French Guiana. The maximum temperature, cumulative rainfall, minimum relative humidity, and global brilliance were identified as determinants of dengue outbreaks, and the precise intervals of their values and variations were recorded. In (Flamand, et al., 2014), there is a significant lag between the occurrence of an outbreak and factors related with meteorological conditions, which is even bigger in this work. In fact, the strongest significant correlations were observed between dengue incidence and meteorological drivers after a 4–6-week lag.

2.2 Mining Association Rules through Genetic Algorithms

Several research works that address the use of GAs to mine ARs have been published, especially over the last few years. Nonetheless, the majority of researchers working on this field are focused on DBs with discrete attributes, although most real-world DBs embody essentially continuous attributes. Furthermore, the majority of the tools that work in the continuous domain merely discretise the attributes and treat them as if they were discrete (Vannucci and Colla, 2004). In (Martínez-Ballesteros, et al., 2009) the authors developed a GA able to find ARs over DBs with continuous attributes, avoiding the discretisation as an early step of the process, which would allow the overlapping of the regions covered by the rules. But even before this work, a GA able to mine quantitative and categorical (qualitative) ARs without a prior discretisation of the data was presented in 2007 on (Salleb-Aouissi, Vrain and Nortet, 2007). Its analysis has been fundamental for this research in its early stages, although the tool implemented by Salleb-Aouissi, Vrain and Nortet is much more complex than what is needed to fulfil the requirements of this research. The developed tool's (QuantMiner) algorithm starts with a set of rule templates and then looks dynamically for the "best" intervals for the numeric attributes present in these templates. Each rule template consists of the left-hand side attributes (conditions), and the right-hand side attributes (conclusions), as well as the respective range of values of each quantitative attribute or the category of each qualitative attribute.

QuantMiner allows the user to filter attributes as well as to choose which attributes form the antecedent and consequent of the rules, through the use of a *User Interface* (UI). It also lets the user choose several parameters, namely the population size, number of generations and percentage of cross-over and mutations, which makes it a very powerful tool. Its interactivity allows avoiding the discovery of hundreds of rules, through establishing minimal values for the support and confidence.

In case studies with frequent item sets containing many items, the candidate item sets will be huge and the algorithm will have to scan the DB millions of times. In (Dou, et al., 2008), the authors call this the low efficient problem of Apriori, and propose an efficient DM technique that allows to overcome it. The tool described in this research gives a quick response to users and provides a friendly UI, said to grant users with real demanded rules. The proposed system consists of two main stages. It starts with the GA, which is used to mine maximal frequent item sets and show them to users. In the second part, ARs are deduced in terms of the maximal frequent item sets and then the DB is scanned for obtaining real support and confidence of those rules. The main characteristic of the implemented GA is that, when judging if an item set is frequent or not, it does not need to scan the DB; it just mines maximal frequent item sets and only scans the sets in which users are interested in. This feature, when properly used, can drastically reduce the mining time, compared with traditional algorithms like Apriori.

The problem addressed in this work will only require mining of single level ARs, but in the Big Data analysis context, strong ARs tend to be in multilevel forms, therefore requiring more complex and efficient methods to be mined efficiently. In (Xu, et al., 2014) the authors presented an innovative algorithm that allows the users to mine good multilevel ARs, and prove its efficiency by testing its performance in several DBs, as well as the classic Apriori algorithm. The results obtained with the newly implemented algorithm were more accurate, and Apriori was even slower than the presented algorithm.

Alatasetal (Alataş and Akin, 2006) is a GA that simultaneously search for intervals of quantitative attributes as well as positive and negative quantitative ARs in a single run of the algorithm. The chromosomes represent rules, with each gene consisting of four parts. The first part serves as the antecedent or consequent of the rule while the second represents the positive or negative ARs and the third and fourth represent, respectively, the lower and upper bound of the item interval. The

proposed GA performs a dataset-independent approach that does not depend on the minSup and minimum confidence (minConf) thresholds. The authors of (Alataş, Akin and Karci, 2008) used Alatasetal to devise a multi-objective differential evolutionary algorithm (MODENAR) capable of mining accurate and comprehensible quantitative ARs without being necessary to specify the minSup and the minConf. This algorithm uses the same coding scheme for the chromosomes as Alatasetal but without the second part. MODENAR considers four objectives to improve the quality of the rules, namely support, confidence, comprehensibility and amplitude of the domain of the attributes. Thus, these objectives form the fitness function.

A GA meant to obtain numeric ARs was introduced in (Yan, Zhang and Zhang, 2009). This algorithm used confidence as a single objective to be optimised in the fitness function. To fulfil this goal, the authors avoided the specification of the actual minimum support, which is the main contribution of this work. Each chromosome of EARMGA, the presented algorithm, encodes a generalised k-rule, where k indicates the desired length. The most interesting rules are returned according to the interestingness measure defined by the fitness function, which relies on the support of the rule and its antecedent and consequent support. The authors demonstrated that the devised algorithm significantly reduces the computation costs and generate interesting ARs only.

Both MODENAR (Alataş, Akin and Karci, 2008) and Alatasetal (Alataş and Akin, 2006), as well as EARMGA (Yan, Zhang and Zhang, 2009) were tested in (Martín, et al., 2014), and its results were compared with the GA presented in the same study: QAR-CIP-NSGA-II. The rules that can be obtained with this algorithm are very strong, and it provides a good trade-off between interpretability and accuracy, maximising three objectives: interestingness, comprehensibility and performance. In fact, the authors show that with QAR-CIP-NSGA-II the obtained rules have better values for interesting measures and a higher coverage of the dataset.

Other GAs aiming to prioritise ARs can be found in (Kumar and Iyakutti, 2011; Papè, et al., 2009).

Chapter 3

3 Understanding and preparing the data

3.1 Business understanding

The first step of the CRISP-DM methodology is to assess what useful knowledge can be obtained through the DM. At this stage, a thorough analysis might help to prevent the loss of important resources.

3.1.1 Background information

Healthcare is the problem area for this case study, which is becoming an increasingly popular, or better saying, increasingly essential area for DM applications. Due to the complexity of healthcare and a slower rate of technology adoption (and even evolution), the healthcare industry is not taking full advantage of DM solutions, at least when compared with industries like the automotive industry, higher education, life sciences, telecommunications and manufacturing. In fact, DM in healthcare today remains, for the most part, an academic exercise with only a few pragmatic success stories. Although many researchers are using DM extensively in the field of healthcare, the industry has always been slow to incorporate the latest research into everyday practice (Crockett, Johnson and Eliason, 2014). In the specific case of this dissertation, the problem is already

presented extensively in the first chapter, as well as the motivation behind it, thus it shall not be debated any longer. CRISP-DM methodology also requires the presentation of advantages and disadvantages of current solutions. This task has already been done in the chapter dedicated to the Literature review, but few works can be selected as being more relevant, and therefore further analysed in this step.

3.1.2 Business objectives

The main goal to be achieved with this project is to discover hidden patterns or relationships on the data towards effective prediction of a dengue outbreak. For example, to better understand the relationship between the environment and transmission dynamics of the 4 dengue virus serotypes, can potentially guide estimates of the timing, location, and magnitude of risk and thus allow effective use of resources (Campbell, et al., 2013). There are, however, other minor objectives that have to be achieved, namely to:

1. Develop one or several mathematical models for the generation of ARs and choose the most relevant ones for the prediction of DF/DHF outbreaks.
2. Devise and/or reuse previously implemented GAs to ease the generation of ARs, and assess what benefits can be obtained through the use of those GAs in comparison to other AR mining strategies.
3. Develop other DM models, towards the success in fulfilling the leading objective.

The first two objectives were requirements provided by the research group responsible for devising the idea of this dissertation. But due to the fact that these objectives are quite restrictive, it is foreseen that slight changes might occur throughout the application of this methodology.

CRISP-DM methodology states that for every objective a success criteria shall be defined. For the main goal of this dissertation to be a success, valuable knowledge about dengue outbreaks must be learned. This know-how might help to estimate the probability of future dengue outbreaks and its final size, as well as recommending possible ways of action. Nevertheless, every small step towards this goal would already make this work a success, and a basis for future research works. The success criteria for all the other objectives can be found below:

1. An AR model shall be devised.
2. Obtain an insight about using GAs to generate ARs, in comparison with using traditional methods to generate those rules.
3. Develop one or more DM models, besides the AR models.

3.1.3 Assessing the situation

A thorough cleaning was done on the dataset beforehand, and it contains not only demographic data related with patients infected with DF/DHF, but also weather data. With 6081 records, it is not a small dataset at all, but the prior cleaning makes this project a feasible task to be done by a small working group.

There are neither security nor legal restrictions on the disclosure of project results, but the original dataset must not be disclosed to public. Nevertheless, statistics about the dataset can be published. Similarly, requirements on results deployment and financial constraints do not exist on this dissertation work, and scheduling is flexible. Notwithstanding, the models that will be devised can, in the future, be deployed as part of a surveillance system. These systems, whose definition has already been discussed in the Contextualisation section, enable public health officials to act more quickly in the event of a potential pandemic, by finding patterns that resemble to previous dengue outbreaks (Tarmizi, et al., 2013b). All the aforementioned factors are motifs for confidence in achieving the proposed objectives, especially when knowing that no extra personnel will be needed and that full access has been granted on the dataset, which comes in an Attribute-Relation File Format (ARFF). However, the academic context of this project makes it necessary that the developed models are fully explained, as well as its results.

To carry out this project, not so many risks are being taken. The data can be of poor quality or coverage, or the technology might not allow good results to be achieved with the application of DM algorithms. Besides that, the risk that suitable models will not be obtained, or that it will not be possible to implement its results, is always present. Nevertheless, since this is an academic assignment, none has monetary implications, and because the available dataset is the only data that can be accessed for this dissertation, no contingency measures for the identified risks need to be planned beforehand. Furthermore, given that the project has no costs at all, any kind of discovery like formulating a very informative DT about DF or advancing in the research about the

generation of ARs from dengue related data, with or without the use of GAs, would already be a good contribution. However, for this work to be a complete success, hidden patterns that lead to a better comprehension about dengue outbreaks have to be found on the dataset, making this dissertation work a step forward in predicting DF/DHF outbreaks in the future, therefore saving many lives.

3.1.4 DM goals

The most important business objective previously identified, namely the discovery of hidden patterns or relationships on the data towards effective prediction of a dengue outbreak, can be translated into the following DM goals:

- Find the attributes, or groups of attributes, that influence the onset of an outbreak, by deriving several DTs (or classification trees) that can predict the existence of an outbreak based on data elements. In other words, classification trees shall be built using classification algorithms, aiming to uncover relationships between the attributes in order to predict the outcome (outbreak or no outbreak).
- Discover similarities in data, that allowing for finding sets of objects (clusters) such that the inter-cluster similarity is low and the intra-cluster similarity is high. Clustering is an unsupervised DM technique, in which it is assumed that nothing is known about the initial data. Using such a technique is an attempt to validate the results obtained in a) and also to analyse the data in a different perspective.
- Obtain ARs from the given data that might allow discovering relationships between seemingly unrelated data.

An example of a breakthrough discovery, that could have been possible through the use of DM techniques, is that the presence of pre-existing antibodies is a major risk factor for DHF. The WHO has warned about the danger of the appearance of DHF epidemics in regions where a particular serotype of dengue has had (or still has) a big impact. This situation has been observed in some countries of Latin America (Esteva and Vargas, 1998; Gubler, 1998). If a similar finding would be possible thanks to the work done on this research, then this dissertation would have been a success.

Since the other objectives are technical requirements, corresponding to gaps in the research carried out so far in Malaysia, the remaining objectives are already DM goals and, naturally, some of them already reflect the technical requirements.

3.1.5 Project plan

The project plan is a summary of what has been discussed so far in terms of resources and risks, and includes an estimate on how much time will be spent during each phase of the CRISP-DM methodology. This plan can be consulted in **Table 1**.

Phase	Time	Risks
Business Understanding	1 month	-
Data Understanding	2 months	Data problems
Data preparation	1 month and 2 weeks	Data problems
Modelling	2 months	Technology problems Inability to devise good models
Evaluation	2 weeks	Changes in some features Inability to implement results
Deployment	2 weeks	Changes in some features Inability to implement results

Table 1 – The project plan

As previously discussed, the risks associated with this project are minimal. The data can place several constraints on this work, and that is a risk to deal with in the data understanding and data preparation stages. Technology can also be limiting, in a way that it might not be possible to achieve good results by applying DM algorithms on the data. Likewise, there is always a risk of not obtaining suitable models, or even obtaining such models but not be able to implement its results. Since this dissertation is an individual academic assignment, there are no resources available.

The time estimates shown in **Table 1** are only that, estimates. Moreover, it is always difficult to make good estimates on the amount of time required for the different phases in projects like this, especially when the chosen methodology is iterative. As a matter of fact, it is expected that several iterations of the first four stages of the project will have to be done due to adjustments made in the models implemented during the modelling phase. But one should not neglect that accurate time estimation is a crucial skill in project management: it allows to know how long projects will take, and to get commitment from the people involved.

The reader might have noticed that not so much time has been allocated to data preparation. In fact, for most DM projects, an experienced data engineer would spend the biggest share of his time in data preparation. Nonetheless, as discussed before in this document, most of that had already been done, and the data requires not so much preparation.

The tool that will be used for model building is R. Its open source nature was an important decision factor, but the main reason for choosing R instead of other tools, like WEKA or Rapid Miner, was its ease of use combined with implementations of all the DM tasks required for this dissertation, namely classification, clustering and ARs mining. Although the R documentation might not be very detailed, there is a vast literature dedicated to R, and that was also an important factor in its choice. Throughout the document there are some excerpts of R code, but the whole code will not be shown here. Notwithstanding, the complete code can be found in a github™ repository[13].

[13] https://github.com/binte/A-data-mining-approach-towards-effective-dengue-outbreak-prediction

3.2 Data understanding

During the second stage of the CRISP-DM methodology, a comprehensive analysis of the data available for mining is performed. This task is critical to avert problems during the data preparation phase.

3.2.1 Collecting Initial Data

Collecting the data has already been done before the beginning of this project, and since there is no need to purchase any data, and the existing data is all that is expected to be studied throughout this project, this step should be very straightforward. Nevertheless, at this point one should also assess which attributes are relevant for the project, and take some considerations about the data.

The most promising attributes in the dataset are, as expected, 'Rainfall', 'Humidity' and 'Average temperature'. Nonetheless, as it is predictable that these attributes will be the ones who will be more revealing, the purpose of this dissertation is to study some socio-demographic factors, that can also be used to determine the people which are more likely to get infected by a variant of dengue, namely age group, gender, race and job category. Some of the attributes, like 'Year' and 'Week', will bring no added value for this study and can therefore be excluded.

Although the dataset is small, both in number of attributes and number of records, it should be enough to draw generalizable conclusions and/or to make accurate predictions. The data has already been cleaned and merged in advance, hence no missing values are expected to be found in the dataset.

3.2.2 Describing Data

The dataset used in this dissertation is the same or similar to ones that have been already used in several other works (Bakar, et al., 2011; Long, et al., 2010; Mousavi, et al., 2013; Tarmizi, et al., 2013a; Tarmizi, et al., 2013b). In each of these articles, a subset of the set of attributes from the original dataset is used, but the attributes used in each work are not the same. In fact, in (Bakar, et al., 2011; Long, et al., 2010; Mousavi, et al., 2013) only socio-demographic data is used, and

each of these works uses a small subset of the original dataset coming from the *Unit Kawalan Vektor* (Vector Control Unit) of the *Pusat Kesihatan* (Health Centre) from the Hulu Langat district, with 134 attributes. An image (**Figure 24**) of this dataset can be found in appendix a, which still contains some attributes in Malay. Since the choice of attributes to be used in this research was done prior to this work, it has been decided not to translate the attributes that will not be used in this dissertation, as well as not to explain the meaning of each of them, because that work goes beyond the scope of this research work.

