UPPSALA
UNIVERSITET

Prediction of Factors Influencing Rats Tuberculosis Detection Performance Using Data Mining Techniques

Joan Jonathan

Subject: Information Systems
Corresponds to 30 hp

Presented: Spring 2019
Main Supervisor: Prof. Andreas Hamfelt

Co-Supervisor: Dr. David Johnson
Examiner: Franck Tétard

Department of Informatics and Media

Abstract

This thesis aimed to predict the factors that influence rats TB detection performance using data mining techniques. A rats TB detection performance dataset was given from APOPO TB training and research center in Morogoro, Tanzania. After data preprocessing, the size of the dataset was 471,133 rats TB detection performance observations and a sample size of 4 female rats. However, in the analysis, only 200,000 data observations were used. Based on the CRISP-DM methodology, this thesis used R language as a data mining tool to analyze the given data. To build the predictive model the classification technique was used to predict the influencing factors and classify rats using a decision tree, random forest, and naive Bayes algorithms. The built predictive models were validated with the same test data to check their classification prediction accuracy and to find the best. The results pinpoint that the random forest is the best predictive model with an accuracy of 78.82%. However, the accuracy differences are negligible. When considering the predictive model accuracy (78.78%) and speed (3 seconds) of the decision tree, it is the best predictive model since it has less building time compared to the random forest (154 seconds). Moreover, the results manifest that age is the most significant influencing factor, and rats of ages between 3.1 to 6 years portrayed potentiality in detection performance. The other predicted factors are Session_Completion_Time, Session_Start_Time, and Av_Weight_Per_Year. These results are useful as a reference to rats TB trainers and researchers in rats TB and Information Systems. Further research using other data mining techniques and tools is valuable.

Keywords

Giant African Tuberculosis in human sputum, Pouched rats Tuberculosis in human sputum, Trained African giant pouched rats in human sputum, Data Mining in healthcare, Data Mining, Classification Technique in Diagnosis of Tuberculosis, Classification Technique.

Acknowledgements

I am sincerely thankful to the almighty heavenly God who strengthened me and consistently protected me together with my whole family during the entire time of my studies. Without a doubt, God is good all the time. Moreover, the success of this thesis is the result of many people support, assistance, and experience.

Heartfelt thanks to the Swedish Institute (SI) for recognizing the potential in me and awarded me the scholarship to pursue the master's studies. I appreciate this great honor and valuable opportunity of financial support that helped me to develop my career and acquire knowledge from a global perspective.

A warm appreciation to my employer the Sokoine University of Agriculture (SUA) and the Center for Information and Communication Technology (CICT) for their appropriate and useful recommendations and assistance concerning the release of my study leave. Moreover, I appreciate the contribution of the Director of CICT (Dr. Jacob A. Churi) for showing concern on the progress of my studies. I also express my deeply grateful to my main supervisor Professor Andreas Hamfelt, co-supervisor Dr. David Johnson, and Dr. Steve McKeever for their principal guidance, motivation, and assistance based on the formal supervisory meetings.

Many thanks to the Program Manager TB Tanzania (Dr. Georgies Mgode) and APOPO TB Training and Research center in Morogoro, Tanzania for granting me access to data for this thesis. I surely appreciate the meetings we had and all the guidance towards the success of the study. Not only that but also the time you used to read up my concept note, and your inputs were consistently helpful.

I am glad to give my special thanks to my husband Mbwiliza Mfumya, my daughter Ishva Mfumya for their love, patience, prayers throughout my studies. Besides, I would like to express sincere gratitude to Mary Jonathan (My mother), Mr. Goldian Mnyambo (My father), Mr. and Mrs. Edwin Kihumbe and Mr. and Mrs. Martine Kihumbe for their encouragements and prayers all the time of my studies. May you receive many blessings and favored protection from our almighty heavenly God.

Based on the hardship of mentioning all people who led to the successful accomplishment of this thesis, suffice it to say thank you all. I heartily appreciate your ideas, encouragements, comments, and advice all over my study.

Index

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
APOPO	Anti-Persoonsmijnen Ontminjnende Product Ontwikkeling
CRISP-DM	Cross Industry Standard Process for Data Mining
DTA	Data Transfer Agreement
DOTS	Directly Observed Treatment, Short-course
FP	False Positive
FN	False Negative
IS	Information System
MAE	Mean Absolute Error
MRCC	Medical Research Coordinating Committee
PBS	Phosphate-buffered saline
RMSE	Root Mean Square Error
SDT	Signal Detection Theory
SDGs	Sustainable Development Goals
SLR	Systematic Literature Review
SUA	Sokoine University of Agriculture
SVM	Support Vector Machine
TB	Tuberculosis
TP	True Positive
TN	True Negative
WHO	World Health Organization

1 Introduction

This chapter aimed to describe a background on the topic of a thesis. Following this, it presents a detailed explanation of a research problem with the research questions. Moreover, it delineates the motive and delimitation. Lastly, it concludes with a disposition which shows a structure of the thesis.

1.1 Background

Data reported annually to the World Health Organization (WHO) by countries shows that Tuberculosis often abbreviated as TB is one of the causes of death worldwide (WHO, 2018). TB is a life-threatening infectious disease caused by bacteria called Mycobacterium tuberculosis that attack the lungs and can also harm other parts of the body (Poling et al., 2011). The transmission occurs from one person to another when a person with TB talks, sneezes, or coughs (WHO, 2014). Millions of people continue to fall sick with TB globally each year, and the estimation was 10 million people in 2017 (WHO, 2018). However, people who are living with HIV in developing countries are more likely to be attacked by TB due to the lack of fast and correct tools that can effectively diagnose people with TB and who can start earlier treatment (Poling et al., 2011). According to WHO (2018), the report shows that 72% of TB transmissions and deaths occur in people with HIV in developing countries.

The 3rd Sustainable Development Goal (SDG) emphasizes good health and well-being for all at all ages. However, this goal has several targets the 3rd target aimed to end the TB by stopping TB deaths and a decrease in new cases by 2030 (WHO, 2018). To meet this target, WHO established a global plan called End TB Strategy to reduce TB deaths by 95%, TB incidence rate by 90%, and make sure that TB affected family sat free from the disaster costs (WHO, 2014). Moreover, WHO gives priority to the need for better adoption and acceptance of new tools to systematic diagnose TB earlier (WHO, 2014). According to Poling et al. (2011), sputum-smear microscopy technique known as microscopy is widely used in developing countries to diagnose TB. Since it has a high specificity of about 90% and a low sensitivity of about 20% to 80%, local clinics apply initiatives for improving sensitivity such as expanding the time for laboratory technicians to view slides. However, the challenge persists due to an absence or limited availability of sensitive diagnostic tools (Ellis et al., 2017).

During the upheaval, APOPO (Anti-Persoonsmijnen Ontmijnende Product Ontwikkeling) or "Anti- Personnel Landmines Detection Product Development" in English

and Sokoine University of Agriculture (SUA) are researching and exploring the trained African giant pouched rats nicknamed "HeroRATS" for detecting TB to complement microscopy and other available diagnostic tools (Mgode et al., 2018). APOPO is a Belgian Non-Government Organization (NGO) based in Morogoro, Tanzania aiming at using rat odor detection technology to solve the humanitarian problems (Poling et al., 2011). Rodents to which Cricetomys rats (HeroRATS) belong have the highly developed olfactory capacity, enabling them to undergo training and do specific detection tasks (Ellis et al., 2017). This technology uses operant conditioning techniques such as a reward or punishment to train rats in detecting Mycobacterium tuberculosis that causes TB disease (Mulder et al. (2017).

It has been in use in Tanzania since 2007 where rats sniff sputum sample from different local hospitals TB clinics under Directly Observed Treatment, Short-course (DOTS). Moreover, in Mozambique, its application is useful since 2013 (Mulder et al., 2017). In Tanzania from Morogoro and Dar es Salaam hospitals the detection rate for presumptive TB patients increased by 44% in 2009, by 43% in 2010, and by 39% in 2014. Moreover, from Maputo, Mozambique detection rate in presumptive TB patients have increased by 53% since 2013 (Poling et al., 2016). Additionally, since 2018 in Ethiopia this technology has been established for research use only. The detection performance of rats often measured in terms of sensitivity and specificity. Sensitivity implies how good the rat is at detecting samples with TB bacteria, while specificity reflects how good the rat is at detecting the samples without TB bacteria (Mahoney et al., 2013). Despite its importance, rats TB detection performance varies (Ellis et al.,2017).

APOPO center generates a large amount of data in its operations, including rats TB detection activities. And as such, it is valuable to discover interesting patterns and incompetence of the generated data for improving their detection performance and decision-making. Moreover, to reduce operating costs and develop husbandry care. Most of the studies (Nagabhushanam et al., 2013; Suresh et al., 2018) used classification as a data mining technique in the diagnosis of tuberculosis to categorize and find the relationships among the manipulated variables. Furthermore, the study of Asha et al. (2011) propose that this technique helps the health sectors to have better decision toward their operations. Therefore, this thesis aimed to predict the factors that influence rats TB detection performance using data mining techniques.

1.2. Problem overview

Since these rats are cost-effective and influence correct TB results, rats successful and consistent training is most important in TB healthcare centers that apply rat as odor-detection technology (Poling et al., 2011). The focal point of their training is on operant conditioning techniques which include reward or punishment. These techniques help to change behavior in such a way that rats improve TB detection performance. Hence, each rat must practice in every stage of training to qualify for detection sessions (Poling et al., 2016). During the clicker training, rats learn to recognize the sound while in discrimination training, they learn to either pause or scratch in a sputum sample (Poling et al., 2011). Since they do respond according to the instructions given during the training, it is useful to have the justifiable rules to avoid incorrect results. Therefore, well experimental setup, quality control, and precise data recording may enhance performance during the detection tasks (Reither et al., 2015).

Despite the justifiable rules, TB detection performance between trained rats does vary. The variation may occur when an individual rat within a group fails to show similarities in diagnostic accuracy such as specificity and sensitivity (Ellis et al., 2017). It often happens when a rat has either high sensitivity and low specificity or low sensitivity and high specificity (Mulder et al., 2017, Reither et al., 2015). Additionally, the variation may depend on patient age groups such as children, adolescents, and adults (Reither et al., 2015). And as such, rats may have detection performance on adults specifically in their most productive years (Mgode et al.,2018). Not only that but also levels of operant conditioning training may decide TB detection performance (Reither et al., 2015).