The dataset with which one will work is an ARFF file, comprised of 6081 records and 14 attributes. These data were coming from two different sources, but its cleaning and integration had already been done beforehand in UKM, more precisely in the centre for AI. It was expected that the quality of the mining model would be improved by combining the two datasets. Carrying out this process ultimately resulted in a dataset containing 6081 records and 14 attributes, namely 'Year', 'Week' (of the year), number of 'DF' and 'DHF' occurrences, 'Average temperature' value, 'Humidity' rate, 'Rainfall' rate, 'Age group', 'Gender', 'Race', 'Job category', 'Town', 'District', 'Epidemic' category. Some of these attributes are discrete, and some are continuous. A more detailed analysis on the dataset will be done in the following sections, but a thorough analysis of each of the attributes contained in the dataset can already be consulted in **Table 2**.

Attribute	***Type***	***Description***	***Range of values***
Year	Numeric	Year in which the case was registered	2003 – 2009
Week	Numeric	Week within that year	1 – 53
DF	Numeric	Sum of DF occurrences in that week in all considered hospitals	1 – 137
DHF	Numeric	Sum of DHF occurrences in that week in all considered hospitals	0 – 9

Average temperature	Numeric	Average temperature in the district of Seremban for that week	24.36 – 29.06
Humidity	Numeric	Average value of humidity in Seremban during that week	60.91 – 84.92
Rainfall	Numeric	Average value of rainfall in Seremban during that week	0 – 173.74
Age group	Nominal	Age group of the infected person	'Senior Citizen', Adult, Children, Youth
Gender	Nominal	Infected citizen's gender	F, M
Race	Nominal	Race of the infected citizen	MALAY, CHINESE, INDIAN, 'OTHER RACES'
Job	Nominal	Job category of the infected person	Non-executive, 'General Worker', Student, Children, Housewife, Independent, Executive, 'Senior Citizen', Unemployed
District	Nominal	District within Negeri Sembilan	SN (Negeri Sembilan), PD (Port Dickson)
Town	Nominal	Place (within the district of Seremban) where the case occurred	AMPANGAN, RANTAU, RASAH, SETUL, SEREMBAN, LABU, PANTAI, 'BDR. SEREMBAN', LENGGENG, NILAI, MANTIN
Epidemic	Nominal	Epidemic Category	DKW, MWB, TKW

Table 2 – The description of the dataset's attributes

3.2.3 Exploring Data

For the task of exploring the data, basic statistics shall be computed for the key attributes, and then hypotheses shall be formulated. For the example mentioned before, in which epidemiological evidence suggested that the presence of pre-existing antibodies was a major risk factor for DHF, the secondary infection or immune enhancement hypothesis was formulated. This hypothesis implies that DHF appears in persons who have had a previous infection with a heterologous dengue virus serotype (Esteva and Vargas, 1998; Gubler, 1998). Albeit this hypothesis cannot be formulated for this study because the dataset does not contain information about the dengue serotype, similar hypotheses shall be formulated after having computed statistics about the attributes.

Through the use of R, more precisely by using its libraries "Hmisc", "r2lh", a lot of information has been obtained about its attributes. As expected, there are no missing values in the dataset. Before going into details, it is important to take in consideration that all the data was obtained in the district of Seremban, from the state of Negeri Sembilan, because this information is crucial in order to understand the statistics that will be presented further on.

For the reader to have a better view about where the data used in this study was collected, the following figures show that in detail, starting with the location of Malaysia (**Figure 9**). Malaysia is a northern hemisphere country, located only a few degrees north of the Equator. With a total landmass of 329.847 square kilometres, this nation is separated by the South China Sea into two regions with similar size, Peninsular Malaysia and East Malaysia (Malaysian Borneo). Peninsular Malaysia is bordered in the North by Thailand and East Malaysia has borders with Brunei and Indonesia. **Figure 10** shows the location of Negeri Sembilan state, 1 of the 13 states of Malaysia, which has 3 more federal territories. This state is comprised of 7 districts, namely Seremban, Jempol, Port Dickson, Tampin, Kuala Pilah, Rembau and Jelebu. Finally, the district of Seremban is also shown **Figure 11**, within a map of Negeri Sembilan (Ooi, 2010).

Figure 9 - Malaysia's location within a world map

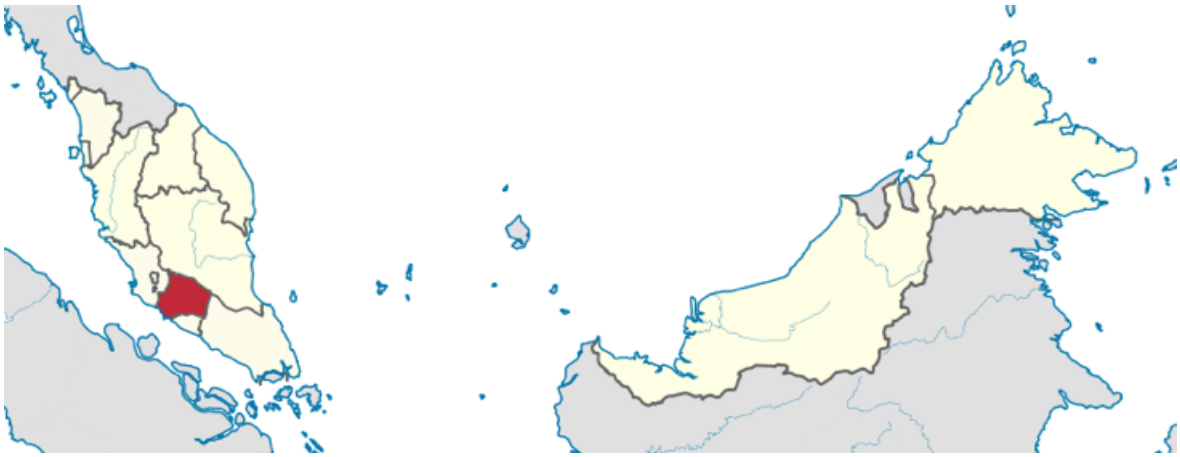


Figure 10 - Negeri Sembilan state, in Malaysia

Figure 11 - Seremban district highlighted in Negeri Sembilan map

The information presented in **Table 3** shows the number of people infected by year ('Year' attribute), as well as the relative percentage, which shows that 2003 was a particularly bad year in terms of dengue infections, with 26% of the people being infected in 2003, from a total of 7 consecutive years (2003-2009). Although it has been decided that this attribute will not be taken in consideration in this study, it is interesting to know how the virus has affected this area of Malaysia throughout the years.

Year	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*

Frequency	1588	998	547	534	633	960	816
%	26	16	9	9	10	16	13

Table 3 – Occurrences of DF/DHF per year

A different perspective about the data grouped by year can be seen in **Table 4**. This table presents the incidence of DF and also DHF in Seremban, and these data, coming from the dataset, shows that almost 96% of the dengue cases in Seremban from 2003 to 2009 corresponded to DF and that the remaining were DHF or evolved into such malady.

	DF	*DHF*
2003	1496	92
2004	972	26
2005	537	10
2006	519	15
2007	623	10
2008	929	31
2009	731	85
Totals	5807	269
Total	6076	

Table 4 - Occurrences of DF and DHF in the dataset throughout the years

Likewise, **Figure 12** shows the occurrences of DF and DHF cases by grouping them by another attribute that will not be considered further on: the 'Week' attribute. It is interesting to notice that there are many more people getting infected in the first 3 weeks of the considered years, 2003-2009. These data gets even more interesting when considering that the Southwest monsoon normally occurs from late May to September. Nonetheless, the Southwest monsoon is not any more severe than the rest of the year on the west coast.

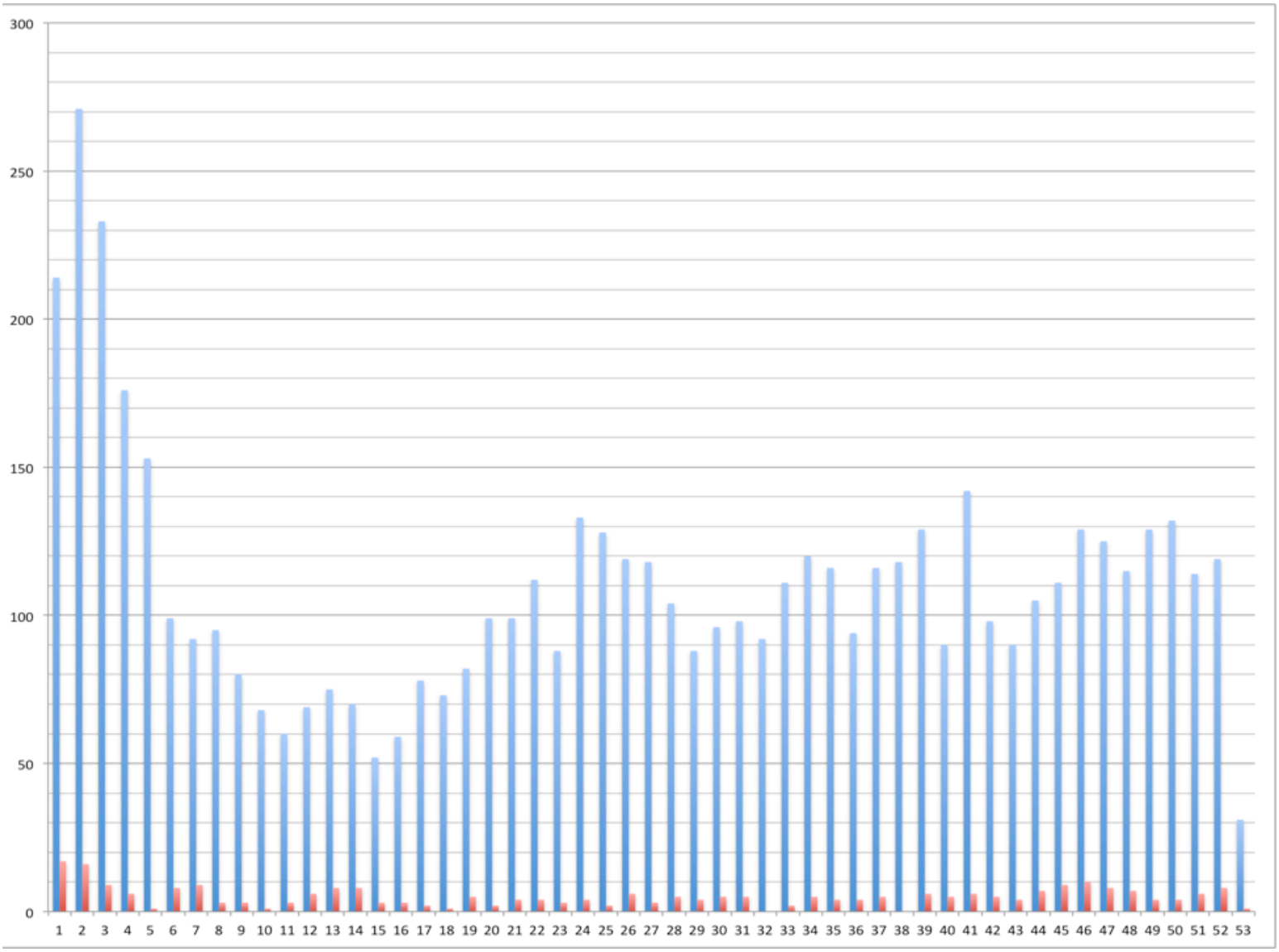


Figure 12 – Occurrences of DF and DHF grouped by week

Information about the remaining continuous attributes is shown in **Table 5**, which contains the means, the highest and the lowest values recorded for the attributes 'DF', 'DHF', 'Average temperature', 'Humidity' and 'Rainfall'.

	Mean	*Lowest*	*Highest*
DF	15.91	1	137

DHF	0.737	0	9
Average temperature	26.66	24.86	29.06
Humidity	76.24	60.91	84.92
Rainfall	43.54	0.00	173.74

Table 5 – Means and lowest and highest values recorded for the continuous attributes

Every decennial, the Malaysian department of statistics performs a census. The last one was done in 2010 (Department of statistics Malaysia, 2010), and much of the statistics that will be presented further on are taken from this census, providing all the statistics concerning age groups, gender distribution, and data related with the races of the people living in Malaysia.

The distribution of the different age groups in the dataset is shown in **Table 6**. Albeit it was not possible to find data related to the district of Seremban, the distribution of people in Negeri Sembilan by age in 2010 (**Table 7** and **Table 8**) will give an approximate estimate to the age distribution of people in Seremban. As the reader might have noticed, the age groups used in the dataset are not compatible with the data obtained in (Department of statistics Malaysia, 2010), because in the dataset, children until 12 stay in one category, and from 13 and above they stay in another, while the data extracted from the aforementioned work arranges people in equal groups of 5. It was therefore decided to consider 3/5 of the people aged from 10 to 14 as being 'Children' (0 - 12) and 2/5 of the same people as being 'Youth' (13-29). The new distribution of people living in Negeri Sembilan in 2010 by age groups is depicted in **Table 9**. It is interesting to notice that the percentage of infected youth people in the dataset is much higher (44%) than the percentage of youth living in Negeri Sembilan (32,8%), and the percentage of infected adults in the dataset (40%) is also slightly higher than the percentage of adults living in Negeri Sembilan (35,5%). Also, the percentage of children infected in the dataset (12%), as well as the percentage of seniors (4%) is much less than, respectively, the percentage of children (22,8%) and seniors (9) that live in Negeri Sembilan.

Age group	*Children (0-12)*	*Youth (13-29)*	*Adult (30-59)*	*Senior Citizen (60+)*
Frequency	746	2654	2445	231
%	12	44	40	4

Table 6 – Frequency and percentage of the 'Age group' attribute

Age	*0-4*	*5-9*	*10-14*	*15-19*	*20-24*	*25-29*	*30-34*	*35-39*
Freq.	78.325	95.620	97.269	106.786	104.621	83.872	72.305	65.268

Table 7 – Distribution by age of people until 39 years old in Negeri Sembilan

Age	*40-44*	*45-49*	*50-54*	*55-59*	*60-64*	*65-69*	*70-74*	*75+*
Freq.	61.299	60.597	56.922	46.260	33.497	21.548	17.537	19.338

Table 8 – Distribution by age of people over 39 years old in Negeri Sembilan

Age group	*Children (0-12)*	*Youth (13-29)*	*Adult (30-59)*	*Senior Citizen (60+)*
Frequency	232306	334187	362651	91920
%	22,8	32,8	35,5	9

Table 9 – Distribution of people by age group in Negeri Sembilan

The gender distribution in the dataset is presented in **Table 10**, which shows that an interesting 60% of the people that were infected with a variant of dengue are men.

Gender	*M*	*F*
Frequency	3673	2403
%	60	40

Table 10 – Frequency and percentage of the 'Gender' attribute

According to the national department of statistics (Department of statistics Malaysia, 2010), there's a ratio of 105 males for every 100 females in Malaysia. Notwithstanding a male majority within the country population, that number falls short (roughly 51%) when compared to the 60% of the people infected in Seremban being males. If the district of Negeri Sembilan is taken in consideration, whose population in 2010 was living, mostly, in Seremban (**Table 11**), the gender

distribution in the whole nation is similar to the gender distribution in Negeri Sembilan (**Table 12**), where 51.58% of the population consists of males. A similar gender distribution would also be expected in Seremban, and the data presented in **Table 13** confirms this proposition, with a 51.9% of male population in Seremban. These data confirms that men in Negeri Sembilan, more precisely in Seremban, are more easily infected with dengue than women, because although the gender distribution is somewhat equal, 60% of the people infected between 2003 and 2009 were male. One should therefore take in consideration that the socio-demographic data were obtained in the 2010 census, and the data related with the people infected with dengue in Negeri Sembilan were registered between 2003 and 2009.

	2000	*2010*
Jelebu	37.194	39.200
Kuala Pilah	63.541	66.092
Port Dickson	106.630	115.361
Rembau	36.848	43.011
Seremban	383.530	555.935
Tampin	77.021	84.889
Jempol	125.010	116.576
Total	829.774	1.021.064

Table 11 - Distribution of population within the towns of Negeri Sembilan

	2000		*2010*	
Gender	Male	Female	Male	Female
Frequency	424.132	405.642	515.293	481.778
%	51,11	48,89	51,68	48,32

Table 12 - Gender distribution in Negeri Sembilan

	Male	*Female*
Population	288.332	267.603

%	51,9	48,1

Table 13 - Gender distribution in Seremban

The distribution of races in the infected people, as given by the dataset, is depicted in **Table 14**. The values presented are very uneven, but so is the population distribution in Malaysia. According to (Department of statistics Malaysia, 2010), there are 56% of Malays in Negeri Sembilan, and even less in Seremban (51.2%), thus having 60% of the people infected in Negeri Sembilan being Malay might indicate that this race is more predisposed to infection. The values presented in **Table 15** and **Table 16**, referring respectively to the distribution of the population in Negeri Sembilan and in Seremban, show also that much less percentage of Chinese people are infected in Seremban than the actual percentage of Chinese people living there. The percentage of Indians registered as being infected with dengue in Negeri Sembilan is slightly bigger when compared with the percentage of Indians living in Seremban, but people from other races show a lesser predisposition for infection, because although 7% of people from other races were registered as infected in the dataset, there were actually 8.8% of them living in Seremban in 2010. Once again, one should take into consideration that the socio-demographic data were obtained in the 2010 census, and the data related with the infected people refers to the period between 2003 and 2009.

Race	*Malay*	*Chinese*	*Indian*	*Other races*
Frequency	3650	1024	977	425
%	60	17	16	7

Table 14 – Frequency and percentage of the 'Race' attribute

Race	*Malay*	*Chinese*	*Indian*	*Other races*
Frequency	572006	223271	146214	79573
%	56	21,9	14,3	7,8

Table 15 – Population distribution in Negeri Sembilan according to race

Race	*Malay*	*Chinese*	*Indian*	*Other races*
Frequency	284564	134572	87663	49136
%	51,2	24,2	15,8	8,8

Table 16 – Population distribution in Seremban according to race

Table 17 shows the job categories carried out by the people infected with dengue. The distribution of people infected with dengue by town is shown in **Table 18** while the distribution by district is presented in **Table 19** and the distribution of the different types of epidemics in **Table 20**.

Job	*General Worker*	*Retired*	*Children*	*Executive*	*Housewife*	*Independent*	*Non-executive*	*Student*	*Unemployed*
requency	1236	176	579	287	572	303	1465	1158	300
%	20	3	10	5	9	5	24	19	5

Table 17 – Frequency and percentage of the 'Job' attribute

Town	*Bdr. Seremban*	*Ampangan*	*Labu*	*Lenggeng*	*Nilai*	*Pantai*	*Rantau*	*Rasah*	*Seremban*	*Setul*
Frequency	4	2495	321	54	7	11	980	928	405	871
%	0	41	5	1	0	0	16	15	7	14

Table 18 – Frequency and percentage of the 'Town' attribute

District	*SN*	*PD*
Frequency	6076	5
%	100	0

Table 19 – Distribution of the 'District' attribute

Type of epidemic	*DKW*	*MWB*	*TKW*
Frequency	1925	376	3775
%	32	6	62

Table 20 – Frequency and percentage of different types of the 'Epidemic' attribute

The analysis of the data and the computed statistics has lead to the formulation of several hypotheses. The necessity of water for the eggs to hatch had to be taken in consideration, as well as the fact that *aedes aegypti* takes 4 to 13 days to develop from egg to an adult mosquito and the intrinsic incubation period of dengue takes 3 to 14 days. Consequently, the time since eggs are laid, until the people that got infected by the mosquitoes that were born from these eggs start to show symptoms vary from 7 to 27 days. If the average incubation period is considered, then it takes from 8 to 20 days. The hypotheses were formulated based on this knowledge and choices, and are presented next:

1. Is there any relation between rainfall in a week and the number of DF/DHF cases in the same week, and 1 and 2 weeks later?
2. Is there any relation between humidity in a week and the number of DF/DHF cases in the same week, and 1 and 2 weeks later?
3. Is there any relation between temperature in a week and the number of cases of DF/DHF cases in the same week, and 1 and 2 weeks later?
4. Can someone's race influence the ease with which he/she contracts DF/DHF in the presence of an outbreak?
5. Does the gender of a person influence the ease with which he/she contracts DF/DHF in the presence of an outbreak?
6. Can someone's age group influence the ease with which he/she contracts DF/DHF in the presence of an outbreak?
7. Does someone's job category influence the ease with which he/she contracts DF/DHF in the presence of an outbreak?

The formulation of hypotheses is essential for the next step in the methodology, but this step of the methodology would not be completed without reviewing the DM goals, based on the newly

acquired knowledge about the data. Throughout the research, these goals had to be changed, as well as the hypotheses, due to inconsistencies or incompatibilities that have arisen and lead to new iterations of the methodology - these attempts will be mentioned later in the **Error! Reference source not found.**.

3.2.4 Verifying Data Quality

It is rare for the data within a dataset to be perfect. A thorough analysis must be performed at the end of the Data understanding before moving through the next stage, the Data preparation. Problems like missing data, data errors like typographical errors, measurement errors, coding inconsistencies (for example, using F and female for gender) and mismatches between the apparent meaning of a field and the meaning stated in a field name or definition.

As previously stated in this dissertation, there are no blank fields in the dataset. Notwithstanding, some entries have 'NIL' as a job category, and that should be dealt with accordingly in the next phase of the methodology.

While Exploring Data with R, some typing errors have been discovered, namely 'MELAYSIAN' as a race and the 'Apidemic' attribute name. Besides that, some entries had incorrect values for the weather data. These mistakes were coming from copying it within a worksheet incorrectly. That was easy to spot because within a week, the temperature, rainfall and humidity are expected to be the same (corresponding to the average values recorded in that week), and in some entries that was not happening.

Coding inconsistencies have also been spotted in the dataset. 'Mantin' and 'Setul', two town names showing up in the 'Address_id' column, are actually the same. Having 'Senior citizen' as a job category can also be considered a coding inconsistency. Some examples of metadata are also present in the dataset. The attribute 'Address_id' is misleading, because the values are used are not identifiers (IDs), they are actually town names.

3.3 Data preparation

Data preparation is probably the most important phase of the CRISP-DM methodology. It is the point where the data engineer finally starts to manipulate the data, preparing and packaging it for mining. Under normal circumstances, most of the project's time would be spent in preparing the data, but a previously performed data preparation already within this research scope, and the thorough Business understanding and Data understanding stages performed earlier, have already minimised this overhead.

The data available for this dissertation work, as already mentioned, had already been pre-processed beforehand. Data cleaning (or data cleansing), in which corrupt or inaccurate records are detected and corrected, was applied to the datasets in several ways, like removing inaccurate rows and translating the attribute names and values from Malay to English. Decimal values were rounded to two decimals to simplify the process and because it was not necessary to have so much precision, and that is also a data cleaning procedure. Another pre-processing task utilised was attribute selection (also known as feature selection), where redundant attributes are removed and a subset of relevant features (variables/attributes) is selected. The original demographic dataset alone had 134 attributes (can be consulted in appendix a), and the dataset with which this work started had only 14 attributes, including the weather related variables, coming from the second dataset. Data transformation, in the form of data aggregation, was also applied on the weather dataset, in order to calculate the means for temperature, humidity and rainfall for each week. The demographic dataset had also its data aggregated, in order for the number of DF and DHF cases per week to be calculated. The two distinct datasets were then merged, and that is a data integration pre-processing scheme. In order for the merger to succeed, the number of rows from the demographic dataset was maintained, having been added to the final dataset the following attributes, whose calculation had already been explained:

- Number of DF cases per week
- Number of DHF cases in a week
- Mean temperature for a week
- Mean humidity within a week
- Mean rainfall for a week

Therefore, for each week, the number of DF and DHF cases were calculated, and then replicated as an attribute for each entry within the same week. Also for each week, the average numbers for temperature, humidity and rainfall were also replicated for every dataset entry. Notwithstanding the exhaustive work done over the datasets before the start of this research, the pre-processing schemes were still far from being complete. By following the CRISP-DM methodology, several other pre-processing tasks had to be done, and will be explained in the remaining of this chapter.

3.3.1 Selecting Data

Attributes that are not relevant for the study, as well as records that bring no added value, shall be disregarded during this step. Some of those attributes have already been identified in earlier stages: 'Year' and 'Week'. The first one is not interesting because it is too broad. All the information that could be obtained by analysing the 'Year' attribute was already fully explored in the previous section. The latter would be much more interesting to study than the 'Year' attribute, but since it is expectable that the same week in different years can have a totally different weather, it has been decided to discard this attribute as well.

The 'Epidemic' attribute was also discarded, because contrary to the intuition that this attribute would properly classify the epidemics, after thoroughly analysing the data in the previous chapter, the conclusion was that its values were not making sense and therefore could not be interpreted. Since the methodology had to be restarted due to this failure, a better insight into the problem is left for the **Error! Reference source not found.**.

Another attribute that shall not be considered further on is the 'Town' attribute because no statistical information regarding the demographics in each of the considered towns could be obtained, and especially due to the size of the dataset, which is manifestly small.

Not only attributes were discarded at this stage, since five rows had to be removed from the dataset. These entries, which are all that have the value 'PD' as a district (**Table 19**), were removed because although PD (Port Dickson) is in fact a district in Negeri Sembilan (as Seremban, where all other cases occurred), the towns in these five rows were located in Seremban. Therefore, it was concluded that these five entries are inconsistent and it has been decided it was better to remove them. Analogously, the "District" column was also removed, because after

removing the aforementioned entries, all entries were coming from the same district, Seremban, thus no added value was coming from having this attribute.

After the removal of the aforementioned five entries, the dataset started to have 6076 entries, which is exactly the sum of all the DF and DHF cases, which indicates that those entries were, as presumed, errors. Besides that, in the weeks containing the five rows the number of dengue cases was inaccurate, and only after their removal the number of entries related to those weeks started to match with the recorded number of occurrences.

Although according to the methodology one should only select data at this stage, an attentive reader would have noticed that the statistics presented earlier were already disregarding the 5 rows that only now were eliminated, except the ones presented for the 'District' attribute (**Table 19**). That approach was taken to avoid showing the statistics twice, therefore it has been decided to, whenever possible, show just the statistics for the dataset immediately prior to the modelling stage.

3.3.2 Cleaning Data

Now it is time to take a closer look to the problems identified earlier, while Verifying Data Quality. In the 'Job' column, the value 'NIL' was replaced by 'Unemployed', because this value was used for the unemployed people and NIL could be misunderstood. Translation errors, coming from the previously incomplete work done beforehand, were also corrected, namely converting the race 'MELAYSIAN' to 'MALAY' and the attribute name 'Apidemic' to 'Epidemic'. The entries with incorrect values for the weather data were also corrected. In the 'Town' attribute, 'Setul' has replaced 'Mantin', and 'Retired' replaced 'Senior citizen' in the 'Job' category column. At last, the attribute 'Adress_id' was renamed to 'Town'. Although some of these changes will not have any influence on the outcome because some attributes were discarded, all the changes made during the iterations of the methodology shall be mentioned.

3.3.3 Constructing New Data

It is common that new data has to be generated before the mining process. There are two ways in which new data are constructed: by deriving attributes or by generating records. For this

dissertation work, the target attribute will be a binary attribute derived from other attributes. It was named 'Outbreak', and is based on the definition of a dengue epidemic by experts. For a specific week, the attribute is generated by calculating the average of dengue cases, both DF and DHF, between the two previous weeks ('Week') and comparing the values with the number of dengue cases in the current week. If the number of dengue cases for the current week is higher than the average number of cases for the two previous weeks, then it is classified as an outbreak (Y), otherwise it is classified as non-outbreak (N). An example is given below, by using the data from **Table 21**. There was no need to generate records in this research.

Week	*DF*	*DHF*
1	34	5
2	15	1
3	75	21
…	…	…

Table 21 - Target class ('Outbreak' attribute) calculation example

The calculation of the target attribute for the 3rd week is done in two steps:

1. Calculate the mean of dengue cases (DF+DHF) of the 2 previous weeks:

   (34+5+15+1) / 2 = 27.5

2. Compare the value obtained with the number of dengue cases in the current week. If the obtained value is higher, then the target class is classified as N, otherwise it is classified as Y.

After deriving the new attribute, the distribution of its values can be consulted in **Table 22**, aggregated by week, and in **Table 23** without aggregation.

Outbreak?	*Y*	*N*
Frequency	186	179
%	51	49

Table 22 – Distribution of values for the 'Outbreak' attribute per week

Outbreak?	*Y*	*N*
Frequency	3732	2344
%	61	39

Table 23 - Distribution of values for the 'Outbreak' attribute without aggregation

After generating the new attribute, the number of DF and DHF cases can also be discarded from the dataset.

But the creation of this attribute is not enough to properly answer the enunciated hypotheses. In order to study the effects of the weather (temperature, humidity and rainfall) in the existence, or not, of an outbreak 1 and 2 weeks after, it was necessary to shift the newly created attribute by 1 and 2 weeks, therefore creating 2 new attributes. All the attributes that were to be studied here were weekly, thus 3 new datasets were created, having just 1 row per week and with the non demographic attributes, namely 'Year', 'Week', 'DF', 'DHF', 'Average temperature', 'Humidity', 'Rainfall' and 'Outbreak'. These datasets have the 'Outbreak' attribute shifted in 0, 1 or 2 weeks and, respectively, 365, 364 and 363 rows. From this moment on, these datasets will be called 'week0', 'week1' and 'week2' datasets, while the first one will be named non-demographic dataset.

3.3.4 Integrating Data

When applying the CRISP-DM methodology, it is also frequent to have multiple data sources for the same entities. If the sources have the same unique ID, they can be merged into a single data source, or they can even be appended into one of the datasets, if the multiple initial datasets have similar attributes. This step had already been done, when the demographic data was merged with the weather data, and since the outcome was a single dataset, there is no further need to integrate data in this study.

3.3.5 Formatting Data

The final step before modelling consists in assessing if the DM algorithms that will be applied to the dataset require a particular format or order to the data. In (IBM Corporation, 2011) the example of a sequence algorithm is given. Even if the model can perform the sorting for you, it may save processing time to use a sort function prior to applying the algorithm.

In order to apply the chosen techniques, it is not necessary to apply any formatting to the data, therefore nothing had to be done on this methodology step.

3.3.6 Implementing

Throughout this section, the changes done over the dataset have been described. Nonetheless, the way they were implemented was not referred to so far, therefore the purpose of this final subsection is to document those modifications and the tools that were used.

As stated several times in this document, R was the DM tool used for this dissertation. But in order to perform some of the aforementioned changes, it was better to use Excel because of its simplicity. Below those modifications are listed, in chronological order:

1. All the typographical errors or inconsistencies were corrected directly in Excel.
2. Using a simple Excel formula, the new attribute was easily created.
3. The 5 erroneous rows were then removed.
4. The new datasets with non-demographic data only and with the 'Outbreak' attribute shifted by 0, 1 and 2 were created.