Moreover, the characteristics of rats may affect their TB detection performance. And as such, its impact may vary according to time. And as such, rats with certain factors may decide the effectiveness in odor-detection tasks (Mgode et al., 2018; Poling et al., 2011). Ellis et al., (2017) conducted a study which used 22 rats (male and female) and grouped them into two groups of 11 rats each with different ages aimed to find the relationship between age and sex with detection performance. Rats median ages were 3.8 years for Group 1 and 2.4 years for Group 2. Detection performance in Group 1 was better than in Group 2. In other words, older rats may do better than less older rats. Following this, age influenced rats TB detection performance but there were no significant differences between male and female detection performance.

However, it is not valuable to generalize this concept at large because detection performance may depend on different factors. Moreover, based on the sample size, it differs from this thesis since the data given and analysed had a sample size of 4 female rats and the median age of 3.71 years. On other hands, based on the experience as rats become older and the weight increases the detection performance decrease. In this conception, there is no empirical evidence on the main influencing factors and the trend of their impact is not clear. Therefore, understanding of factors such as age, weight, the session starts time of day, and session completion time of the trained rats is of importance to improve the rats TB detection performance. Thus, to further contribute to this body of knowledge this thesis focuses on the following research questions:

1.2.1 Research questions

- Can the factors that influence rats TB detection performance be predicted using a classification technique?
- If yes, to what extent do different factors affect rats TB detection performance?
- If age is one of the significant factors, which one provides peak performance and why?

1.3 Motivation

Using the African giant pouched rats as odor-detection technology is of importance not only as a complement for the other TB diagnostic tools but also in detecting land mines. As a result, this applicability makes the technology more significant and of interest. Not only that but also, the potential cost-effectiveness and speed in detecting many sputum samples than microscopy makes it of interest for TB new cases finding particularly in areas with high populations. Thus, it is a benefit to understand and consider rats significant factors that influence TB detection performance in all stages of their life.

The advantage of this thesis is to show empirically that rats' TB detection performance depends on specific factors. Thus, the center must take care of and support TB detection factors to increase rats' detection performance throughout their lifetime. Moreover, based on presented preprocessed data in plots, the center can visualize clear the correlation or trend of rats' TB data detection performance and utilize its usefulness. The result of this thesis is valuable as a reference to rats TB trainers and researchers in rats TB. Since this thesis implements data mining techniques in a social setting by identifying factors that enhance rats in detecting TB disease, it is also helpful to the academic society of Information System (IS).

1.4 Delimitation

The focal point of this thesis is rats' TB detection performance data from APOPO research and training center in Morogoro, Tanzania. These data are between 2014-2018 years and has a small sample size of 4 rats. However, the given datasets consisted of data from 2011 to 2019 and five female rats. The fifth rat in the RAT_WEIGHT dataset had no corresponding detection performance variables in the other two datasets and thus was disqualified. The sample size for characterizing a TB rat is therefore only four since was the ones found with the requested data and was expected to address the aim of the thesis. The range of five years may decide the desired results and answer the research questions.

Furthermore, there is no comparative analysis of rats' TB detection performance between the years. The data mining technique and algorithms used in this thesis to predict the main factors that influence rats TB detection performance and a class of every rat are classification, decision tree, random forest, and naïve Bayes respectively. The classification technique and these algorithms are proposed to yield desired results in TB diagnosis. However, it is not a design science study. Besides, other classification algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), and logistic regression can offer the solution.

On other hands, instead of using the weight of rats, this thesis applied Av_Weight_Per_Year to reduce bias since most of the rats' data in DetectionRatsData datasets missed their corresponding weights in the RAT_WEIGHT dataset. Despite several classification performance measures for comparing the built predictive models' performance, this thesis applied accuracy metric since it is used most in classification problems.

1.5 Disposition

This section depicts six chapters of the thesis aiming to present a detailed explanation and knowledge of the topic. These chapters include Introduction, Theory, Method, Results and Analysis, Discussion, and Conclusion.

Theory (Chapter 2) provides a theoretical framework on the topic of the thesis which includes existing research and an approach of the literature review. Method (Chapter 3) explains and motivates the overall methodology and the choice of methods. Also, it defines how the thesis implements these methods. Results and Analysis (Chapter 4) explains the results, its analysis and interpretation based on the formulated research questions. Discussion (Chapter 5) entails strength of the results and the analysis based on the methodology and

methods used. Conclusion (Chapter 6) draws conclusions by answering the three research questions and presents recommendations for future research and the limitations of the thesis.

2 Theory

This chapter provides a theoretical framework on the topic of the thesis. Initially, it describes existing research and an approach of literature review called systematic literature review. Afterward, a means applied to collect and select relevant and of quality literature. Following this, it alienates the outcome of the literature review which focuses on the aim of the thesis. This outcome includes a detailed literature review and the theory which oversee the rats TB detection performance. In conclusion, it presents related studies which motivate data mining application in healthcare and classification technique in tuberculosis diagnosis.

2.1 Existing Research

This section consists of kinds of literature that are relevant to the topic of the thesis. It describes and motivates an overview of the literature review. Boell & Cecez-Kecmanovic (2015) propose that the literature review is a significant part of all aspect of research. In other words, it helps to find related research and bring together their results. Moreover, the literature review assists to show the knowledge gaps that may lead to identifying further research.

Further research is established initially by scrutinizing the overview and the critical analysis of earlier research. Additionally, criticizing the existing knowledge to discover the problem and relevant research questions of future research (Alvesson & Sandberg, 2011). For beginner researchers, the literature review process may have difficulties due to the vast amount of literature. However, Webster & Watson (2002) suggest the use of a systematic literature review (SLR) approach to present, classify, and evaluate the literature. Hence, this thesis uses SLR to search and select the relevant literature.

2.1.1 Systematic Literature Review (SLR)

General strategies to conduct a literature review emphasize on specifying a topic and number of literature sources involved in the study. In other words, they are not clear on the systematic means to collect and interpret information or results. Similarly, traditional literature reviews aim to give a detailed explanation of the topic problem and critical assessment of research knowledge without identifying a proper way to answer the research questions and acquire the evidence of the results (Finfgeld-Connett & Johnson, 2013). However, for rigor research literature review should not only check for the strengths and weaknesses of the research. Literature reviews must give evidence that answers the research questions (Boell & Cecez-Kecmanovic, 2015).

The SLR is a complement to traditional literature reviews which strengthen the scientific rigor of the research (Okoli & Schabram, 2010). It is the type of literature reviews that use systematic methods to gather and find the kinds and of quality literature (Morrell, 2008). SLR offers current and best evidence available from earlier research that is relevant to the identified research questions. Thus, its advantage is on the ability to find, select, assess and combine evidence from the earlier related literature (Webster & Watson, 2002).

Furthermore, the SLR is a standardized method for literature reviews which are rigorous, replicable, unbiased and clear (Okoli & Schabram, 2010; Oates et al., 2012). It also specifies a systematic means to search for relevant and quality literature. Following this, the SLR is superior to other approaches for conducting literature reviews. However, since there is no such justification, it remains as a general approach to conducting literature reviews (Boell & Cecez-Kecmanovic, 2015).

2.1.2 Critical Analysis of the Literature

This sub-section focuses on identifying the purpose of the thesis by gathering different literature from earlier related researches. Also, it aims at creating knowledge on the topic by describing several topic concepts such as rats' TB detection performance, signal detection theory, data mining application in healthcare, and classification technique in TB diagnosis. Based on Webster & Watson (2002), this thesis applies two steps to collect, assess and select the quality and relevance of the collected existing related research. These steps are as follows:

- Literature search

This step aims to present and motivate the approach used to the search process and the search terms applied. Furthermore, it describes in detail the various sources of literature including articles and books and exclusion of other unpublished sources.

- Literature selection

Since the literature sources usually provide a vast amount of information, it is of great benefit to involving only the relevant literature. To implement this, it is of importance to use a set of criteria to control the access of information.

2.1.3 Literature Search

The literature search aimed to collect information from earlier related studies and focused on the literature with the evidence on the topic for the finding's consistency. The terms giant African tuberculosis in human sputum and using "classification technique" in the diagnosis of

tuberculosis were used separately to search for literature from Google Scholar and Uppsala University databases. The reasons behind the usage of the above sources are; easy to use and their ability to give a large amount of literature. Moreover, the accessibility in Google Scholar is free to use and, the Uppsala University databases are accessible all the time. Additionally, this search excludes sources such as newspaper articles and blog posts.

Furthermore, the search process focused on the published literature to make use of only relevant and quality literature on the topic for review. Google Scholar and Uppsala University database yielded 16,800 and 942 results respectively for the first search term. Also, for the second search term, the results were 2,130 and 146 respectively. Since a Google Scholar search engine is connected to the other databases like Uppsala University, it is true that most of the literature were found in both the databases. Hence, the thesis selected only the first 41 and 23 literature that appeared to both databases for the first and second search terms, respectively since they were relevant to the subject and presumed to address the aim of the thesis.

2.1.4 Literature Selection

It was of great benefit to using only the literature out of 64 that influences the consistency of the results. And as such, 2-stages used to acquire the relevant and of quality literature on the topic (Smith et al., 2011).

- Identifying relevant literature

This stage involved a comprehensive evaluation of the 64 kinds of literature obtained from the literature search process. In other words, this assessment enabled to prove the literature strength and support in making recommendations for future research. The relevance identification started by examining the title and abstract of the chosen 64 kinds of literature. The relevance of the literature was obtained when the title and the abstract related to the topic and it was peer-reviewed. Afterward, only 44 out of the 64 were identified as relevant for the next phase of the quality appraisal.

- Quality assessment of the literature

The literature quality appraisal process started by perusing the literature and finally concentrating on the abstract and conclusion to confirm its consistency with the topic. When the author had previous related publications and were easy to read, these kinds of literature manifested as of quality. Following this, only 33 relevant and quality-appraised kinds of literature was used as the source of information to support the topic of the thesis.

2.2 Tuberculosis Detection Rats

Microscopy is the most used device to detect TB in developing countries. However, its effectiveness is still a problem (Poling et al., 2011). Tanzanian Ministry of Health permitted APOPO to use rats in a second-line screening of sputum samples (Weetjens et al., 2009). This screening is aiming at confirming people with or without the TB. It usually happens after the first diagnosis performed by laboratory technicians from Directly Observed Treatment, Short-course (DOTS) centers. These rats are found in sub-Saharan Africa and have features such as large-size and long-lived for up to eight years in captivity versus two years in the wild (Mgode et al., 2018). The study conducted by Ellis et al. (2017) shows that they are resistant to local parasites, diseases and need relatively simple care. Not only that but also a highly developed sense of smell increases their potential for use to detect TB bacteria in sputum samples (Poling et al., 2011).