This left little work to be done in R, since only the removal of attributes was still to be done. The code in R can be found below, where the remaining non-demographic attributes are also discarded, in order to proceed with the modelling in the first dataset.

```
# Read the main dataset, stored in dengue.csv file, into a data structure named 'dat_demographic'
dat_demographic = read.csv("dengue.csv", header = TRUE)

# Discard attributes: 'Year', 'Week', 'DF', 'DHF', 'Town', 'District', 'Epidemic'
dat_demographic$year = NULL
dat_demographic$week = NULL
dat_demographic$df = NULL
dat_demographic$dhf = NULL
dat_demographic$temp_avg = NULL
dat_demographic$humidity = NULL
dat_demographic$rainfall = NULL
dat_demographic$town = NULL
dat_demographic$district = NULL
dat_demographic$epidemic = NULL
```

Chapter 4

4 Modelling the data

It is at this stage that practical results appear, and the "hard work starts to pay off". It is expected that several iterations will have to be done in order to fine-tune the parameters, and even to go back to the Data preparation phase so that the data can be properly prepared for models and parameters that, in the previous stages, were not thought to be necessary.

4.1 Selecting Modelling Techniques

The selection of modelling techniques is preceded by a reflection about the data types available for mining, the DM goals and the specific modelling requirements. The DM goals did not change since its enunciation, therefore the modelling techniques were all previously identified in the DM goals section: DTs, clustering and ARs mining.

A DT is a hierarchical structure that represents a classification model. Internal tree nodes correspond to splits applied to decompose the domain into regions, and terminal nodes assign class labels to regions believed to be sufficiently small or sufficiently uniform. For convenience, the term node will be reserved to internal nodes only, and terminal nodes will be referred to as leaves (Cichosz, 2015).

The basis of operation of any DT-based algorithm is the "divide and conquer" strategy. These algorithms successively divide the problem into several sub problems with a smaller number of dimensions, until a solution for each of the simpler problems is found. Having this strategy as a principle, the classifiers based on DTs divide the data into subgroups, creating nodes for each of them, until all the observations within each node address only one class or until one of the classes shows a clear majority, therefore not justifying further divisions. In this situation, a leaf containing the class majority is generated, and it is said that the stopping criteria has been reached for that node, therefore becoming a leaf.

A DT algorithm starts with all the data at the root node and scans all the variables for the best one to split on. This is measured by applying a splitting criterion to all the variables and each of its values, testing the purity of each possible node. This results in the most pure nodes below it, i.e., containing a majority of observations of one of the possible values of the dependent variable.

DTs have several advantages. They are what is known as a glass-box model, because after the model has found the patterns in the data and grown the tree, one can see exactly what decisions will be made for data that is still to be predicted. They are also very intuitive and can be read by people with little experience in machine learning after a brief explanation. Finally, they are the basis for some of the most powerful and popular machine learning algorithms (Stephens, 2014).

But DTs have some drawbacks as well: they are greedy. It is often the case that when selecting the most pure node it will not lead to the better, more pure tree. Notwithstanding, a DT algorithm will always make the optimal decision by taking only its two sons in consideration, and is unable to go back as new nodes are grown. This greedy algorithm is used because exploring every possible version of a tree is extremely computationally expensive (Stephens, 2014).

A cluster is a group of objects belonging to the same class. The members of a cluster are more like each other than they are like members of other clusters. Thus the task of grouping similar objects from a dataset in one cluster and dissimilar objects in a different one is called clustering. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high (Oracle, 2008).

Similarly to classification (a DT is an example of a classification technique), clustering is used to segment data. But unlike classification, clustering models divide the data into groups that were not previously defined. Classification models segment data by assigning it to previously defined classes (target classes) while in clustering targets are not used (Oracle, 2008).

The main advantages of clustering over classification are its capability to single out useful features that distinguish different groups, and its adaptability to changes. Being too sensitive to outliers is the biggest drawback of these techniques.

4.2 Generating Test Designs

The generated DTs should be complex enough to allow useful information to be obtained from them, and simple enough to make sense from a business perspective. The test design that will guide the modelling process is presented below.

DT test design

1. Divide the dataset in training and testing data.
2. Fit the tree model for the training dataset with the 'rpart' package.
3. Predict the dependent variable against the testing set.
4. Obtain statistics about the prediction and the previously grown tree, like the percentage of accurate predictions and the complexity parameter (CP) table of the tree.
5. Improve the first tree by calibrating its parameters and, ultimately, growing a final tree.

As previously mentioned, the generated clusters should have a low inter-cluster similarity and a high intra-cluster similarity. The test design for clustering is as follows.

Clustering test design

1. Remove the 'Outbreak' attribute, because it is the attribute that will be predicted.
2. Convert the non-numeric attributes in numeric attributes.

3. Rescale the attributes.
4. Assess which algorithm, K-means or PAM (Partitioning Around Medoids) is appropriate to derive the model.
5. Apply the chosen algorithm to the dataset in order to create the clustering model.
6. Tune the model, until an optimal model has been found.

4.3 Building the Models

DT growing

As previously mentioned, the chosen models will be derived based on the 4 datasets that have been constructed from the original one. Starting with the non-demographic dataset, the age category, gender, race and job category of the infected people (independent attributes) have been used to predict the dependent attribute, 'Outbreak'. The DTs grown from the demographics dataset indicate the predominance of the studied demographics, namely race, job category, gender and age group of people infected with the disease in weeks where there has been an outbreak, and in weeks where an outbreak did not occur.

The dataset was first divided in training and testing sets, and the R code used for that is shown below:

```
smp_size <- floor(2/3 * nrow(dat_demographic))
set.seed(123)
train_ind <- sample(seq_len(nrow(dat_demographic)), size = smp_size)
train <- dat_demographic[train_ind, ]
test <- dat_demographic[-train_ind, ]
```

The next step consists in growing the tree, by fitting the tree model to the training set. For that effect, the following R code was used:

```
formula = outbreak ~ age+gender+race+job
tree <- rpart(formula, data=dat_demographic, method="class")
```

Subsequently, the DT was visualized with the 'prp' function, using the code presented below.

```
prp(tree, extra=101, type=3, digits=4, fallen.leaves=TRUE, faclen=0)
```

The tree that has been grown is shown in **Figure 13**.

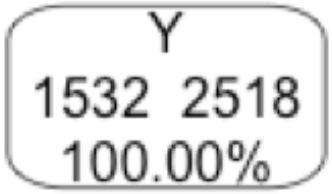


Figure 13 - DT grown with the default 'rpart' parameters for the demographic dataset

As can be seen in the figure, the newly grown tree has a single node, and classifies all occurrences as being an outbreak. This happened because the 'rpart' routine has decided that the best model it can find is the "intercept only" model, i.e., a tree with no branches at all. Hence it is necessary to calibrate the parameters. The 'rpart' package allows users to parameterise DTs in several ways, but throughout all the DT modelling process only three of them revealed worthy of tuning: 'minsplit', 'minbucket' and 'cp'. The first is the minimum number of observations in a node for which the routine will even try to compute a split. It defaults to twenty. The second corresponds to the minimum number of observations in a terminal node, and its default value is minsplit/3. The latter (CP) is the amount by which splitting a node improved the relative error, and defaults to 0.01. The splitting criteria that R uses by default is the gini index, and it allows users to use also the information gain criteria, but the outputted trees were not improved by using the latter, so all the DTs that have been grown used the gini index (Therneau and Atkinson, 2015).

Several adjustments had to be made to the single node tree (**Figure 13**), previously grown with the default 'rpart' parameters, and after executing the R code shown below, a promising DT has been obtained (**Figure 14**). This tree, as well as all DTs grown from the demographics dataset, indicates the predominance of certain demographics, namely race, job category, gender and age group of people infected with the disease in weeks where there has been an outbreak, and in weeks where an outbreak did not occur.

```
tree <- rpart(formula, data=train, method="class", control=rpart.control(cp=0.0005, minsplit=50))
```

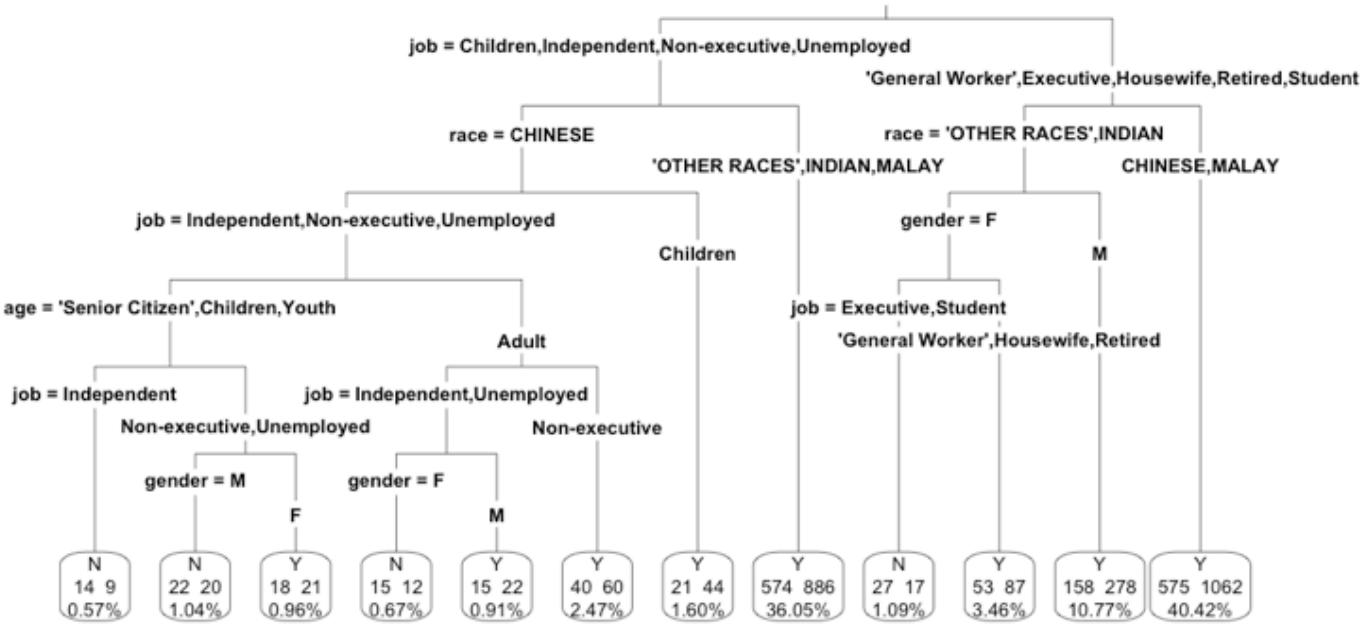


Figure 14 - DT grown after calibrating for the demographic dataset

The tree was visualized by using the 'prp' command again. The tree leaves are plotted in round squares, containing the outcome value, the percentage of rows that fall in that leaf, as well as the number of correctly and incorrectly classified entries. When the outcome is 'Yes', the number of correctly classified entries is to the right, and to the left when the outcome is 'No'. This particular tree has 12 leaves and 6 levels, and classified correctly 2538 cases but failed to classify 1512 cases. It is now time to predict the dependent variable against the testing set, by using the 'pred' command.

```
pred = predict(tree, test, type="class")
```

The prediction obtained with the 'predict' command is fairly unbalanced, having only predicted 66 cases as occurring during a non-outbreak week, but an outstanding majority of 1960 as having occurred during an outbreak. Notwithstanding, the percentage of accurate predictions is greater than 60%. Following, the reader can consult the commands used to calculate all the statistics

mentioned in this paragraph about the prediction. The last command has further outputted a table with the True Negatives (TN), False Negatives (FN), True Positives (TP), and False Positives (FP), which can be seen in **Table 24**.

```
summary(pred)
mean(pred == test$outbreak)
table(pred, test$outbreak)
```

True value	*N*	*Y*
Prediction: N	37	29
Prediction: Y	775	1185

Table 24 – Predicted TN, FN, TP and FP with calibrated parameters for the demographic dataset

The tree that has been grown is not satisfactory at all. Since the vast majority of predictions in the test dataset is 'Y', the tendency would be to increase the tree size by reducing the value of the CP threshold, in order to obtain more end nodes with value 'N'. After trying this approach, despite the predictions had been more balanced, the difference was minimal and the generated trees were extremely large. Plus, the leaves with value 'N' would typically have very few observations. This was happening because the data are distributed in a very well balanced manner.

As an example, consider the training dataset. It is known that in the training set there are 2518 entries with the value 'Y' in the 'Outbreak' variable and the remaining 1532 have 'N' as the value in the 'Outbreak' column, which makes a total of 62% of entries during a dengue outbreak. If this dataset is divided into two datasets, one with all the rows having 'Y' as the value of the 'Outbreak' attribute and the second with 'N' as the value of 'Outbreak', and if it is considered for each of those datasets, for example, the percentage of Malaysians, 63% of them lie in the dataset whose 'Outbreak' attribute value is 'Y'. This percentage balance between the demographics of people who went to hospital with symptoms of DF/DHF in weeks where an outbreak occurred and the demographics of people who got infected with dengue in weeks where this did not happen, and knowing beforehand that there is a higher prevalence of 'Y' in the 'Outbreak' column of the training dataset (derived naturally from a greater prevalence of 'Y' in the variable 'Outbreak' also in the original dataset) results in an overwhelming majority of forecasts of 'Y' in relation to 'N' for the outcome variable.

Constructing a model for the 'week0' dataset started with growing the tree with the 'rpart' package, after having divided the dataset in training and testing datasets. Since the commands used are the same as the ones used with the non-demographic dataset, having changed only the formula and the dataset used, the new commands will not be mentioned. The tree that has been generated with the default 'rpart' parameters is depicted in **Figure 15**.

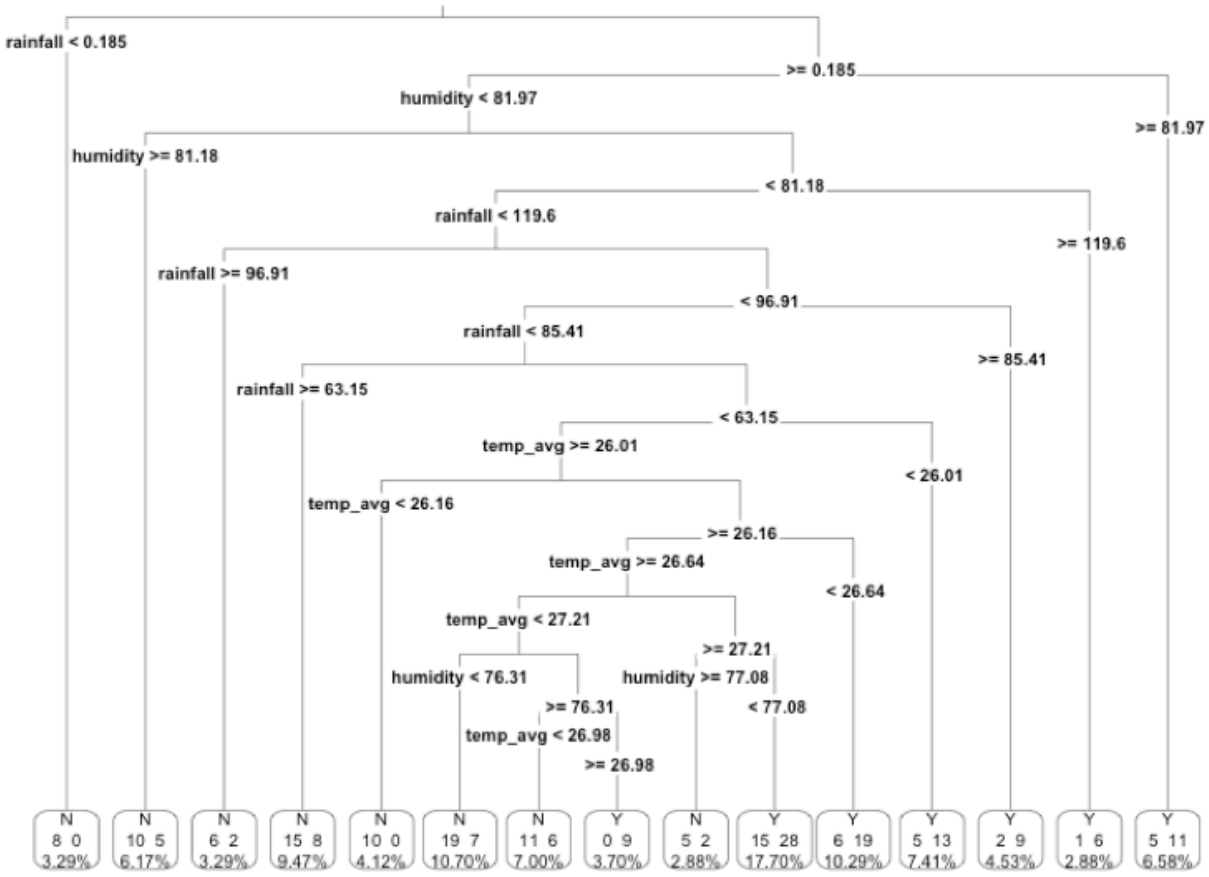


Figure 15 - DT grown with the default 'rpart' parameters for the 'week0' dataset

This tree has 15 leaves and 13 levels, which is quite a deep tree, and classified correctly 179 cases but failed to classify 64 cases. When the 'predict' command is used to obtain a prediction about the dependent variable in the testing set, the output shows that 48 cases were classified as not

happening in a week with an outbreak, and 74 were classified positively. Nonetheless, the percentage of accurate predictions is only 50%. The TN, FN, TP and FP can be seen in **Table 25**.

True value	*N*	*Y*
Prediction: N	24	24
Prediction: Y	37	37

Table 25 – Predicted TN, FN, TP and FP with default parameters for the 'week0' dataset

A few adjustments were undertaken in the tree grown with the default 'rpart' parameters, until a promising tree was found. With the CP threshold at 0.03, the minimum number of cases in a leaf at 3 and having at least 8 observations to consider a split, **Figure 16** presents such tree.

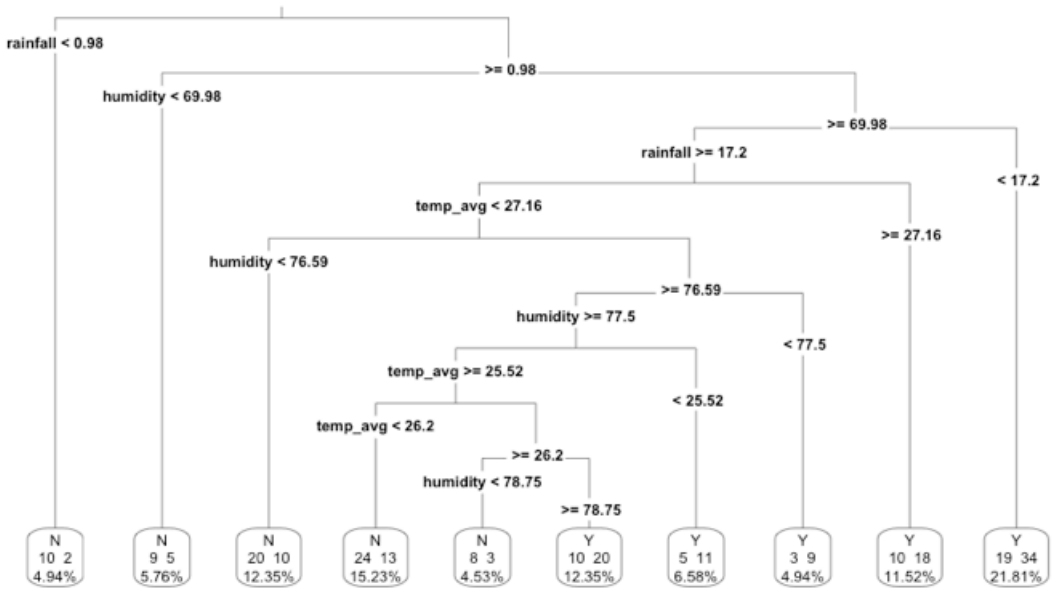


Figure 16 - DT grown after tuning for the 'week0' dataset

The DT grown after calibrating has 10 leaves and 9 levels, and was able to classify correctly 163 cases, having failed to classify 80. By predicting against the test set, the model chose an outcome of no outbreak in 21 situations and the opposite for 101 observations, making a good guess in

almost 55% if these predictions. **Table 26** shows data related to the TN, FN, TP and FP of the DT grown after calibrating when attempting to predict against the test set.

True value	*N*	*Y*
Prediction: N	24	24
Prediction: Y	37	37

Table 26 – Predicted TN, FN, TP and FP with tuned parameters for the 'week0' dataset

Devising a model for the 'week1' dataset followed the exact same steps that have been taken to generate the two previous models. The tree that has been grown with the default 'rpart' parameters is shown in **Figure 17**. It has 15 leaves and 10 levels, contains 179 correct classifications and 64 incorrect classifications.

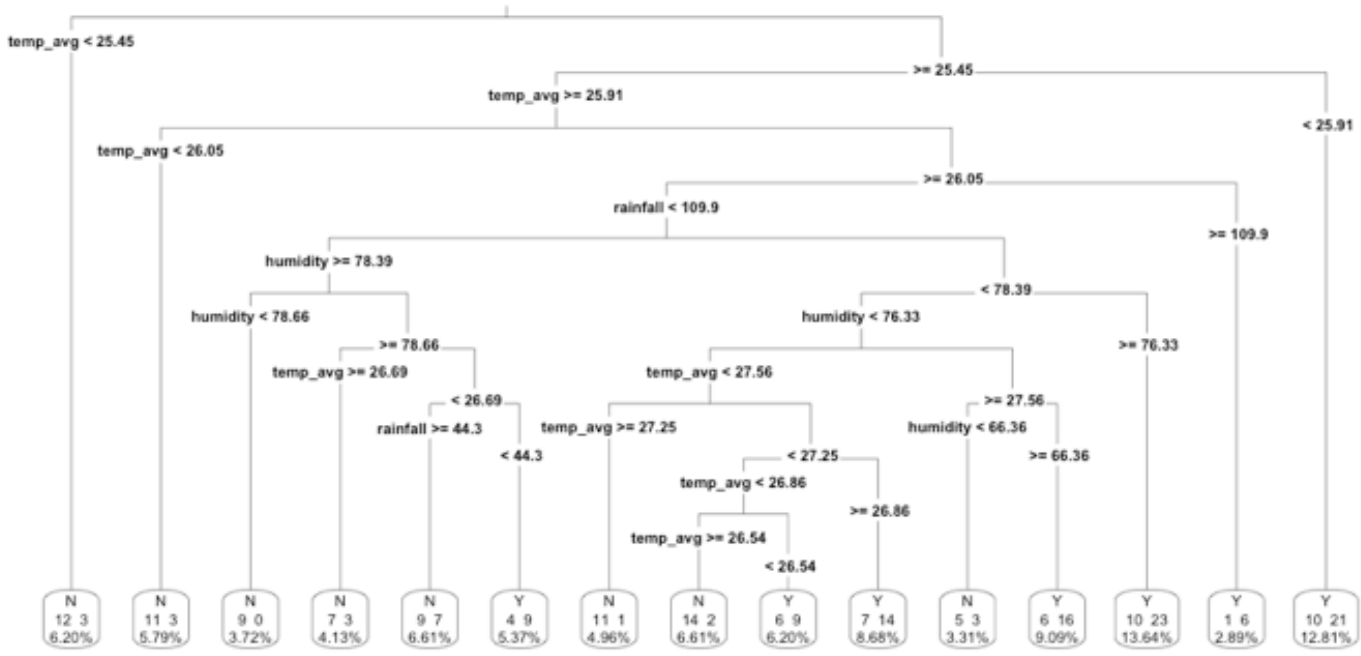


Figure 17 - DT grown with the default parameters for the 'week1' dataset

When the DT model was tested against the testing dataset, its predictions were 44 negative and 78 positive, and has managed to classify correctly 55% of the observations.

True value	*N*	*Y*
Prediction: N	23	21
Prediction: Y	34	44

Table 27 – Predicted TN, FN, TP and FP with default parameters for the 'week1' dataset

The tree that has been devised after calibrating is shown in **Figure 18**. It has been found with the CP threshold at 0.025, and no other change in the default 'rpart' parameters.

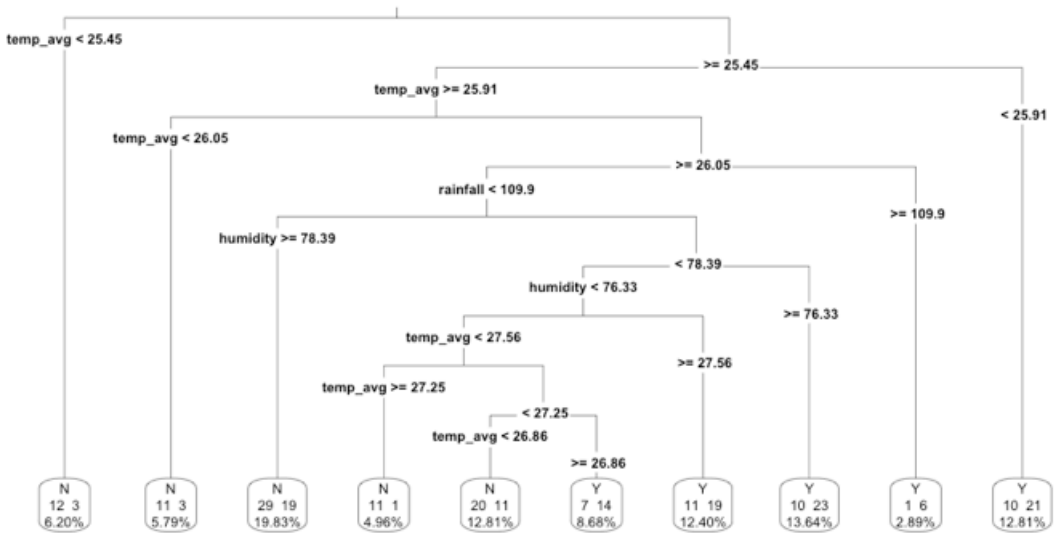


Figure 18 - DT grown after tuning for the 'week1' dataset

This tree has 10 leaves and 9 levels, classified correctly 163 observations and failed to classify 80. When the accuracy of the model was tested against the testing set, it chose an outcome of no outbreak in 56 situations and in 66 it chose the opposite. From these predictions, more than 56% were accurate. The following table presents the TN, FN, TP and FP of the predictions.

True value	*N*	*Y*
Prediction: N	30	26
Prediction: Y	27	39

Table 28 – Predicted TN, FN, TP and FP with calibrated parameters for the 'week1' dataset

Finally, the DT model for the 'week2' dataset, grown by using the default 'rpart' parameters, is shown in **Figure 19**. The DT has 11 leaves and 5 levels, having classified correctly 173 cases, and failing to classify 69 observations.

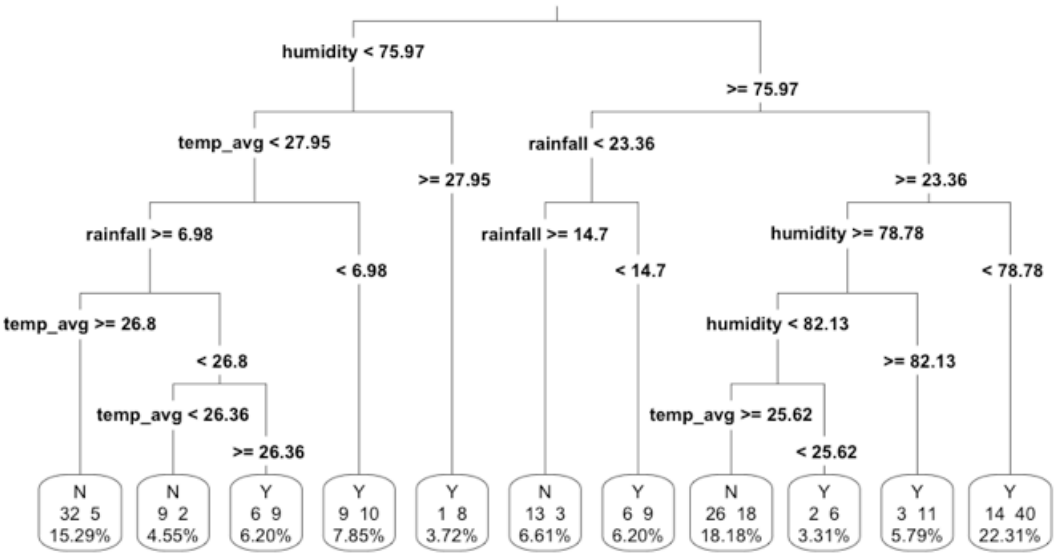


Figure 19 - DT grown with the default 'rpart' parameters for the 'week2' dataset

The above DT performed slightly better than the previous ones. When tested against the testing set, it classified 61 cases as occurring in a week where an outbreak took place, and 60 as not occurring within an outbreak situation. It got more than 58% of accurate predictions, from which the TN, FN, TP and FP can be consulted in the table below.

True value	*N*	*Y*
Prediction: N	34	26
Prediction: Y	24	37

Table 29 – Predicted TN, FN, TP and FP with default parameters for the 'week2' dataset

Few adjustments were made in the DT above, in order to improve that tree. Similarly with what happened while tuning the model for the 'week1' dataset, only the CP threshold had to be changed, and the value of 0.024793 was used. The DT grown with the new parameter change can be consulted below.

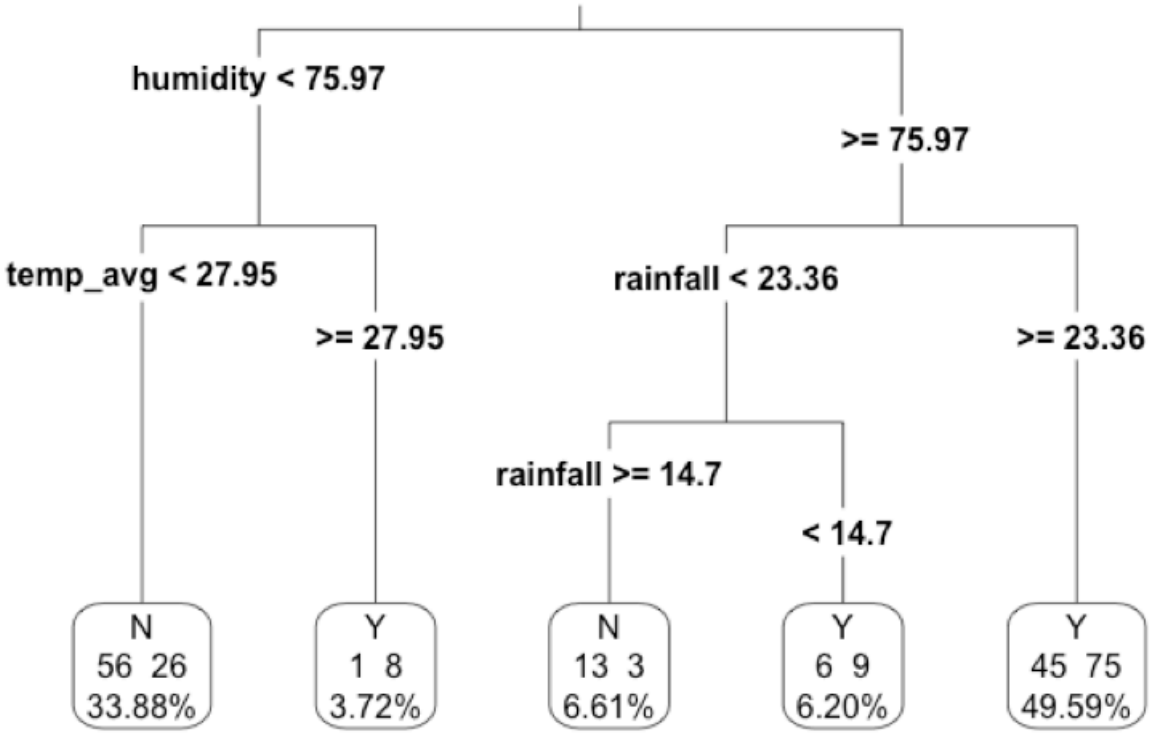


Figure 20 - DT grown after calibrating for the 'week2' dataset

The newly formed tree has 5 leaves and only 3 levels, was able to classify correctly 161 cases and failed to classify 81. When predicting against the test set, the model chose an outcome of 47 situations not having occurred in an outbreak scenario, while 74 did occur in such circumstances, making a good guess in more than 59% of the times. The TN, FN, TP and FP of those predictions are shown in **Table 30**.