As APOPO center reproduces these rats, it is of great importance for every rat to carry out various stages before presented to the metal cage for TB detection tasks (Ellis et al., 2017). At the preliminary stage and young ages, rat pups start to interact with humans. However, during 3 to 6 weeks of age rats are getting familiar to everyday smells, sounds, and sights; and given food by trainers. Afterward, clicker training and discrimination training begin by using operant conditioning techniques such as reward or punishment. During clicker training, rats learn to approach the trainer when the click sounds to get food as a reward from trainers through a plastic tube attached to a syringe (Poling et al., 2011).

Clicker training is the first training after socialization which uses a metal cage of a 2-cm hole in the floor below to present a sample with TB bacteria (Poling et al., 2011). And as such, rats learn to differentiate frequently sounds just before getting food. The food reflects as a reward when approaching trainers. However, in this training, there are no TB negative samples presented in the cage rather TB positive because the aim is to support rats to recognize sounds. After the success in clicker training, the second training called discrimination training starts (Poling et al., 2016).

Discrimination training uses a three-hole cage to present TB positive and TB negative samples to trained rats. This training is aiming at enabling rats to distinguish between TB positive and TB negative samples from the presented samples in a cage (Mulder et al., 2013). The study conducted by Mgode et al. (2018) pinpoints that during the training, rats learn to pause for about 3 seconds to the sample hole with TB bacteria and take 1 second to the sample

hole without bacteria. Furthermore, rats receive food (bananas mixed with food pellets, both mashed) as a reward only if they paused correctly at holes with positive control samples. After rats' success in both kinds of training, they become eligible for TB detection tasks (Poling et al., 2011). However, sputum sample from DOTS centers first undergoes heat inactivation to inactivate microorganisms. Following this, trainers present the containment bars with inactivated samples to trained rats for TB detection tasks (Mgode et al., 2018).

Using these rats in sniffing sputum samples together with microscopy have much-increased TB detection case findings in developing countries (Ellis et al., 2017). In Tanzania from Morogoro and Dar es Salaam hospitals the detection rate for presumptive TB patients increased by 44% in 2009, by 43% in 2010, and by 39% in 2014. Moreover, from Maputo, Mozambique detection rate in presumptive TB patients have increased by 53% since 2013 (Poling et al., 2016). Therefore, to date, there is an increase in new case findings of about 40% in Mozambique (Ellis et al., 2017). Local hospital TB clinics as DOTS centers missed finding the TB positive samples during the first screening. As a result, the trained rats increased these new cases by identifying samples with and without TB (Mgode et al.,2018).

Moreover, the study conducted by Mgode et al., (2018) shows that rats succeeded to detect new TB results from sputum samples. However, the DOTS centers microscopy failed to recognize the TB positive results during the first screening from the local hospital TB clinics in Tanzania. The new TB positive cases identified by these rats were 4,793 TB patients from January 2011 to June 2015. Following this, rats have high sensitivity and low specificity in detecting the TB samples than microscopy.

The usefulness of this technology is due to the rats' rapid diagnostic speed which can test up to 100 samples in 20 minutes where a laboratory technician can take about four days when using the microscopy. The speed of a single rat to test hundreds of samples reveals that this testing is not expensive. Detection rat technology is of great benefit to the community and public health hospitals. It provides quick results on the same-day and reduces the high workload of TB samples from local hospital TB clinics which may lead to many false-negative results (Ellis et al., 2017). However, the detection performance of rats may either depend on the success of the kinds of training or the characteristics of an individual rat or a group of rats (Poling et al., 2016).

Detection performance may depend on rats' characteristics such as age and sex (Brushfield et al., 2008). However, no significant difference in detection performance between

male and female rats (Ellis et al., 2017). Moreover, the study conducted by Ellis et al. (2017) identifies that time of the day of training may influence the detection performance. Additionally, Mgode et al. (2018) propose that rats are more precise on samples with a lower bacterial count. Nevertheless, an individual rat may fail to detect samples. The increase of failure may either depend on the characteristics of the rats or unsuccessful early training. As a result, it may show poor reliability on both TB-positive and TB-negative samples. However, other rats may have consistent training, but an unidentified health problem may result in their inability (Ellis et al., 2017).

2.3 Signal Detection Theory

Signal detection theory (often abridged as SDT) used to present the theoretical concepts that guide the analysis of the thesis. SDT states that the detection of a stimulus depends on both the intensity of the stimulus and the physical or psychological state of the individual (Green & Swets, 1966). Moreover, SDT is a theoretical and empirical framework that offers to understand how features of the stimulus, individual factors, and background stimuli affect performance on stimulus-discrimination detection tasks (Mahoney et al., 2013).

SDT is used most in various fields, including medical diagnosis that needs better strategies for decision making. Primarily, the psychophysics experiment used SDT to pinpoint the relationship between stimulus characteristics. As a result, the characteristics influenced the discriminative responses. Therefore, in medical diagnosis, SDT helps to distinguish between the discrimination-sensitivity of an individual-participant and the discriminative response in the task. Hence, SDT has become of interest in separating and learning discriminative responses (Samuel & Snodgrass, 2015).

In an experimental study, SDT assessed individuals' discriminative response abilities based on the given stimuli and conditions. Individuals' were exposed to many detection trials to make discriminative responses (Yes or No) (Samuel & Snodgrass, 2015). Similarly, rats discriminative-detection abilities in the same stimulus intensity may decide their TB detection responses. In other words, this evaluation depends on the operant conditioning techniques such as reward or punishment which enforce rats to portray discriminative detection responses (Poling et al., 2011).

SDT in rats TB detection task has four concepts (Hit, false alarm, miss, and correct rejection) which govern its operation on several stimulus detection tasks. Hit means detection responses when TB bacterium is present in the sample while false alarm entails detection

responses when the TB bacterium is absent. Moreover, miss inferring no detection responses when the TB bacterium is present. Additionally, correct rejection implies no detection responses when the TB bacterium is absent (Poling et al., 2011). However, the analysis of formulated research questions focused on the two concepts. These concepts include Hit and correct rejection. Hit and correct rejection (sensitivity and specificity) are correct detection responses which imply rats TB detection performance used for analysis while false alarm and miss reflect only incorrect. In other words, the hit means the fraction of the sum of hits and the sum of total hits and total misses. Also, multiplying the result by 100% for a percentage measure. Additionally, correct rejection entails the fraction of the sum of misses and the sum of total misses and total false alarms and multiplying the result by 100% (Mahoney et al., 2013).

As mentioned before that stimulus is one of the dependent factors for SDT which facilitates the detection responses. Phosphate-buffered saline (PBS) added to the sample and it was easier to detect the TB positive sample. In other words, sensitivity was higher than specificity due to the extra PBS which increased the strength of the stimulus (Poling et al., (2013). These variations may occur between trained rats and may influence either high sensitivity and low specificity or low sensitivity and high specificity (Mulder et al., 2017). During the experiment rats offered per-patient high sensitivity and low specificity of about 81.9% and 56.9% respectively (Mahoney et al., 2013). On other hands, other rats have detection performance with a low sensitivity of 56.9% and high specificity of 80.5% (Reither et al., 2015). However, TB diagnoses tools with high sensitivity and low specificity are of utmost importance (Ellis et al., 2011).

As SDT also depends on the physical or psychological state of the individual-participant for the detection responses, other rats may or may not have many trials to give the detection performance. TB detection performance may vary based on the factors of an individual-participant rat (Ellis et al., 2017). Following this, TB detection performance between rats sometimes may vary due to various factors and thus affect sensitivity and specificity of a rat. These rats' factors may include age, sex, time of day, and bacterial count. Based on the experience, as rats become older and the weight increases the detection performance decrease (Brushfield et al., 2008, Ellis et al., 2017, Mgode et al., 2018). However, other rats might not show a significant difference between male and female detection performance (Ellis et al., 2017).

Based on these two concepts, rats TB detection performance ability varies between trained rats. Like SDT which proposes that an individual participant may fail to show detection

response depending on their sample criterion. However, the sample may either have or not have the stimuli (Samuel & Snodgrass, 2015). Therefore, it is of importance to understand the most factors of rats which enhance TB discriminative-detection performance.

Table 1: Rat responses in Tuberculosis detection task. (Green, D.M. and Swets, J.A.1966)

Rats responses	TB bacteria (stimuli) present	TB bacteria (stimuli) absent
Detection response	Hit	False alarm
No detection response	Miss	Correct rejection

2.4 Data Mining in Healthcare

Technological advancement has led to the growth and rapid increase of medical data generated from different operations in the healthcare sectors. These operations may include diagnosis, medication, prescription or prevention of disease (Chaurasia & Pal, 2014). Healthcare sectors produce a vast amount of data and make it difficult to find interesting patterns or inefficiencies from data by using the traditional approach (PrasannaDesikan et al., 2011). Therefore, data mining is more important to describe and identify the relationship between the data pattern to enhance better strategies and decision-making (Dubey et al. 2016).

Data mining is the process of discovering knowledge by extracting and identifying useful information and patterns from a structured dataset. It is used mostly in many fields such as scientific discovery, marketing, surveillance, fraud detection, and medical to discover hidden patterns (Sharda et al., 2014). In medical services, data mining supports various areas such as medical tests, the discovery of relationships among clinical and diagnosis data, and medication Moreover, it helps healthcare sectors to learn more about their operations, generate effective strategies and reduce costs (Chaurasia & Pal, 2014). Furthermore, data mining enables healthcare sectors to achieve their goals (Dubey et al. 2016).

The data mining process in healthcare depends on different techniques which include classification, clustering, and association for its operation. These techniques help to learn past data and detect knowledge patterns. Classification technique builds predictive models to predict future events from the manipulated data. The clustering technique groups each instance into a specific group or category based on the characteristic's commonalities. Moreover, the association technique provides the rules which help to find a relationship between the interesting patterns. Therefore, they all offer solutions to real-world health problems like diagnosis and treatment of diseases (PrasannaDesikan et al., 2011). However, the applicability of these techniques depends on the context or aim of the task (Chaurasia & Pal, 2013).

Classification is the data mining technique which operates by building predictive models that categorize and assign a label to manipulated and newly encountered instances. These predictive models help to solve multi-classification problems through prediction and analysis. In healthcare sectors classification is of great value to find interesting patterns in various aspects like diagnosis, treatment or planning (Sharda et al., 2014). The study conducted by Ramana et al., (2011) used the classification technique with bagging and boosting to build the predictive model for the diagnosis of Liver disease.

Not only that but also the classification technique used bagging algorithm to diagnose the heart disease from the past medical data set. Hence, classification is mostly used in healthcare sectors to predict the presence or absence of any disease. One of the advantages of classification is that it helps the medical specialists to make decisions on patients based on the diagnosis. As a result, there is a possibility of preventing disease transmission since suspected patients can start treatment earlier. Fundamentally, the classification technique supports healthcare sectors to predict future events and gain insight into the patients' data based on past data to improve their operations (My Chau et al., 2009).

2.4.1 Classification Technique in the Diagnosis of Tuberculosis

The studies conducted by Ameri et al. (2014) and Nagabhushanam et al. (2013) suggest a classification technique in the diagnosis of tuberculosis disease. However, this technique depends on various algorithms for the implementation. The classification algorithms recently used in the diagnosis of tuberculosis include Decision Tree, Random Forest, Naive Bayes, Support Vector Machine (SVM), and Artificial Neural Network (ANN) (Suresh et al., (2018). These algorithms have been successfully implemented and recognized in solving several classification problems in tuberculosis diagnosis due to their high generalization performance (Nagabhushanam et al., 2013).

The decision tree algorithm was used to build a predictive model which learned patterns from the past medical data. The predictive model aimed at helping the doctors to predict and diagnosis the tuberculosis disease. And as such, the predictive model played roles of a decision support tool in clinics (Asha et al., 2011). In other hands, the decision tree algorithm was applied to find the characteristics of the disease. The decision tree generated rules that enabled the prediction of the patient's disease status. These rules are simple and easy to understand and interpret (Ameri et al., 2014). Furthermore, Suresh & Arulanandam (2018) applied decision tree algorithm purposely to support on the diagnosis decision when suspected tuberculosis

patients must start treatment. As a result, the built predictive model helped to find patients with or without tuberculosis based on the patients manipulated suspected variables.

Not only that but also Asha et al. (2011) used classification algorithms such as Random Forest and naive Bayes to detect tuberculosis. Following this, the algorithm comparison is of great importance to find a reliable algorithm in the given data. The study conducted by Ayas & Ekinci (2014) used random forest algorithm to categorize the Mycobacterium tuberculosis based on local color distributions and regions. Moreover, the study of Maniya et al. (2011) used a naive Bayes algorithm to classify patients affected by tuberculosis into two classes which are least probable and most probable. This algorithm learned patterns from the past data to discover and extract hidden interesting information by building the predictive model that categorized tuberculosis into two classes (Yes and No).

Besides, Support Vector Machine (SVM) was used as the classification algorithm to build a model for categorizing the patient's samples into TB and non-TB. Apart from classifying the samples, SVM also helped to decide which compounds in the database was useful to classify these samples (Kolk et al., 2012). Furthermore, Ameri et al. (2014) performed a study which identified the most influential factors on osteoporosis using the C.5.0 algorithm and artificial neural network. Additionally, the study of Nagabhushanam et al. (2013) used the classification technique to predict tuberculosis by using the multi-layer Neural Networks. Moreover, neural networks are used to categorize tuberculosis patients based on the laboratory and demographic characteristics (Tamer et al., 2012).

Therefore, prior studies conducted by Brushfield et al., (2008); Ellis et al., (2017); and Mgode et al., (2018) using rats for TB detection proposed that TB detection performance sometimes may depend on the rat's factors such as age, sex, time of day, and bacteria count. However, based on the experience as rats become older and the weight increases the detection performance decrease. In other words, there is no empirical evidence on these factors and the trend of impact is not clear. Despite the prior studies, there are no studies related to using data mining techniques to predict factors for rats TB detection performance. Since the rat's technology is advantageous, it is useful to understand the factors that influence their TB detection performance. One of the rat's TB technology advantages is that it helps to reduce false results. Following this, the center must consider these factors to maintain its usefulness. Thus, this thesis aimed to predict the factors that influence rats TB detection performance using data mining techniques.

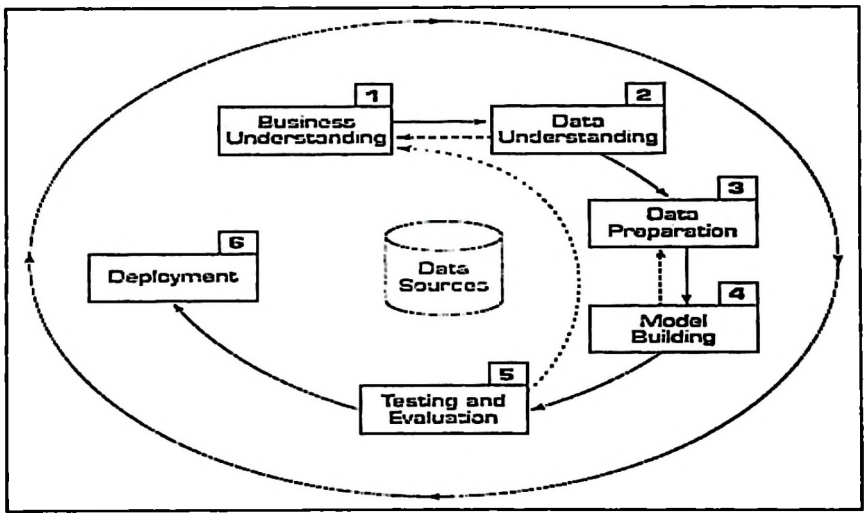
3 Method

This chapter describes the overall methodology used when conducting this thesis. It includes the data mining process, technique, and algorithms. Moreover, justifies the selected choices and presents ethical approval to use the data.

3.1 Data Mining Process

This thesis applied the CRISP-DM methodology (Cross Industry Standard Process for Data Mining) to describe the systematic and organized approach in the data mining process. CRISP-DM is an open standard process which consists of six phases for data mining. These phases are business understanding, data understanding, data preparation, model building, testing and evaluation, and deployment. Moreover, CRISP-DM is the most popular and non-proprietary standard data mining process (Sharda et al., (2014). Figure 1 shows that these phases often operate in sequential order. Despite the mentioned order, this analysis used back and forth to influence the desired results. Therefore, CRISP-DM was used as a framework to answer the formulated research questions.

Figure 1: The Six-Phases of CRISP-DM Data Mining Process. (Sharda et al., 2014)



3.1.1 Business Understanding

The first phase of CRISP-DM aimed to know what the center needs from a business perspective. To answer this question, it was of importance to start with a detailed understanding of the aim, requirements, and constraints of the center. This understanding was created by

extracting the knowledge of the application domain. This knowledge helped to formulate the thesis problem and research questions. Moreover, it was easier to choose a suitable data mining method that might answer the formulated research questions and meet the purpose.

3.1.2 Data Understanding

Following the business understanding, this phase concentrated on the access, description, and identification of the relevant data. And as such, it helped to find the value of the given data. The data source for this thesis was the APOPO TB training and research center in Morogoro Tanzania. The given datasets composed of rats TB detection performance variables from 2011 to 2019 years and rats weight variables from 2012 to 2019. They were three datasets in excel files with quantitative and qualitative data. Table 2 describes the four weight variables of one of the datasets named RAT_WEIGHT dataset that contained five female rats with 1438 observations. However, the fifth rat in the RAT_WEIGHT dataset had no corresponding detection performance variables in the other two datasets and thus was disqualified. Hence, this thesis used only the four female rats which were consistent with the other two datasets.

Focusing on the two remained datasets named DetectionRatsData, each dataset had two worksheets with two female rats' detection performance data. One with Happy and Catia data and the other with Sofia and Mkuta data. Table 3 describes 18 variables (1 dependent and 17 independent) found on these datasets and a separate DOB variable inclusive depicted in a summary table from the original data. The DOB variable helped to find the age of rats in a specific session day. Independent variables mean the rats factors that may influence TB detection performance while the dependent variable implies TB detection performance (HIT) as a target class for classification.

Based on the independent variables, Table 3 shows that DOTS_NAME, RAT_NAME, and GENDER had string data except for the rest such as SESSION_DATE, Age, START_TIME, END_TIME, and DOB that contained integer, DateTime, and Date data. TB detection performance variable (HIT) is the dependent and categorical variable with boolean data in the form of TRUE or FALSE. Considering the signal detection theory, TRUE means rats' sensitivity (Hit) while FALSE implies rat's specificity (Correct rejection). Therefore, the dependent variable was entirely used to predict the influencing factors and assign the detection performance class of every given rat into either TRUE or FALSE.

Table 2: RAT_WEIGHT Dataset Description

Number	Variable Name	Data Type	Description
1	ID_RAT	Integer	Identification of rat
2	RAT_NAME	String	Name of rat
3	WEIGHT_DATE	Date	Date when the weight of the rat was measured
4	WEIGHT	Integer	Weight of the rat

Table 3: DetectionRatsData Dataset Description

Number	Variable Name	Data type	Description
1	DOTS_NAME	String	Name of the DOTS center
2	DOTS_PATIENTS_NUMBER	Integer	Number of patients from DOTS center
3	ENTRY_YEAR	Integer	A year when patient attend DOTS center
4	ID_SAMPLE	Integer	Identification of the sample
5	ID_BL_DOTS	Integer	Identification of the blood from DOTS center
6	HIT	Boolean	Rats TB detection performance (categorical variable)
7	ID_BL_APOPO	Integer	Identification of the blood from APOPO center
8	ID_CONFIGURATION	Integer	Identification of the cage during training
9	ID_BL_FM	Integer	Identification of the fluorescence microscope
10	ID_EVALUATION_SESSION	Integer	Identification of evaluation session
11	SESSION_DATE	Date	Date when a session performed
12	ID_RAT	Integer	Identification of the rat
13	RAT_NAME	String	Name of rat
14	GENDER	String	Sex of rat
15	Age	Integer	Age of rat
16	START_TIME	DateTime	Date and time when the detection task started
17	END_TIME	DateTime	Date and time when the detection task ended
18	DOB	Date	Date when rat was born

3.1.3 Data Preparation

This data preparation or preprocessing phase was used to prepare the data obtained from the second phase into a useful structure for data mining methods. Fundamentally, it covered the approach of data analysis. This phase enhanced data consistency from the raw given data. Therefore, it used the four main steps which are data consolidation, data cleaning, data transformation, and data reduction as shown in Figure 2 suggested by Sharda et al. (2014) to prepare the given raw data into the required format.