True value	*N*	*Y*
Prediction: N	28	19

Prediction: Y	30	44

Table 30 – Predicted TN, FN, TP and FP with tuned parameters for the 'week2' dataset

Clustering

Before applying a clustering algorithm, the data needs to be prepared. Since clustering is an unsupervised learning technique, the target attribute will not be considered. Clustering algorithms can only be applied to data with numeric attributes, and these attributes need also to be rescaled, in order for all of them to have similar significance. The R code displayed below was used to undertake all these transformations on the demographic dataset.

```
# Create a copy of the dataset to be used with Clustering
dat_demographic_cl = dat_demographic

# Remove the attribute that is to be predicted
dat_demographic_cl$outbreak <- NULL

# Convert the non-numeric attributes in numeric attributes
dat_demographic_cl$age <- as.numeric(dat_demographic_cl$age)
dat_demographic_cl$gender <- as.numeric(dat_demographic_cl$gender)
dat_demographic_cl$race <- as.numeric(dat_demographic_cl$race)
dat_demographic_cl$job <- as.numeric(dat_demographic_cl$job)

# Rescale variables to assure compatibility
dat_demographic_cl = scale(dat_demographic_cl)
```

The next step consists of choosing the most appropriate algorithm to derive the model. For that, the "clValid" package has been used, because it contains functions for validating the results of a clustering analysis, and compares the performance of several algorithms for different numbers of clusters (Brock, et al., 2008). Notwithstanding, this last feature will not be used, since it is known beforehand that only two clusters exist. Also, only two algorithms will be considered, namely K-means and PAM. The R code necessary to run those algorithms and for inspecting the results afterwards is as follows.

```
# Compute internal validation measures about the clustering models obtained with K-means and
PAM algorithms
intern <- clValid(dat_demographic_cl, 2, clMethods=c("kmeans","pam"), validation="internal")

# View the computed results
summary(intern)
```

The table below presents the internal validation measures calculated for the generated clusters.

	Connectivity	*Dunn*	*Silhouette*
K-means	0.0000	0.0787	0.2518
PAM	12.3933	0.0724	0.2114

Table 31 – Internal validation measures of the clusters generated for the demographic dataset

These measures reflect the compactness, connectedness, and separation of the cluster partitions. Connectedness relates to what extent observations are placed in the same cluster as their nearest neighbours in the data space, and the "clValid" package measures it by the connectivity. Compactness assesses cluster homogeneity, usually by looking at the intra-cluster variance, while separation quantifies the degree of separation between clusters (usually by measuring the distance between cluster centroids). Since compactness and separation demonstrate opposing trends (compactness increases with the number of clusters but separation decreases), popular methods combine the two measures into a single score. The Dunn Index and Silhouette Width are both examples of non-linear combinations of the compactness and separation, and with the connectivity comprise the three internal measures available in "clValid".

Connectivity has a value between zero and ∞ and should be minimized. For a particular clustering partition $\mathrm{P} = \{C_1, \dots, C_K\}$ of the N observations into K clusters, the connectivity is defined as:

$$C(\mathrm{P}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

In the above equation, $nn_{i(j)}$ is defined as the jth nearest neighbour of observation i, and L is the number of nearest neighbours to use. Then $x_{i,nn_{i(j)}}$ is zero if i and j are in the same cluster and $1/j$ otherwise (Brock, et al., 2008).

The Silhouette Width is the average of each observation's Silhouette value. The Silhouette value measures the degree of confidence in the cluster attribution of a particular observation, with well-

clustered observations having values near 1 and poorly clustered observations having values near −1. Thus, it lies in the interval [−1, 1], and should be maximised.

The Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It has a value between zero and ∞, and should be maximised.

The measures presented in **Table 31**, referring to the demographic dataset, were also calculated for the remaining datasets, and are hereby shown, respectively, in **Table 32**, **Table 33** and **Table 34**.

	Connectivity	*Dunn*	*Silhouette*
K-means	37.7488	0.0278	0.3725
PAM	32.3329	0.0486	0.3712

Table 32 – Internal validation measures of the clusters generated for the 'week0' dataset

	Connectivity	*Dunn*	*Silhouette*
K-means	37.7488	0.0278	0.3722
PAM	32.3329	0.0487	0.3707

Table 33 – Internal validation measures of the clusters generated for the 'week1' dataset

	Connectivity	*Dunn*	*Silhouette*
K-means	35.3556	0.0482	0.3732
PAM	31.7802	0.0487	0.3723

Table 34 – Internal validation measures of the clusters generated for the 'week2' dataset

These results show that K-means is better suited to analyse the demographic dataset, whilst for the remaining datasets PAM is the choice to be made. The R output already points out this choice, but in order for the reader to better understand it, the table below shows the most suited algorithm for the datasets, according to the studied measures.

	Demographic	*Week0*	*Week1*	*Week2*
Connectivity	K-means	PAM	PAM	PAM
Dunn	K-means	PAM	PAM	PAM

Silhouette	K-means	K-means	K-means	K-means

Table 35 - Algorithm chosen for each dataset according to the studied measures

The connectivity value was lower with the K-means algorithm for the demographic dataset, while the opposite happened with the remaining datasets. Since this measure must be minimised, K-means performed better when applied to the demographic dataset, while PAM had better results when applied to the remaining ones. Similarly, PAM performed better for the 'week0', 'week1' and 'week2' datasets, according to the Dunn index, because it had higher values when compared to the ones obtained with K-means, and for the demographic dataset PAM was once again outperformed by K-means. The silhouette measure did not follow the trend of the other measures, and K-means performed better than PAM in all the datasets, having had bigger values for the silhouette value on every run.

The demographic dataset was clustered using K-means as the clustering algorithm, as was instructed by 'clValid'. K-means is the most common partitioning algorithm in cluster analysis. It aims to find the best division of n entities in k groups, so that the total distance between the group's members and its corresponding centroid, representative of the group, is minimised. Given the number of centroids K and a set of data points $x_1 \dots x_n$ (N is the size of the dataset), K-means proceeds as follows:

1. Select K centroids, randomly.
2. Assign each data point to its closest centroid.
3. Recalculate the centroids as the average of all data points in a cluster.
4. Assign data points to their closest centroids.
5. Continue steps 3 and 4 until the observations are not reassigned or the maximum number of iterations (R uses 10 as a default) is reached.

In steps 2 and 4, R uses an efficient algorithm (Hartigan and Wong, 1979) that partitions the observations into K groups, by minimising the sum of squares of the observations to their assigned cluster centres. Thus, each observation is assigned to the cluster with the smallest value of:

$$SS(k) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - \bar{x}_{kj}\right)^2$$

Where k is the cluster, x_{ij} is the value of the j^{th} variable for the i^{th} observation, and $\bar{x}_{kj}$ is the mean of the j^{th} variable for the k^{th} cluster. The number of observations is given by n, and p is the number of variables.

The code below illustrates the R code used, after calibrating the parameters. R does not provide many ways to tune the K-means algorithm, but through the 'nstart' parameter, the algorithm attempts multiple initial configurations and reports on the best one. For example, with $\text{nstart} = 25$, the algorithm will try 25 initial configurations. The number of clusters is also a K-means parameter, but since in this study case the number of clusters was known in advance, there was nothing to calibrate here. Different results were obtained by changing the seed though, which also allowed tuning the algorithm. The algorithm to be used in the K-means command is also a parameter (Hartigan-Wong, Lloyd, Forgy or MacQueen), alongside with the maximum number of iterations, but the results were not improved when these parameters were changed.

```
# Set the seed for the random number generators, as an assurance that the results will be
reproducible
set.seed(1)

# Apply the kmeans clustering algorithm
kmeans.result <- kmeans(dat_demographic_cl, centers=2, nstart=25)
```

To better understand how the clustering algorithm performed, the number of correct and incorrect observations in each cluster was calculated with the following R command. **Table 36** presents the output, showing that there were 3227 correct observations in 6076 observations.

```
table(dat_demographic$outbreak, kmeans.result$cluster)
```

	1	*2*
N	1395	949
Y	2278	1454

Table 36 – Observations per cluster after applying K-means to the demographic dataset

The following command allows the K-means output to be looked at in more detail. Its output revealed that, among others, 2 clusters of sizes 3673 and 2403 were found, and the ratio of the between sum of squares to the total sum of squares ($\frac{between\ SS}{total\ SS}$) equals 25.5%. The latter is a measure of the total variation in the dependent variable explained by the cluster means. Since this value is pretty low, it suggests a poor outcome by the K-means algorithm.

```
print(kmeans.result)
```

Having four dimensions being studied through clustering techniques makes it difficult to properly plot the data, therefore no graph will be shown hereby.

As suggested by the output of the 'clValid' package, the 'week0', 'week1' and 'week2' datasets were clustered using PAM as the clustering algorithm, which is a modern alternative to K-means clustering. PAM is an acronym for "Partitioning around Medoids", and the term medoid refers to an observation within a cluster for which the sum of the distances between it and all the other members of the cluster is a minimum. These observations (one per cluster) are a representative example of the members of that cluster. Like K-means, PAM requires that you know the number of clusters in advance, but in order to insure that the medoids it finds are truly representative of the observations within a given cluster, it involves considerably more computation than K-means. While in the K-means algorithm the centres of the clusters (which might or might not actually correspond to a particular observation) are only recalculated after all of the observations have had a chance to move from one cluster to another, with PAM the sums of the distances between objects within a cluster are constantly recalculated as observations move around. It is expected that this procedure will provide a more reliable solution, but as the results obtained with the "clValid" show, that does not happen for every study case (Maechler, et al., 2015).

In R, not much calibration could be done while using PAM. But although the attempts to tune the model for 'week0' dataset produced no changes in the results, it was possible to obtain better results in the models related to 'week1' and 'week2' datasets. That was accomplished by setting the metric to be used for calculating dissimilarities between observations through the "metric" parameter to "manhattan", therefore using the sum of absolute differences instead of Euclidean distances.

The command used in the 'week0' dataset, with no parameters other than the mandatory ones, can be found below. **Table 37** shows that there have been 191 correct observations in a total of 365.

```
# Apply the pam clustering algorithm
pam.result <- pam(week0_cl, k=2)
```

	1	*2*
N	87	92
Y	99	87

Table 37 - Observations per cluster after applying PAM to the 'week0' dataset

The PAM command in R provides a plot known as a silhouette plot. This feature calculates a measure for each observation, that allows seeing how well it fits into the cluster that it has been assigned to. It compares how close the object is to other objects in its own cluster with how close it is to objects in other clusters. A value near 1 indicates that the observation is well placed in its cluster, while values near 0 mean that it is likely that an observation might belong in some other cluster. If the silhouette plot shows values close to 0 for each observation, the fit was not good; but if there are many observations closer to 1, it can be concluded that the fit was good. The silhouette plot is thus very useful in assessing about the results obtained by the PAM command, and it will be used to determine the goodness of the models obtained for the 'week0', 'week1', and 'week2' datasets. The following command plots, among others, the silhouette plot, while **Figure 21** presents the plot outputted by the command.

```
plot(pam.result)
```

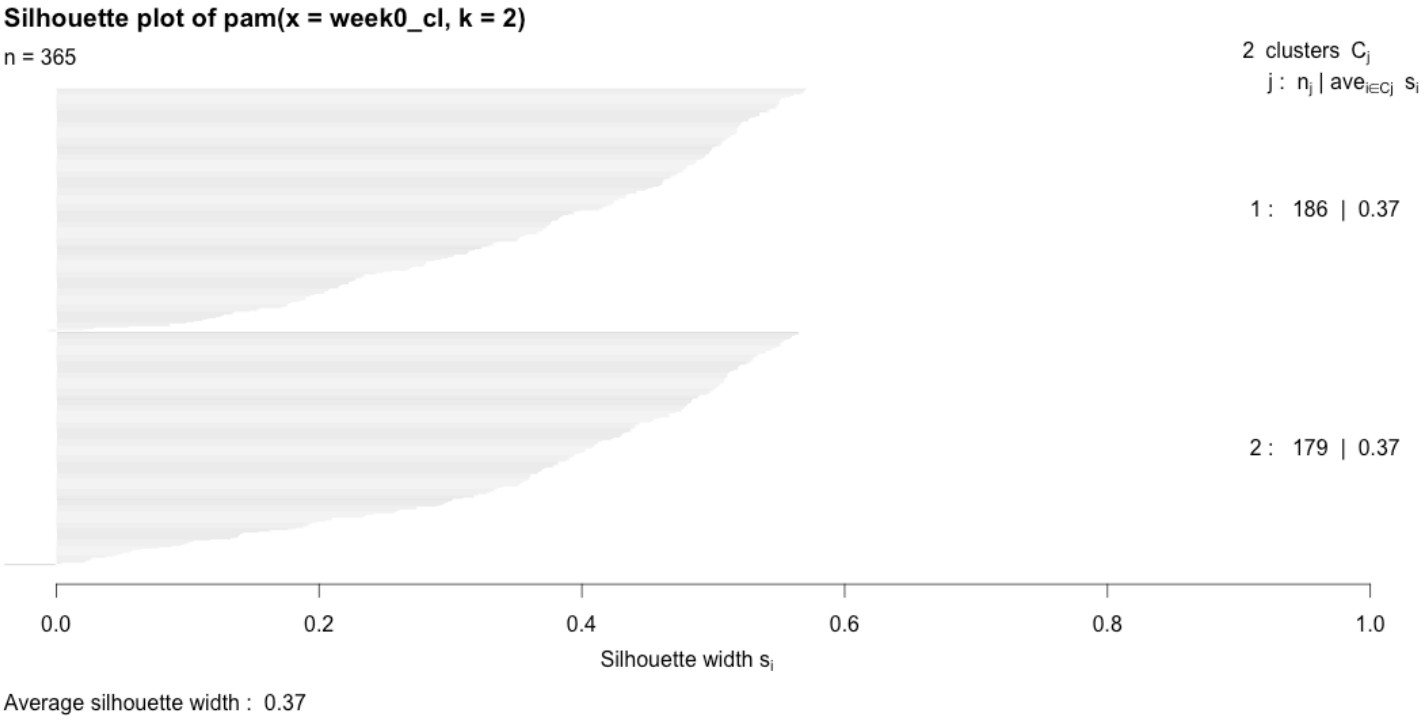


Figure 21 - Silhouette plot related to the 'week0' PAM model

The following tables (**Table 38** and **Table 39**) and figures (**Figure 22** and **Figure 23**) present the results obtained for the 'week1' and 'week2' datasets.