The data consolidation step focused on access, selection, and integration of the given data. After the access of data from the second phase, it was of benefit to identify the relevant variables from the given raw data. To enhance clarity, merging of the four separate worksheets

from the other two DetectionRatsData datasets into a single file called Rats dataset was performed. Hence, the new merged data file (Rats) consisted of 17 columns and a total of 471,133 observations from 2011 to 2019 years.

The second step of data cleaning dealt with the removal of irrelevant variables and empty rows from the Rats dataset to prevent noises, outliers, and inconsistencies from the data. This step started by eliminating undesirable variables, rows, and observations from 2011 to 2013 and 2019 years to make use only the intended data from 2014 to 2018 years. Following this, it was useful to find and remove inconsistencies from the START_TIME and END_TIME rows to prevent outliers. These variables had the same date values and were incorrect since they differ from the SESSION_DATE values. However, the time values were correct. In this conception, there was no imputation of the missing values. On other hands, from the RAT_WEIGHT dataset, since it consisted of data from 2012 to 2019 years and the disqualified fifth rat, it was of importance to remove data from 2012 to 2013, and 2019 and the fifth rat from the RAT_WEIGHT dataset to maintain consistency with the DetectionRatsData datasets.

The third step of data transformation focused on creating other variables and converting data variables from one data type to another to ease the data mining process. And as such, this step began by creating new variables which include Age, Av_Weight_Per_Year, Session_Start_Time, and Session_Completion_Time. The difference between rats SESSION_DATE and DOB (Date of Birth) resulted in the new Age variable. However, the original two datasets (DetectionRatsData) had the Age variable with incorrect values since all rats contained the same current age throughout the years while the detection activities happened in different years (2011-2019), but it was renamed as AgeOrg to make it different from the new Age variable created.

The Av_Weight_Per_Year variable implies the average weight of each rat per year created from the WEIGHT variable of the RAT_WEIGHT dataset. Moreover, this Av_Weight_Per_Year variable was merged into the new file (Rats dataset) to acquire the desired dataset for the analysis. All rats within a specific year were given the same average weight to enhance data consistency since the two files (DetectionRatsData datasets) consisted of many observations describing the daily detection tasks. However, most of them missed their corresponding weights since the weight of the rats from the RAT_WEIGHT dataset was measured on every week while detection tasks performed daily. The average weight might reduce bias in variables. Moreover, Session_Start_Time is the variable created after renaming Start_Time variable from START_TIME variable of the original given data which contained

values in date and time format since data was recorded based on sessions as indicated in the SESSION_DATE variable. Therefore, the Session_Start_Time consisted of converted values in hours and not in date and time to enhance the data mining process.

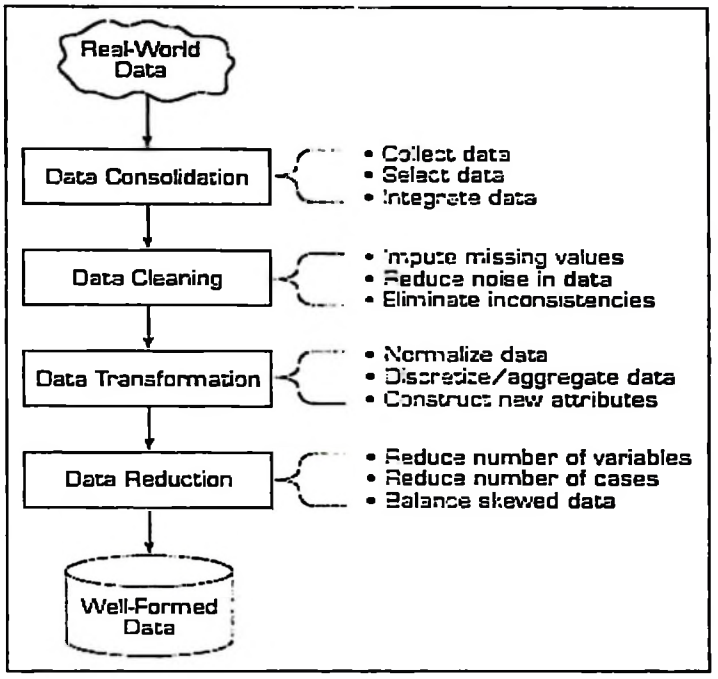
Additionally, Session_Completion_Time variable entails the difference between rats Start_Time and End_Time that were renamed from START_TIME and END_TIME variables respectively from the original given data. However, Start_Time variable was also renamed to Session_Start_Time for consistency with SESSION_DATE variable since it was the factor to investigate than End_Time. The data in Start_Time and End_Time variables consisted of values in date and time format. In other words, it was vital to transforming them into hours to get their differences in minutes and per sessions as Session_Completion_Time variable. And as such, it was created purposely to help in finding whether the session completion time influences detection performance.

Not only that but also, it was useful to rename the categorical variable HIT to Performance for understanding. In transforming data, the conversions applied to DOTS, Rat, and Performance variables into a factor data type. Moreover, Age, Session_Start_Time, Session_Completion_Time, and Av_Weight_Per_Year variables converted into a numeric data type. Therefore, the data transformation process helped to acquire well-formatted data for the data mining process. Following the three above steps, Rats dataset contained 7 variables with 365,843 observations.

The data reduction step focused on reducing the number of observations into the manageable size to ease the processing process and analysis. Random sampling was applied to deduct the given records from 365,843 to 200,000. Based on personal computer RAM space and processing abilities, it was difficult to process the given large dataset when creating the random forest model. Therefore, the data used for this analysis consisted of 7 variables and 200,000 observations.

Moreover, in data analysis, this thesis applied the R-language as a data mining tool to transform the given data into useful data types and reduced the number of observations. RStudio is a data mining tool and an integrated development environment for R, is a free programming language with extensive modeling and quality graphs resources (Zhao, 2013). Furthermore, most data miners used R in scientific research. The following figure shows the four steps applied in data preparation.

Figure 2: Data Preprocessing Steps. (Sharda et al., 2014)



The data preparation phase was of benefit to identify well-formatted data for rats' factors that may influence TB detection performance. Table 4 shows the variables names and their data types before and after data preprocessing. Moreover, it depicts the newly created variables (the ones with dashes) and renamed variables (remaining ones) used in the analysis from the original APOPO datasets (RatsDetectionData) for consistency.

Table 4: Variables names before and after data preprocessing

Variables				
Number	Name before data preprocessing	Data type	Name after data preprocessing	Data type
1	DOTS_NAME	String	DOTS	Factor
2	HIT	Boolean	Performance	Factor
3	SESSION_DATE	Date	Session_Date	Date
4	RAT_NAME	String	Rat	Factor
5	GENDER	String	GENDER	String
6	Age	Integer	AgeOrg	Integer
7	-	-	Age	Numeric
8	START_TIME	DateTime	Start_Time	Time
9	END_TIME	DateTime	End_Time	Time
10	-	-	Session_Start_Time	Numeric/Time
11	-	-	Session_Completion_Time	Numeric/Time
12	DOB	Date	DOB	Date
13	WEIGHT	Integer	WEIGHT	Integer
14	-	-	Av_Weight_Per_Year	Numeric

Moreover, the following tables depict descriptive summary for independent variables (continuous and nominal) and the dependent variable (categorical) used in the analysis.

Table 5: Summary for rats' continuous variables (Independent variables)

	Age	Av_Weight_Per_Year (g)	Session_Start_Time (hrs)	Session_Completion_Time (min)
Min	0.79	843.7	8:00	1.00
Max	7.95	1054.8	18:00	129.00
Mean	3.83	899.4	12:16	10.49
Median	3.71	866.8	12:00	10.00

Table 5 depicts data for rats' continuous variables where the younger and older rats have ages of 0.79 and 7.95 years respectively with the mean and median age of 3.83 and 3.71 years. Moreover, the rats' lowest and highest average weight per year are 843.7g and 1054.8g respectively, with the mean and median of 899.4g and 866.8g. Besides, the table shows that their lowest and highest session start time are 8:00 and 18:00 hours, with the mean and median of 12:16 and 12:00. Furthermore, the minimum and maximum session completion time is 1 and 129 minutes, with the mean and median of 10.49 and 10.00 minutes. Since the mean and median are not equal, it manifests that the data used for this analysis lack normal distribution.

Apart from the continuous independent variable's tables, below are tables for nominal data variables.

Table 6: Summary for rats' name and number of observations detected.

Rat	Gender	Number of observations
Sofia	F	50448
Catia	F	50271
Happy	F	50035
Mkuta	F	49246

The data from Table 6 depicts four female named rats used in this analysis with their observations completed during the detection tasks. Sofia completed many numbers of observations when compared to all. Besides, Mkuta has few numbers of observations whereas its youngest may have caused this performance. Moreover, Happy is older and has a few observations than Sofia and Catia. The removal of irrelevant data from 2011 to 2013 and 2019 could have led this since there is a possibility that Happy had many observations in the irrelevant years. Furthermore, the table shows that there is sex inequality in the data given from the DOTS center, since all rats are female. Therefore, the following table shows the first four DOTS centers and their number of observations completed in the second screening.

Table 7: Summary for DOTS names and number of number of observations detected

DOTS	Number of observations
Mwananyamala	20930
Mbagala Rangi 3	17385
Ukonga	15973
Amana	10136

Table 7 indicates some of the DOTS centers out of 81 which performed the first screening of the given data. These are the only first four with many completed observations than the rest. However, Mwananyamala consisted of many observations when compared to others. The high population of the area presumed high transmission of the disease. Since Table 5,6, and 7 contained data for independent variables, while Table 8 shows the dependent variable used in this analysis.

Table 8: Summary for rats' performance (Dependent variable) and the number of observations detected

Performance	Number of observations
FALSE	157686
TRUE	42314

From Table 8, there is performance inequality in the distribution of rats' detected observations. TRUE observations are far less by 21.2% than FALSE of about 78.8% for all observations. And as such, this analysis used more FALSE than TRUE.

Despite the descriptive summary of the variables identified above. Figure 3 shows the extent to which each variable relates to one another. Additionally, it pinpoints the correlation between each independent variable and the dependent variable (Performance). In statistics, one of the measurements of variables correlation is Correlation coefficients. These are used to measure the strength of a relationship between two variables (Sharda et al., 2014). One of the types of correlation coefficients used most in linear regression is the Pearson's correlation (Pearson's R). Pearson's correlation coefficient formula has a correlation coefficient (r) between -1 and 1. And as such, 1 and -1 imply strong positive and negative relationships, respectively. However, there is no relationship between two variables when the correlation coefficient is 0.

Figure 3: Dependent and Independent variables correlation

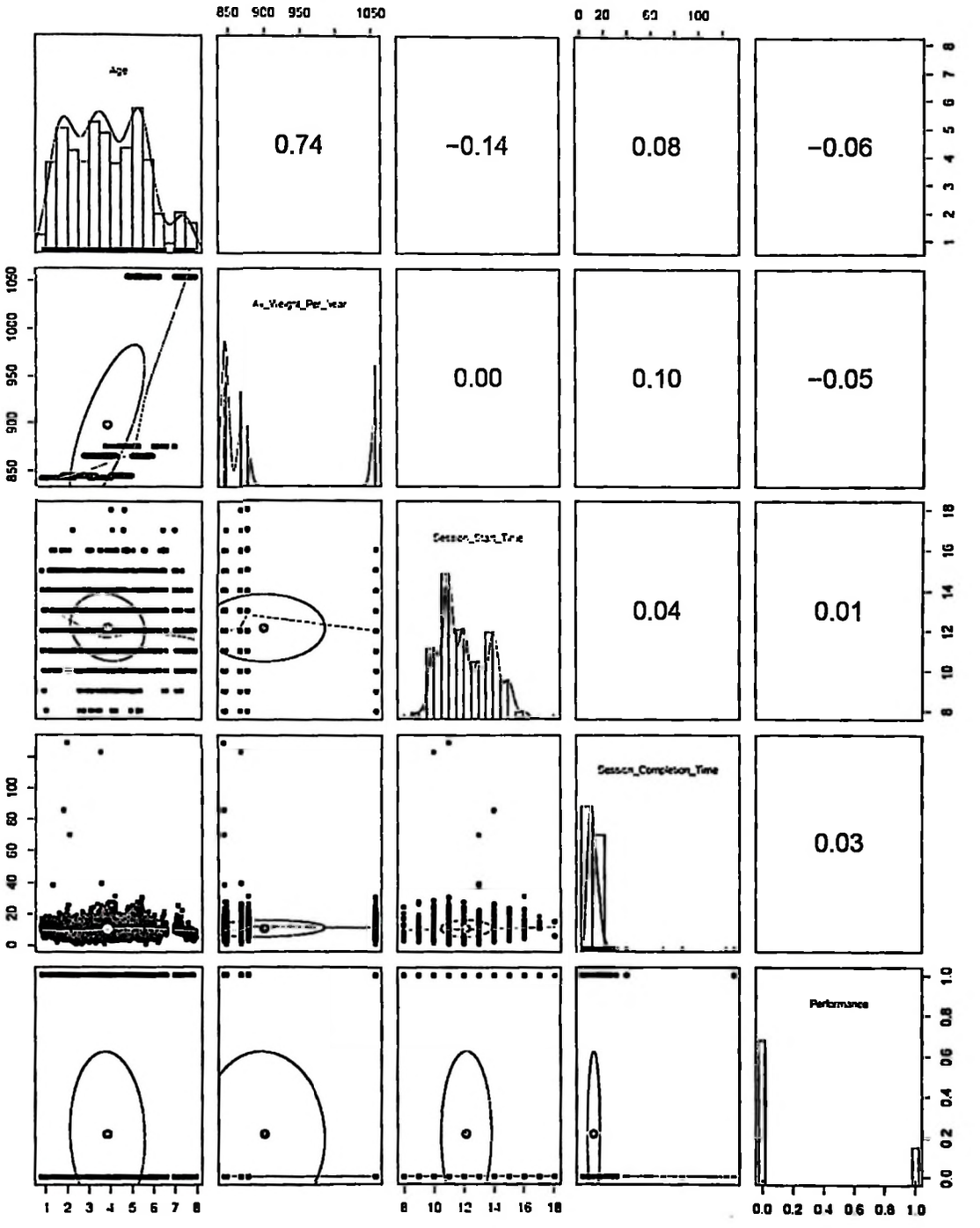


Figure 3 shows that there is a weak positive relationship between Age and Av_Weight_Per_Year since it has the colleration coefficient of 0.74. Moreover, Age and Session_Start_Time variables have a weak negative relationship with a coefficient of -0.14. However, there is no relationship between Av_Weight_Per_Year and Session_Start_Time since the correlation coefficient is 0.00. Not only that but also, there is a weak positive relationship of 0.10 correlation coefficient between Av_Weight_Per_Year and Session_Completion_Time. On other hands, Session_Start_Time and Session_Completion_Time pinpoint a weak positive relationship of 0.04 correlation coefficient.

Based on the dependent variable (Performance), Figure 3 shows that there are weak positive and negative relationships between dependent and independent variables. In this sense, Performance relates weakly with Age, Av_Weight_Per_Year, Session_Start_Time, and Session_Completion_Time with the correlation coefficients of -0.06, -0.05, 0.01, and 0.03, respectively.

3.1.4 Model Building

After obtaining the data with the required format, this phase used to select and apply the data mining technique and algorithms based on the nature of the data that might address the need of the center. This phase applied classification technique to build predictive models that assigned a class for each rat in the given data and predicted the factors that influence rats TB detection performance. Not only that but also the predictive models might be useful to place and predict the new instances (rats) with unknown labels into their respective classes.

Classification is a supervised learning technique in data mining and machine learning that learn the relationship or patterns between independent variables (input) and the dependent variable (output) from the past data and classify each data item into a predefined class label. Moreover, it is the most often used data mining technique for real-world problems (Chaurasia & Pal, 2014). Before presenting the preprocessed data to the algorithms for the learning process, R was entirely used to partition the data into training data and test data by using a simple split estimation method as shown in Figure 4. It is the most popular method which divided two-thirds of the data in the training data and the one-third in the test data (Sharda et al., 2014).

Figure 4: Simple Random Data Splitting. (Sharda et al., 2014)

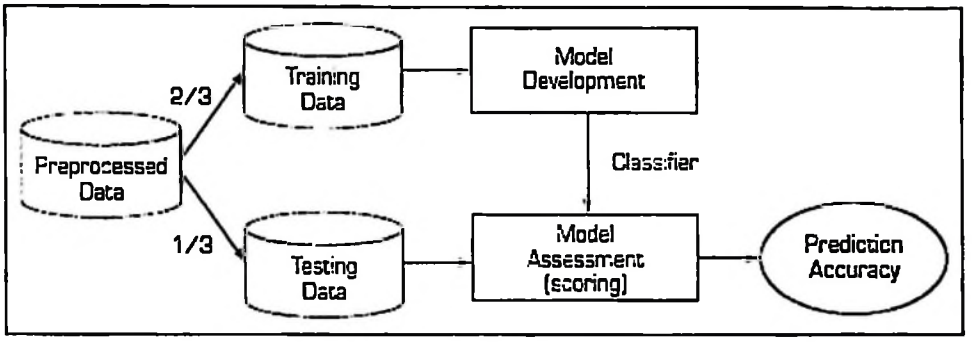


Figure 4 shows that ⅔ of the preprocessed data was used to build a predictive model while the remaining ⅓ used to assess the predictive model classification accuracy. However, the classification process applied three steps. First, the learning step responsible for building the predictive model that assigned predefined classes for each instance by learning and analyzing the training data variables manipulated. Second, the validating step that estimated the accuracy of the predictive model using test data. Third, the application step to allow the real application of the predictive model to predict new unseen data and unknown label class since the predictive model accuracy is considered acceptable.

The acceptable Confidence Interval (CI) in the medical field is 95% which implies that the probability of observing the differences in the data by chance is just 5%. Following this, the predictive model can act as a classifier in the decision-making process. Despite many classification algorithms used for prediction, this analysis used Decision Tree, Random Forest, and Naïve Bayes. Hence, Table 9 shows the training data and test data used for all the applied classification algorithms.

Table 9: Summary of a simple random data splitting

Type of data	Number of observations
Training data	134000
Testing data	66000

The data from Table 9 indicate that this analysis consisted of 67% training data and 33% test data. The training data were given many observations to build the predictive model while test data were used only to assess the performance of the model generated.

A decision tree algorithm is a supervised classification algorithm which generates the decision tree automatically by examining the weight of each variable used to the extent that each leaf node has the same class (Sharda et al., 2014). It is one of the most commonly used algorithms in classification technique for research purposes. Moreover, this algorithm generates rules that are easy to interpret and understand (Chaurasia & Pal, 2014). The decision tree is a tree-shaped diagram comprises many input variables that may have an impact on classifying different patterns. Additionally, it is known as a decision support classifier which depends on the input to show the possible outcomes (Asha et al., 2011).

The decision tree was generated by recursively dividing the training data until each division consisted of the variables of the same class or values based on conditions. Following this, a split point used in each node to test the manipulated variables and decide the way to divide the data. The split decision focused on the amount of information a computed variable offered in the class (information gain) and its randomness (entropy). As a result, the variable with the highest information gain and the lowest entropy split and tested. The information gain and entropy determined the decision on the split of data and construction of the decision tree. However, the growth of the decision tree influenced deep learning. Control on the parameters used to overcome this problem through pruning first (Sharda et al., 2014). Pruning is the process of reducing the size of decision trees by removing sections of the tree that provide little power to classify instances. And as such, it might increase the generalization and enhanced the prediction accuracy on test data (Chaurasia & Pal, 2014).

The generated decision tree consists of a root node, branches, and leaf nodes. The root node is the node at the top of the tree which implies the most important factor responsible for classifying the observations. The branches represent the pattern classification outcome of a test using one of the variables based on conditions. The leaf nodes placed either before or at the end of the decision tree imply the nodes without children. And as such, they identify the last class choice for a pattern. Moreover, the decision tree formed rules (IF-THEN statements) from the root node to the leaf nodes which are easy to interpret and understand. As a result, they enhanced the discovery of exploratory knowledge on the factors that influence rats TB detection performance. Furthermore, the random forest algorithm was also applied to predict the influencing factors and compare the prediction accuracy of the models.