	1	*2*
N	91	88
Y	87	98

Table 38 - Observations per cluster after applying PAM to the 'week1' dataset

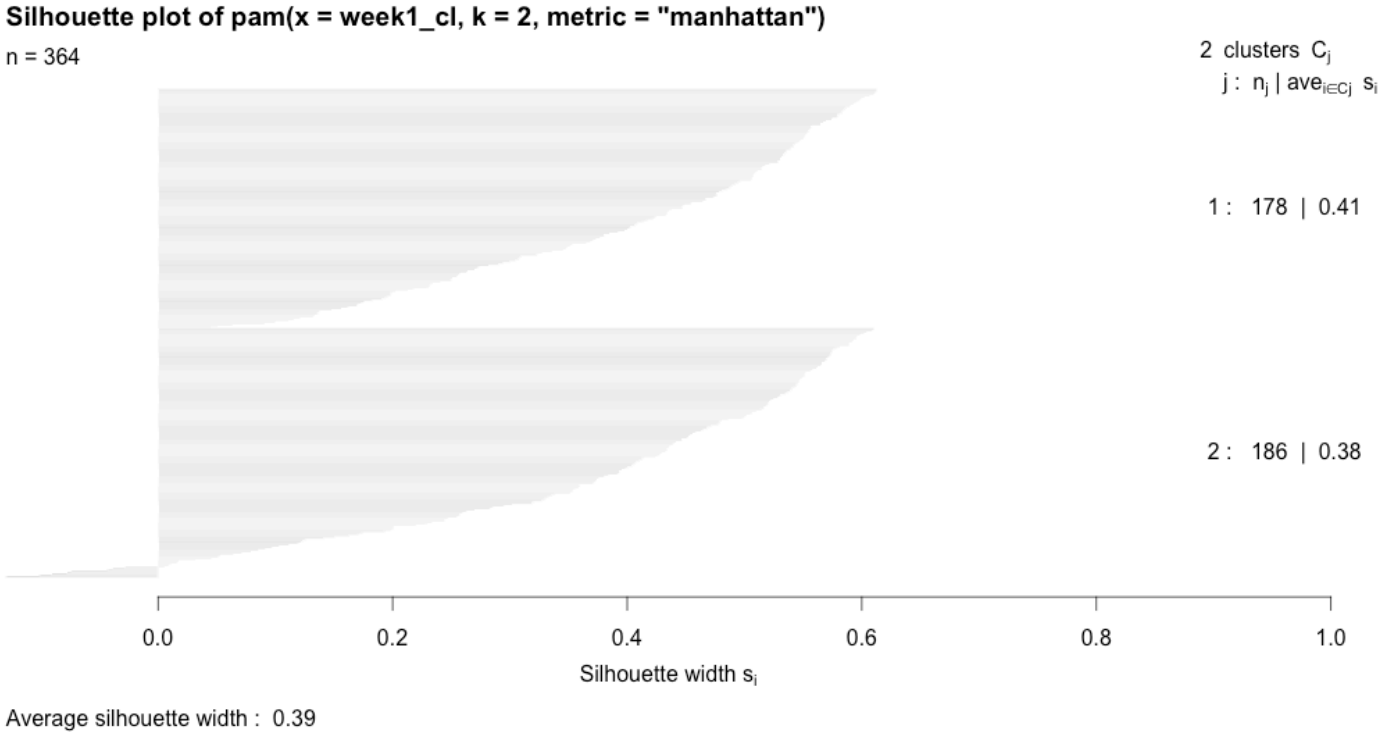


Figure 22 - Silhouette plot related to the 'week1' dataset

	1	*2*
N	79	100
Y	99	85

Table 39 - Observations per cluster after applying PAM to the 'week2' dataset

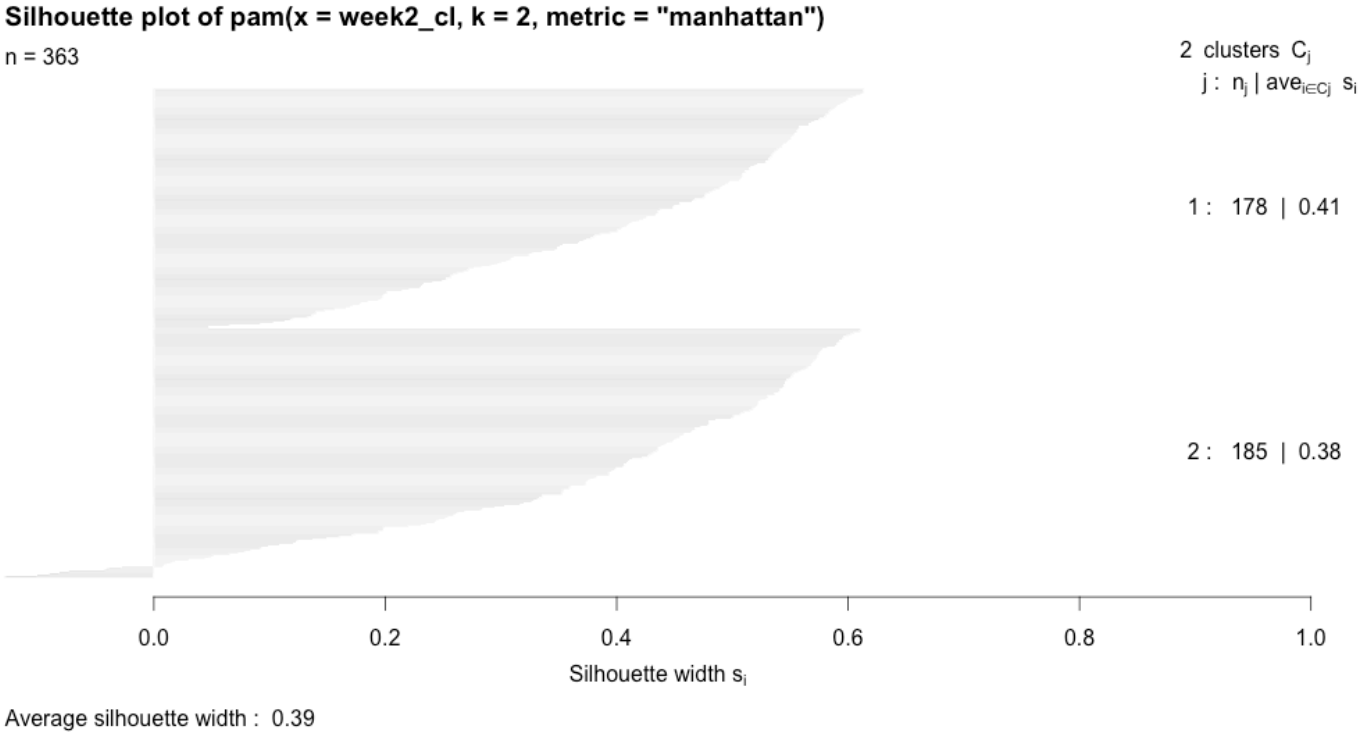


Figure 23 - Silhouette plot related to 'week2' dataset

Similarly to what happened with the K-means model for the demographic dataset, the results were not good at all. The amount of misclassified observations is high in the three models, and having average values given by the silhouette plot in the generated models of 0.37, 0.39 and 0.39 for the 'week0', 'week1', 'week2', respectively, indicates that the clusters are far from being accurate.

Chapter 5

5 Conclusions and future work

In a first iteration of the methodology, and according to the available information (that had to be translated from Malay language), the target attribute (also known as category or class label) in the application of the classification algorithms was the type of epidemic. It was believed that this attribute was carrying information about, as the name suggests, the type of outbreak, and that it would allow to properly classify an outbreak as being a new outbreak, a controlled outbreak or an uncontrolled outbreak. Nonetheless, by analysing the data after creating graphs and calculating several statistics about the data, it has been concluded that that attribute brings no added value to the research. Occurrences of the three cases are common during the same week and within the same town, which indicates that the initial information about the attribute was erroneous. Therefore, the decision to not consider that attribute has been made, and since there was no other discretised attribute that could be used as target attribute, the number of occurrences of DF and DHF have been discretised, and used as target attributes in a second iteration of the methodology. As a consequence, the initial DM goals and the initial hypotheses had to be updated in this new iteration of the methodology.

Discretisation, which is commonly used in classification algorithms but not so much in regression or clustering algorithms, consists in replacing an originally continuous attribute by a discrete attribute, with different values assigned to particular intervals of the original attribute's range. The target attribute can be used during discretisation to observe the predictive utility of the attribute being discretised and its loss due to discretisation. When the discretisation algorithm exploits this

possibility, it is referred to as supervised, and if it does not, it is called unsupervised. Since the attributes being discretised are the target attributes, the chosen algorithm was unsupervised. That algorithm is called equal-frequency intervals, and in appendix b the R code used in its implementation is presented (Cichosz, 2015). It is not included in the repository because it is just a small function.

The problem with the equal-frequency intervals algorithm is that it requires the number of intervals in which the attribute should be discretised as input. Having two target attributes to study and still having to infer the best possible combination of intervals at this point, left many possibilities open and since the initial results were not convincing and especially because there was no medical background supporting the decision to discretise the chosen attributes, another approach was taken. In (Tarmizi, et al., 2013a; Tarmizi, et al., 2013b) which, as mentioned before, use the same or a similar dataset as the one used in this study, the authors have decided to derive a new attribute, based on the definition of outbreak, and the same was done in this research. This step was explained in the Data preparation section. Once again the target attribute has changed and, consequently, the DM goals and the initial hypotheses had to be changed accordingly.

The available literature, as well as the other tools already developed, are a good basis for the creation of an application that will support the generation of ARs, and therefore will help to achieve the objectives of this dissertation. QuantMiner (Salleb-Aouissi, Vrain and Nortet, 2007) can be used to generate the ARs the author is looking for, and several modifications and improvements were made in the application so that it becomes as useful as possible. Notwithstanding, the complexity shown by QuantMiner, associated with not so promising results in the beginning of the research, has led to postponing the first two business objectives, namely the development of AR models for the development of DF predictive models and devising GAs to assist the generation of these ARs. Following the initial guidelines, coming from the research group whose research led to this dissertation work, made it very difficult to follow the chosen methodology, and that was also a factor for postponing the business objectives number 1 and 2. The DM goals were already outlined considering these decisions. In order to comply with the self-imposed deadline to deliver this dissertation, it was not possible to get back to those objectives in due time.

The work done over the same or a similar dataset (Bakar, et al., 2011; Long, et al., 2010; Mousavi, et al., 2013; Tarmizi, et al., 2013a; Tarmizi, et al., 2013b) was focused on different attributes,

therefore they could not be used as a basis for this work. That alone is already a major factor that makes starting the dissertation by deriving AR models a very difficult task, without having neither DT-based nor clustering models. Moreover, shifting the focus of the research to socio-demographic factors rather than meteorological drivers distanced even more this dissertation work from the work done before over the same dataset.

The models that have been derived (classification and clustering) are not good enough to be deployed. There are several reasons that can explain this. First of all, based on the Literature review, more specifically in (Flamand, et al., 2014; Pathirana, Kawabata and Goonatilake, 2009), it was necessary not only to consider the previous two weeks to assess about the likelihood of a DF outbreak, but several more. As previously mentioned, the authors of these works found a correlation between dengue outbreaks and meteorological factors that occurred from 3 to 6 weeks before the outbreak actually took place.

The lack of good results is also a consequence of not having done enough iterations of the methodology, as well as not having used enough different algorithms in each model. Having assessed the models and therefore concluding that they did not comply with the requirements, it is necessary to go one step backwards, in order to improve the results, and that has not been done. The methodology foresees even the redefinition of the business requirements if the assessment is not as expected.

It is important to note, however, that Malaysia is an endemic country for DF, that is very difficult to extract good knowledge from the socio-demographic data, and also that the definition of outbreak that has been used is very broad. The combination of these factors can also explain the results, and highlights the need to further continue this research in future works.

Bibliography

[1] Agrawal, R. and Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pp. 487-499. Morgan Kaufmann Publishers Inc.

[2] Alataş, B. and Akin, E. 2006. An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules. *Soft Computing*, vol. 10, no. 3, pp. 230-237.

[3] Alataş, B., Akin, E. and Karci, A. 2008. MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. *Applied Soft Computing*, vol. 8, no. 1, pp. 646-656.

[4] Azevedo, A. and Santos, M.F. 2008, 'KDD, SEMMA and CRISP-DM: a parallel overview', in *IADIS European Conference on Data Mining*, ed. Abraham, A., Amsterdam, Netherlands, pp. 182-185.

[5] Bakar, A.A., Kefli, Z., Abdullah, S. and Sahani, M. 2011, 'Predictive models for dengue outbreak using multiple rulebase classifiers', in, IEEE, pp. 1-6.

[6] Balmaseda, A., Hammond, S.N., Pérez, L., Tellez, Y., Saborío, S.I., Mercado, J.C., Cuadra, R., Rocha, J., Pérez, M.A., Silva, S., Rocha, C. and Harris, E. 2006. Serotype-specific differences in clinical manifestations of dengue. *The American journal of tropical medicine and hygiene*, vol. 74, no. 3, pp. 449-456.

[7] Bee, T.K., Lye, K.H. and Yean, T.S., 2009. Modelling Dengue Fever Subject to Temperature Change. In: *Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 61-65. Tianjin, China, 14-16 August 2009 DOI: 10.1109/FSKD.2009.761.

[8] Bertolucci, J., 2014. Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive, InformationWeek. http://www.informationweek.com.

[9] Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O., Myers, M.F., George, D.B., Jaenisch, T., Wint, G.R.W., Simmons, C.P., Scott, T.W., Farrar, J.J. and Hay, S.I. 2013. The global distribution and burden of dengue. *Nature*, vol. 496, pp. 504-507.

[10] Bocchini, B. 2015, Butantan inicia última etapa de desenvolvimento da vacina da dengue. Agência Brasil (2015). Published electronically in December 12th. http://agenciabrasil.ebc.com.br/geral/noticia/2015-12/butantan-inicia-ultima-etapa-de-desenvolvimento-da-vacina-da-dengue.

[11] Brock, G., Pihur, V., Datta, S. and Datta, S. 2008. {clValid}: An {R} Package for Cluster Validation. *Journal of Statistical Software*, vol. 25, no. 4, pp. 1-22.

[12] Campbell, K.M., Lin, C.-D., Iamsirithaworn, S. and Scott, T.W. 2013. The complex relationship between weather and dengue virus transmission in Thailand. *The American journal of tropical medicine and hygiene*, vol. 89, no. 6, pp. 1066-1080.

[13] Caribbean Epidemiology Center, Pan American Health Organization and World Health Organization, 2001. Clinical and laboratory guidelines for dengue fever and dengue haemorrhagic fever/dengue shock syndrome for health care providers.

[14] Castro, L.N.d. and Timmis, J., 2002, *Artificial immune systems - a new computational intelligence paradigm*, Springer.

[15] Chapman, P., 1999. *The CRISP-DM User Guide* in 4th CRISP-DM Special Interest Group Workshop, NCR Systems Engineering Copenhagen, Brussels, Belgium.

[16] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

[17] Cichosz, P. 2015, John Wiley & Sons, Ltd.

[18] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L. 2007, 'Data Mining: A Knowledge Discovery Approach', Springer, New York, pp. 9-24.

[19] clicRBS, 2015. *Anvisa autoriza uso de primeira vacina contra dengue no país* in Zero Hora, http://zh.clicrbs.com.br/rs/vida-e-estilo/.

[20] Crockett, D., Johnson, R. and Eliason, B., 2014. What is Data Mining in Healthcare?

[21] Dean, J., 2014, *Big Data, Data Mining and Machine Learning - Value creation for business leaders and practitioners*, 1st Edition, Wiley Publishing, Inc.

[22] Department of statistics Malaysia, 2010. Population and housing census of Malaysia - Population distribution and basic demographic characteristics.

[23] Derouich, M., Boutayeb, A. and Twizell, E.H. 2003. A model of dengue fever. *BioMedical Engineering OnLine*, vol. 2, no. 1.

[24] Dou, W., Hu, J., Hirasawa, K. and Wu, G., 2008. Quick Response Data Mining Model Using Genetic Algorithm. In: *Society of Instrument and Control Engineers (SICE) Annual Conference*, pp. 1214-1219. Tokyo. DOI: 10.1109/SICE.2008.4654843.

[25] Esteva, L. and Vargas, C. 1998. Analysis of a dengue disease transmission model. *Mathematical biosciences*, vol. 150, no. 2, pp. 131-151.