A random forest algorithm is a supervised classification algorithm used to build multiple decision trees called forest in random during the training process. The choice of most of the trees determined the final decision of the algorithm based on the given manipulated

variables. There is a relationship between the number of trees in the forest and the results. Thus, many trees, the more accurate the result. The motives behind this algorithm are that it can be used for both classification and regression problems and lowers the risk of overfitting (Sharda et al., 2014). Overfitting is a modeling error which occurs when the outcome of the analysis is limited only to specific data. As a result, instead of predicting the whole manipulated data, the model predicts only for that set of data (Chaurasia & Pal, 2014).

In the random forest algorithm, the process of determining the root node and the splitting of the variable nodes were performed randomly from the training data. During the training, no control of parameter (pruning) involved preventing the decrease of the relationship between trees. However, pruning is of importance for the reduction of complexity in variable computation during the training. As a result, the algorithm handled about 500 trees in the ensemble and identified the error rate based on the training data. Following this, the random forest algorithm predicted the factors for detection by pinpointing the mean decrease in Gini values for each variable. Based on this, it was easy to find the association between the most significant factors and other factors. Furthermore, the Naive Bayes algorithm was applied to compare their predictive accuracy by using the same test data as shown in Tables 11, 13, and 15. Moreover, finding the best algorithm with high classification accuracy rates for the given data.

Naive Bayes is the supervised classification algorithm that uses a probability theory (Bayesian Theorem) to generate the classification model. Moreover, to place an instance in the desired class. This theory supported to calculate a set of probabilities by counting the frequency and values of the manipulated variables from the given data (Sharda et al., 2014). The Naive Bayes algorithm is a well-performed algorithm owing to its simplicity in execution time. And as such, it can build a final model that can learn rapidly different classification problems (Ameri et al., 2014). However, this algorithm assumed that all variables were independent of the given data while few real-world applications may agree with this (Asha et al., 2011).

The main advantages of the Naive Bayes algorithm compared to the other two algorithms are the run-time speed on large and complex datasets. Hence, most healthcare field researchers across the world use this algorithm due to its better speed and accuracy. This algorithm identified a priori probability for the dependent variable and conditional probabilities for every independent variable based on the manipulated data. The Naive Bayes algorithm does not show the weights of each variable included in the classification, but it has been used purposely to compare its prediction performance with the results generated from the decision

tree and random forest algorithms (Ameri et al., 2014). According to Chaurasia & Pal. (2014) the chosen algorithms have potential in yielding the desired results in researches

3.1.5 Testing and Evaluation

This phase was used to test and assess the classification performance of the three generated predictive models. The assessment was based on the accuracy metric to show the predictive accuracy of the model from the confusion matrix. However, the confusion matrix has several assessment measures such as confidence interval (CI), sensitivity, and specificity. The confusion matrix is a table used to describe a classification model performance based on the test data. Moreover, it is the most used classification performance measure for predictive models. Furthermore, other measures including speed and interpretability were used to assess the models' performance (Sharda et al., 2014). Thus, the following is the description of the three measurement factors which offer a thorough view of models' classification performance.

- Accuracy

This measure was used to assess the ability of the models' to accurately predict the class label of the test data. And as such, the accuracy reflected the matching between actual class labels of the test data and the class labels of the predicted models. This measurement focused on the accuracy rate, the percentage of test instances that were accurately classified by the predictive model. Based on the confusion matrices as shown in Table 10, 13 and 15 it was easy to compare their classification accuracy rates. The following are the confusion matrix tables for the applied algorithms which show the correct and incorrect predictions based on training data and test data.

Table 10: Decision Tree Algorithm Confusion matrix for training data

Prediction	Actual	
	FALSE	TRUE
FALSE	105773	28227
TRUE	0	0

From Table 10 in the diagonal, the data indicate that 105773 observations were FALSE, and the model predicts them to be FALSE. In other hands, there were 0 TRUE observations and the model predicts them to be TRUE. These are True Positive (TP) and True Negative (TN) observation values which imply correct prediction (classification). Moreover, and not in the diagonal, the data show that 28227 observations TRUE but the model predicts them to be FALSE. Also, 0 FALSE observations but the model predicts them to be TRUE. These are False

Positive (FP) and False Negative (FN) observation values entail incorrect prediction (misclassification). Training data used to build the predictive model and not to assess the classification performance inaccuracy rate. Table 11 shows the confusion matrix for test data based on the decision tree algorithm.

Table 11: Decision Tree Algorithm Confusion matrix for test data

Prediction	Actual	
	FALSE	TRUE
FALSE	51997	14003
TRUE	0	0

Table 11 in the diagonal, the data show that 51997 observations were FALSE, and the model predicts them to be FALSE. At the same time, there were 0 TRUE observations and the model predicts them to be TRUE. These are True Positive (TP) and True Negative (TN) observation values which mean correct prediction (classification). Furthermore, and not in the diagonal, the data pinpoint that 14003 TRUE observations but the model predicts them to be FALSE. In other hands, 0 FALSE observations but the model predicts them to be TRUE. These are False Positive (FP) and False Negative (FN) observation values which entail incorrect prediction (misclassification). The test data used to assess the predictive model performance accuracy rate. Table 12 shows the confusion matrix for training data based on the random forest algorithm.

Table 12: Random Forest Algorithm Confusion matrix for training data.

Prediction	Actual	
	FALSE	TRUE
FALSE	105590	27959
TRUE	183	268

From Table 12 in the diagonal, the data indicate that 105590 observations were FALSE, and the model predicts them to be FALSE. In other hands, there were 268 TRUE observations and the model predicts them to be TRUE. These are True Positive (TP) and True Negative (TN) observation values which imply correct prediction (classification). Moreover, and not in the diagonal, the data show that 27959 observations TRUE but the model predicts them to be FALSE. Also, 183 FALSE observations but the model predicts them to be TRUE. These are False Positive (FP) and False Negative (FN) observation values which are not in diagonal and entail incorrect prediction (misclassification). Thus, there is a difference between this matrix

table and one of the decision tree. The ability to overcome overtraining problem might have led to this. Table 13 depicts the confusion matrix for the test data.

Table 13: Random Forest Algorithm Confusion matrix for test data

Prediction	Actual	
	FALSE	TRUE
FALSE	51887	13871
TRUE	110	132

Table 13 in the diagonal, the data show that 51887 observations were FALSE, and the model predicts them to be FALSE. At the same time, there were 132 TRUE observations and the model predicts them to be TRUE. These are True Positive (TP) and True Negative (TN) observation values which mean correct prediction (classification). Furthermore, and not in the diagonal, the data pinpoint that 13871 TRUE observations but the model predicts them to be FALSE. In other hands, 110 FALSE observations but the model predicts them to be TRUE. These are False Positive (FP) and False Negative (FN) observation values which entail incorrect prediction (misclassification). The test data used to assess the predictive model performance accuracy rate. Hence, there is also a difference between this matrix table and one of the decision trees. The ability to overcome overtraining problem might have led to this. Table 14 is the confusion matrix for the training data used to build the naive Bayes predictive model.

Table 14: Naive Bayes Algorithm Confusion matrix for training data

Prediction	Actual	
	FALSE	TRUE
FALSE	105630	28183
TRUE	143	44

From Table 14 in the diagonal, the data indicate that 105630 observations were FALSE, and the model predicts them to be FALSE. In other hands, there were 44 TRUE observations and the model predicts them to be TRUE. These are True Positive (TP) and True Negative (TN) observation values which imply correct prediction (classification). Moreover, and not in the diagonal, the data show that 28183 observations TRUE but the model predicts them to be FALSE. Also, 143 FALSE observations but the model predicts them to be TRUE. These are False Positive (FP) and False Negative (FN) observation values which are not in diagonal and entail incorrect prediction (misclassification). The values in the confusion matrix are different

from the other two algorithms. Therefore, the randomness of the given data might have led to this since Naive Bayes fits in the normally distributed data. Table 15 shows the confusion matrix for the test data used to assess the performance of the naive Bayes predictive model.

Table 15: Naive Bayes Algorithm Confusion matrix for test data

Prediction	Actual	
	FALSE	TRUE
FALSE	51923	13980
TRUE	74	23

Table 15 in the diagonal, the data show that 51923 observations were FALSE, and the model predicts them to be FALSE. At the same time, there were 23 TRUE observations and the model predicts them to be TRUE. These are True Positive (TP) and True Negative (TN) observation values which mean correct prediction (classification). Furthermore, and not in the diagonal, the data pinpoint that 13980 TRUE observations but the model predicts them to be FALSE. In other hands, 74 FALSE observations but the model predicts them to be TRUE. These are False Positive (FP) and False Negative (FN) observation values which entail incorrect prediction (misclassification). The test data used to assess the predictive model performance accuracy rate. Hence, there is also a difference between this matrix table and the other two algorithms. The randomness of the given data might have led to this since Naive Bayes fits in the normally distributed data.

- Speed

This measurement factor reflected the amount of time in seconds used for each algorithm in generating the predictive model. According to Sharda et al. (2014), the predictive model built with short computational time implies the best predictive model. However, based on the aim of the thesis, this measure is not much considered as only used for comparing their speed in building models.

- Interpretability

This measure used to show how easy the model interprets the findings. It was also for identifying the level of understanding of the models generated. Following this, the predictive model specifically decision tree and random forest provided insight into the factors that influence rats' TB detection performance. However, the rules generated from the decision tree make it easier to understand and interpret the influencing factors for detection performance. Moreover, the random forest identified the variable importance with decrease mean Gini.

Besides, as naive Bayes does not measure the weights of each variable manipulated, its model is not intuitive compared to the rest. Therefore, these measures helped to find the best predictive model that fits the given data and aim of the thesis.

3.1.6 Deployment

It was the last phase used to organize and present the knowledge gained to the end-user for real application. The knowledge obtained is explicitly aimed at helping users to predict rats' factors that influence TB detection performance and the classes of new data instances (where the class label is unknown). As a result, visualization techniques such as histogram were used to assess the relationship between rat independent variable (Age) and the dependent variable (Performance). Moreover, the variable importance plot used for proper interpretation and ease understanding of the knowledge gained. Furthermore, the variable correlation plot used to show the extent to which each variable depends on one another.