[26] Esteva, L. and Vargas, C. 1999. A model for dengue disease with variable human population. *Journal of mathematical biology*, vol. 38, no. 3, p. 21.

[27] Esteva, L. and Vargas, C. 2000. Influence of vertical and mechanical transmission on the dynamics of dengue disease. *Mathematical biosciences*, vol. 167, no. 1, pp. 51-64.

[28] Flamand, C., Fabregue, M., Bringay, S., Ardillon, V., Quénel, P., Desenclos, J.-C. and Teisseire, M. 2014. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. *Journal Of The American Medical Informatics Association*, vol. 21, no. e2, pp. 232-240.

[29] Focks, D.A., Daniels, E.A., Haile, D.G.A. and Keesling, J.E.A. 1995. A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. *American journal of tropical medicine and hygiene*, vol. 53, no. 5, pp. 489-506.

[30] Focks, D.A., Haile, D.G., Daniels, E. and Mount, G.A. 1993. Dynamic life table model for Aedes aegypti (Diptera: Culicidae): analysis of the literature and model development. *Journal of medical entomology*, vol. 30, no. 6, pp. 1003-1017.

[31] Ghazali, A.N.M., Hod, R., Sahani, M., Ali, Z.M., Othman, H.F., Amin, F.M., Artika, I.N. and Er, A.C. 2012. Dengue infections and circulating serotypes in Negeri Sembilan, Malaysia. *Malaysian journal of public health medicine*, vol. 12, no. 1, pp. 21-30.

[32] Globo, 2015. *Primeira vacina contra a dengue é aprovada no Brasil* in Bom dia Brasil.

[33] Goldberg, D.E., 1989, *Genetic Algorithms in search, optimization and machine learning*, Addison-Wesley Longman Publishing Co., Inc.

[34] Gubler, D.J. 1998. Dengue and Dengue hemorrhagic Fever. *Clinical Microbiology Reviews*, vol. 11, no. 3, pp. 480-496.

[35] Hartigan, J.A. and Wong, M.A. 1979. Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108.

[36] Hartley, L.M., Donnelly, C.A. and Garnett, G.P. 2002. The seasonal pattern of dengue in endemic areas: mathematical models of mechanisms. *Transactions of the royal society of tropical medicine and hygiene*, vol. 96, no. 4.

[37] Holland, J.H., 1975, *Adaptation in Natural and Artificial Systems*, University of Michigan Press.

[38] IBM Corporation, 2011. *IBM SPSS Modeler CRISP-DM Guide*.

[39] Indira, K. and Kanmani, S. 2012. Performance Analysis of Genetic Algorithm for Mining Association Rules. *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 2.

[40] Kalayanarooj, S. and Nimmannitya, S., 2000. Clinical and laboratory presentations of dengue patients with different serotypes, *WHODengue Bulletin*, vol. 24, pp. 53-59.

[41] Kumar, R.M. and Iyakutti, K. 2011. Application of Genetic algorithms for the prioritisation of Association Rules. *IJCA Special Issue on Artificial Intelligence Techniques - Novel Approaches and Practical Applications*.

[42] Long, Z.A., 2014. Malaysia dengue detection model using frequent outlier. In: *Recent advances in electrical engineering and educational technologies*, pp. 93-101. Athens, Greece, November 28-30.

[43] Long, Z.A., Bakar, A.A., Hamdan, A.R. and Sahani, M. 2010, 'Multiple attribute frequent mining-based for dengue outbreak', in *Advanced Data Mining and Applications (ADMA)*, Springer, Chongqing, China, pp. 489-496.

[44] M., A.N., 2010. *Tips for prevention of breeding of Aedes mosquito in your urban neighbourhood* in A malayali doctor's blog, Kerala, India, pp. A blog by a secondary care Internal Medicine Specialist from India about Health, Medicine, Patients, Hospitals, etc.

[45] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K., 2015. Cluster: Cluster Analysis Basics and Extensions.

[46] Marbán, Ó., Mariscal, G. and Segovia, J. 2009. A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications*, pp. 1-17.

[47] Maron, D.F. 2015, First dengue fever vaccine gets green light in three Countries. Scientific American (2015). Published electronically in December 30th. http://www.scientificamerican.com/article/first-dengue-fever-vaccine-gets-green-light-in-three-countries.

[48] Martín, D., Rosete, A., Alcalá-Fdez, J. and Herrera, F. 2014. QAR-CIP-NSGA-II: A new multi-objective evolutionary algorithm to mine quantitative association rules. *Information Science*, vol. 258, pp. 1-28.

[49] Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A. and Riquelme, J.C., 2009. Quantitative association rules applied to climatological time series forecasting. In: *Intelligent Data Engineering and Automated Learning (IDEAL)*, pp. 284-291. Burgos, Spain. Springer Berlin Heidelberg.

[50] Ministry of health Malaysia, 2002. Annual report. Vector-borne disease control section.

[51] Ministry of health Malaysia, 2003. Annual report. Vector-borne disease control section.

[52] Ministry of health Malaysia, 2004. Annual report. Vector-borne disease control section.

[53] Ministry of health Malaysia, 2005. Annual report. Vector-borne disease control section.

[54] Ministry of health Malaysia, 2006. Annual report. Vector-borne disease control section.

[55] Ministry of health Malaysia, 2007. Annual report. Vector-borne disease control section.

[56] Ministry of health Malaysia, 2008. Annual report. Vector-borne disease control section.

[57] Ministry of health Malaysia, 2009. Annual report. Vector-borne disease control section.

[58] Ministry of health Malaysia, 2010. Annual report. Vector-borne disease control section.

[59] Ministry of health Malaysia, 2011. Annual report. Vector-borne disease control section.

[60] Ministry of health Malaysia, 2012. Annual report. Vector-borne disease control section.

[61] Mousavi, M., Bakar, A.A., Zainudin, S., Long, Z.A., Sahani, M. and Vakilian, M. 2013. Negative selection algorithm for dengue outbreak detection. *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, no. 2, pp. 2345-2356.

[62] Newton, E.A.C. and Reiter, P. 1992. A model of the transmission of dengue fever with an evaluation of the impact of ultra-low volume insecticide applications on dengue epidemics. *The American Journal of Tropical Medicine and Hygiene*, vol. 47, no. 6, pp. 709-720.

[63] Normile, D., 2013. Surprising new dengue virus throws a spanner in disease control efforts, *Science*, 25 October 2013, vol. 342, p. 415.

[64] Ooi, K.G., 2010, *The A to Z of Malaysia*, Rowman & Littlefield.

[65] Oracle, 2008. Oracle data mining.

[66] Otero, M., Schweigmann, N. and Solari, H.G. 2008. A stochastic spatial dynamical model for Aedes aegypti. *Bulletin of mathematical biology*, vol. 70, no. 5, pp. 1297-1325.

[67] Otero, M. and Solari, H.G. 2010. Stochastic eco-epidemiological model of dengue disease transmission by Aedes aegypti mosquito. *Mathematical biosciences*, vol. 223, no. 1, pp. 32-46.

[68] Otero, M., Solari, H.G. and Schweigmann, N. 2006. A stochastic population dynamics model for Aedes aegypti: formulation and application to a city with temperate climate. *Bulletin of mathematical biology*, vol. 68, no. 8, pp. 1945-1974.

[69] Papè, N.F., Alcalá-Fdez, J., Bonarini, A. and Herrera, F. 2009, 'Evolutionary extraction of association rules: a preliminary study on their effectiveness', in *Hybrid Artificial Intelligence Systems (HAIS) 2009*, Springer Berlin Heidelberg, Salamanca, Spain, pp. 646-653.

[70] Pathirana, S., Kawabata, M. and Goonatilake, R. 2009. Study of potential risk of dengue disease outbreak in Sri Lanka using GIS and statistical modelling. *Journal of Rural and Tropical Public Health*, vol. 8, pp. 7-17.

[71] Pongsumpun, P. and Tang, I.M. 2003. Transmission of dengue hemorrhagic fever in an age structured population. *Mathematical and Computer Modelling*, vol. 37, no. 9, pp. 949-961.

[72] Pressman, R.S.X., 2005, *Software engineering: a practitioner's approach*, Palgrave Macmillan.

[73] Racloz, V., Ramsey, R., Tong, S. and Hu, W. 2012. Surveillance of Dengue Fever Virus: A Review of Epidemiological Models and Early Warning Systems. *PLoS neglected tropical diseases*, vol. 6, no. 5, p. e1648.

[74] Reiter, P. 2001. Climate Change and Mosquito-Borne Disease. *Environmental Health Perspectives*, vol. 109, pp. 141-161.

[75] Rosen, L., 1982. Dengue - an overview. In: Mackenzie, J.S. (ed.), *Viral Diseases in Southeast Asia and the Western Pacific*, pp. 484-493. Canberra, Australia, 8-12 February 1982. Sydney, Australia: Academic Press. DOI: 0124848206, 978-0124848207.

[76] Rupnik, R. and Jaklič, J. 2009, 'Data mining and knowledge discovery in real life applications', eds Ponce, J. and Karahoca, A., InTech, pp. 373-388.

[77] Salleb-Aouissi, A., Vrain, C. and Nortet, C., 2007. QuantMiner: a genetic algorithm for mining quantitative association rules. In: *Proceedings of the 20th International Joint Conference on Artificial intelligence (IJCai)*, pp. 1035-1040. Hyderabad, India. Morgan Kaufmann Publishers Inc.

[78] Santos, L.B.L., Costa, M.C., Pinho, S.T.R., Andrade, R.F.S., Barreto, F.R., Teixeira, M.G. and Barreto, M.L. 2009. Periodic forcing in a three-level cellular automata model for a vector-transmitted disease. *Physical Review E*, vol. 80, no. 1, p. 016102.

[79] Simovici, D.A. Data Mining of Medical Data: Opportunities and Challenges in Mining Association Rules.

[80] Stephens, T., 2014. *Getting started with R - part 3: decision trees* in trevorstephens.com, San Francisco, California, USA.

[81] Tan, P.-N., Steinbach, M. and Kumar, V. 2005, Addison-Wesley Longman Publishing Co., Inc., pp. 327-414.

[82] Tarmizi, N.D.A., Jamaluddin, F., Bakar, A.A., Othman, Z.A. and Hamdan, A.R. 2013a. Classification of dengue outbreak using Data Mining models. *Research Notes in Information and Service Sciences (RNIS)*, vol. 12, pp. 71-75.

[83] Tarmizi, N.D.A., Jamaluddin, F., Bakar, A.A., Othman, Z.A., Zainudin, S. and Hamdan, A.R. 2013b. Malaysia Dengue Outbreak Detection Using Data Mining Models. *Journal of Next Generation Information Technology (JNIT)*, vol. 4, no. 6, pp. 96-107.

[84] Therneau, T.M. and Atkinson, E.J., 2015. An introduction to recursive partitioning using the RPART routines. Mayo Foundation.

[85] Tran, A. and Raffy, M. 2006. On the dynamics of dengue epidemics from large-scale information. *Theoretical Population Biology*, vol. 69, no. 1, pp. 3-12.

[86] Vannucci, M. and Colla, V., 2004. Meaningful discretization of continuous features for association rules mining by means of a SOM. In: *12th European Symposium on Artificial Neural Networks (ESANN) 2004*, pp. 489–494. Bruges, Belgium, 28–30 April 2004.

[87] Wearing, H.J. and Rohani, P.G., 2006. *Ecological and immunological determinants of dengue epidemics*, vol. 103, National Academy of Sciences, Washington, DC, USA, p. 6.

[88] World Health Organization, 1991. The urban crisis, World Health Statistics Quarterly vol 44.

[89] World Health Organization, 2009, *Dengue : guidelines for diagnosis, treatment, prevention and control*, New ed. 2009 Edition, World Health Organization Geneva.

[90] Xu, Y., Zeng, M., Liu, Q. and Wang, X. 2014. A Genetic Algorithm based multilevel association rules mining for big datasets. *Mathematical Problems in Engineering*, vol. 2014, p. 9.

[91] Yan, X., Zhang, A.C. and Zhang, A.S. 2009. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert systems with applications*, vol. 36, no. 2, pp. 3066-3076.

[92] Yusof, Y. and Mustaffa, Z. 2011. Dengue outbreak prediction: A least squares support vector machines approach. *International Journal of Computer Theory and Engineering*, vol. 3, no. 4.

[93] Zhao, Q. and Bhowmick, S.S. 2003. Association rule mining: a survey. *Nanyang Technological University, Singapore*.

Appendices

a. Original demographic dataset

1	YEAR	21	OCCUPATION	41	RN	61	PCV2	81	VOMIT	101	SECCODE	121	D_ONSET
2	MONTH	22	SCHOOL	42	WARD	62	BT	82	RASHES	102	SINCE1_D	122	HOSP_RAW
3	WEEK	23	ADDRESS1A	43	TIMEKNW	63	CT	83	COMATOS	103	POSTOAP	123	RUM_LAW
4	NAME	24	ADDRESS1B	44	ONSET	64	PT	84	MANIC	104	B_MEMBER	124	PEND_LAW
5	CASE	25	ADDRESS1C	45	IN	65	TSERO	85	BLEEDING-GUM	105	T_SURVEY	125	C_FOG
6	CASENO	26	ADDRESS1D	46	DIAGNOSE	66	RESULT	86	BLEEDING-HIB	106	POSCONT	126	MASAFOG
7	CASENODIS	27	ADDRESS2A	47	INFORM	67	RESULT_IGG	87	GANGGUAN	107	AI_INDEX	127	GROYONG
8	WEEKONSET	28	ADDRESS2B	48	OUT	68	SERUM	88	ECCHY	108	BI_INDEX	128	NOTE
9	DATE	29	ADDRESS2C	49	DEAD	69	FIBRI	89	PURPU	109	D_FOG	129	NTEAM
10	ICNO	30	ADDRESS2D	50	DEADCODE	70	COMPLE	90	HAEMETE	110	TEL_CASE	130	TINVEST
11	DISTRICT	31	ADDRESS3B	51	EPIC TYPE	71	TRANSA	91	MALAEMA	111	POSCODE	131	ETHNIC
12	CLASSIFIC	32	ADDRESS3C	52	EPICCASE	72	MOVEMENT	92	SHOCK	112	LONGITUDE	132	TSERO2
13	PLACE	33	CITY	53	TEMP	73	SENDER	93	HAEMATU	113	LATITUDE	133	RESULT2
14	AGE	34	CITY2	54	BP	74	POS1	94	SEROTYPE	114	FOREIGN	134	PDS
15	A_BABY	35	CITY3	55	BP2	75	TELFAX	95	AKTIOLEH	115	TEL_TRANS	135	
16	SEX	36	PBT	56	HESSTEST	76	FEVER	96	SPCOPS	116	D_SINCE2	136	
17	RACE	37	AREA	57	HT2	77	MUSCLE	97	ACCEPT_D	117	D_UNTIL	137	
18	CITIZEN	38	LOCALITY	58	PC	78	HEAD	98	ACCEPTTIME	118	TEL_OFF	138	
19	STATE	39	HOSPITAL	59	PC2	79	JOINT	99	INFORM2	119	D_SINCE3	139	
20	MAJDAE	40	HOSPNAME	60	PCV	80	PETECHIE	100	ENVIRONMENT	120	RECURRENT	140	

Figure 24 – List of attributes from the original demographic dataset – taken from (Bakar, et al., 2011)

b. Equal-frequency algorithm implementation

The following function, written in R, implements the equal-frequency intervals algorithm by taking in consideration the distribution of the values from the attribute being discretised. It yields increased interval resolution wherever higher value density is observed by outputting interval breaks that aim to achieve the same number of training set instances corresponding to each bin, or since this cannot be usually achieved exactly, to minimise the differences among bin frequencies (Cichosz, 2015).

```
disc.eqfreq1 <- function(v, k)
{
        unique(quantile(v, seq(1/k, 1-1/k, 1/k)))
}
```