3.2 Ethical issues and considerations

One of the data access requirements was to apply for the dataset and confirm for the data privacy referred to Appendix E. Considering Appendix F and G, the main and co-supervisors introduced me to the Data Provider as the student of Uppsala University and I would be writing my MSc thesis under their supervision during the spring semester 2019. Afterward, both the researcher (student) and the provider signed the Data Transfer Agreement (DTA) as the contract for the requested data at the first step. Moreover, with the referenced Appendix H, the remaining step was for the Medical Research Coordinating Committee (MRCC) to certify the Data Transfer Agreement for the approval to use the data. After this approval to use data, I made great efforts to protect its privacy to the extent that I used these data for this MSc study only. The thesis abandoned informed consent since there was no need for direct contact with participants such as TB patients.

4 Results and Analysis

This chapter presents and describes the results obtained from the data analysis. Firstly, a brief description of the structure of the data used for analysis. Secondly, the results exploratory analysis according to the research questions. The chapter concludes with a classification accuracy comparative analysis of each predictive model generated from the given data.

4.1 Structure of Data for Analysis

The analysed data consisted of 7 variables and 200,000 observations. However, in the analysis, this data was divided into two parts that are 67% of the data used for training to build predictive models and 33% used for test to assess the predictive models. The data also comprised of 4 female rats. Considering Table 5, the median of the numeric variables was 3.71 years, 866.8g, 12:00 hours, and 10.00 minutes respectively. Moreover, the data had 81 DOTS centers. However, Mwananyamala had many samples of 10.5% as shown in Table 7. Furthermore, Table 8 shows that 157,686 were FALSE and 42,314 were TRUE. Table 16 shows a summary of the variables used for building the classification predictive models.

Table 16: Description of the variables used to build predictive models

Variable	Description	Data type	Values
DOTS	Name of DOTS center	Factor	Mwananyamala, Mbagala Rangi 3, etc
Rat	Name of rat	Factor	Sofia, Happy, Catia, Mkuta
Age	Age of rat	Numeric	0.79, 2.04, 3.22, etc
Av_Weight_Per_Year	The average weight of rat per year	Numeric	846.35, 866.80,etc
Session_Start_Time	Time of day when detection session started in 24 hours	Numeric	12:05,13:34,14:00,etc
Session_Completion_Time	Differences in minutes between session start time and session end time	Numeric	1,2,3,etc
Performance	Performance of rat during the session	Factor	TRUE, FALSE

4.2 Results exploratory analysis

The data mining process aimed to elicit knowledge from the given structured data and present it to the end-user for the real application. And as such, this process was managed by the classification technique and algorithms that helped to learn the relationship between the patterns. However, the classification technique used three algorithms which are decision tree, random forest, and naïve Bayes to build the predictive models. Thus, this section presents results and analysis based on the formulated research questions as follows:

- *Can the factors that influence TB detection rats' performance be predicted using a classification technique?*

The classification technique was used to build predictive models that predicted the class for each rat and the factors that influence TB detection performance. However, this technique applied three algorithms to learn the relationship between variables. These variables are also called rats factors that may associate with TB detection performance. The independent variables (input) include Age, Av_Weight_Per_Year, Session_Start_Time, and Session_Completion_Time while the dependent variable (output) is Performance. Therefore, all three algorithms applied these variables separately to build predictive models on factors that influence rats' TB detection performance.

Starting with the decision tree algorithm, it generated a decision tree where the top node (root node) shows the most significant factor that influences TB detection performance. The ability of the algorithm to seek optimal splits in variable values has led to this. As a result, Figure 5 depicts the hierarchy of variables where the variable with a high correlation with the prediction split on first. The following other nodes show the remaining factors. Moreover, the leaf nodes indicate the class of every instance from the observations. Therefore, the decision tree algorithm results depicted the most significant factor and other factors that influence TB detection performance and revealed in Figure 5.

Figure 5: Decision tree with rats factors that influence TB detection performance.

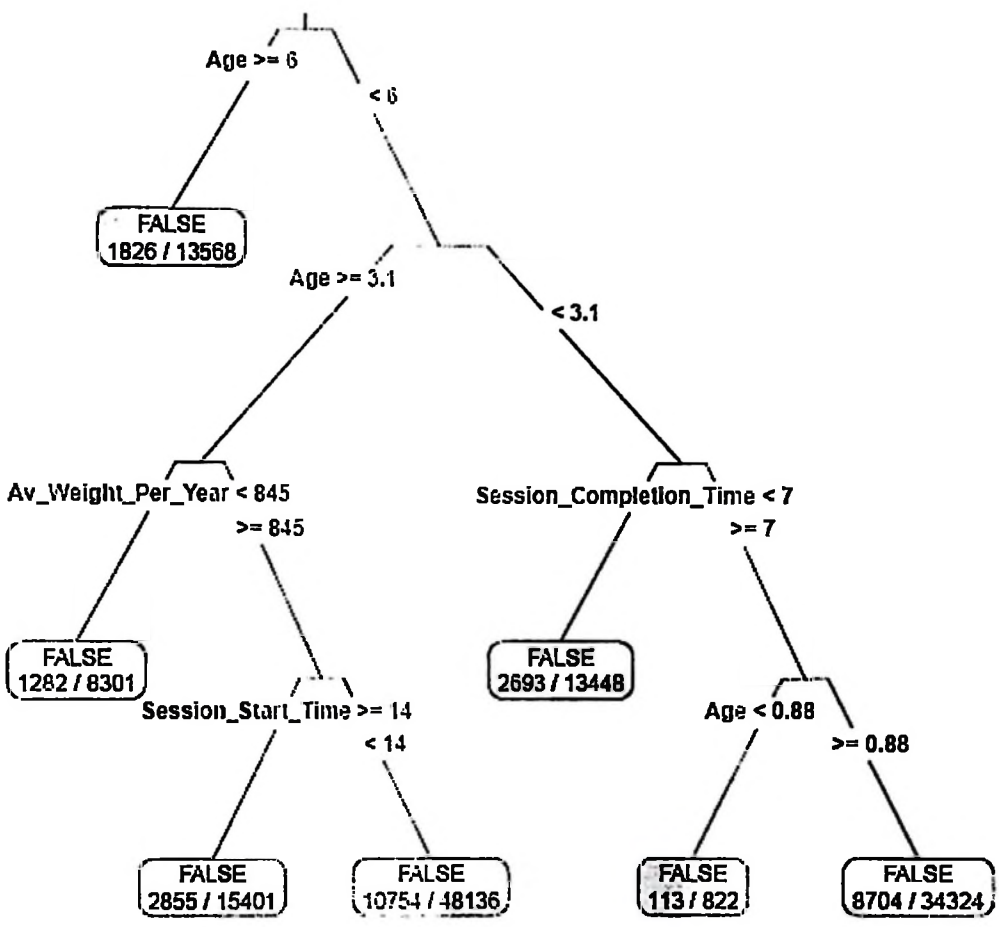


Figure 5 demonstrates the four variables recognized as the factors for rats TB detection performance. However, the variables are in the hierarchy since the information gain and entropy determined the split decision of variables in the class. Following this, the results depicted in Figure 5 show that the Age variable returned the highest information gain (the most homogenous values) and the lowest entropy when compared to the remaining variables. In this sense, the age of the rat is the most significant factor.

One of the potential advantages of the decision tree algorithm is that it generates rules that are easy to interpret and understand. These rules are the result of the IF-THEN statements from the root node to the leaf nodes based on the number of observations. The two parts (IF-THEN) imply the condition (s) on the value of predictor variables and the prediction (Performance decision) respectively. And as such, the general formula of the rules is IF

condition THEN conclusion. Thus, Table 17 manifests the rules generated from the decision tree in Figure 5.

Table 17: Classification rules generated from decision tree algorithm

Rule Number	Rule	Performance Decision		
		TRUE	FALSE	Number of observations in %
1	IF Age >= 6 ⇒	0.13	0.87	10%
2	IF Age < 0.88 & Session_Completion_Time >= 7 ⇒	0.14	0.86	1%
3	IF Age is 3.1 to 6 & Av_Weight_Per_Year < 845 ⇒	0.15	0.85	6%
4	IF Age is 3.1 to 6 & Av_Weight_Per_Year >= 845 & Session_Start_Time >= 14 ⇒	0.19	0.81	11%
5	IF Age < 3.1 & Session_Completion_Time < 7 ⇒	0.20	0.80	10%
6	IF Age is 3.1 to 6 & Av_Weight_Per_Year >= 845 & Session_Start_Time < 14 ⇒	0.22	0.78	36%
7	IF Age is 0.88 to 3.1 & Session_Completion_Time >= 7 ⇒	0.25	0.75	26%

The first rule says that older rats had a performance chance of 0.13, TRUE and 0.87, FALSE. The rats with ages greater or equal to 6 years detected fewer observations since the total percentage of the number of observations was 100% but 10% detected by rats in this range of ages. In other words, rats out of this range of ages classified in the remaining 80%. Moreover, considering the second rule which says that rats with the age of fewer than 0.88 years, and at least 7 minutes as the session completion time had a performance chance of 0.14 TRUE and 0.76 FALSE. However, 1% of the observations were classified in this rule. Following these two rules, older and less young rats portrayed the low performance.

The sixth rule has 36% of the detected observations than others. In other words, rats with ages of 3.1 to 6 years, at least 845g of the average weight per year, and the session start time before 14:00 hours had a detection performance chance of 0.22 TRUE, and 0.78, FALSE. This rule is consistent with the fourth one except for the session start time. Since the sixth rule had many observations than the fourth, the session starts time before 14:00 hours is the most performed one.

Furthermore, the fifth rule has 10% detected observations and states that rats with ages of 3.1 years and session completion time of fewer than 7 minutes had a performance chance of

0.20 TRUE and 0.80 FALSE. In other words, rats which detected observations at this range of time were potential. When comparing this rule with the second one, rats with a session completion time of fewer than 7 minutes depicted potentiality in detection since this rule had many observations compared to the second one. Therefore, the results pinpointed in Table 17 from the generated rules manifest that rats with ages of 3.1 to 6 years, at least 845g of the average weight per year, the session start time before 14:00 hours, and fewer than 7 minutes as the session completion time performed well.

Following this, the generated rules enhanced interpretation and understanding of the decision tree reported in Figure 5 to discover the knowledge. And as such, they ease predictions of the factors that influence rats TB detection performance. At this point, it is true that the classification technique predicted the influencing factors using a decision tree algorithm. Not only that but also the random forest identified factors that influence rats TB detection performance using the same manipulated variables. The following part presents and describes the predicted factors (importance of the variables) based on the random forest algorithm.

- *If yes, to what extent do different factors affect rats TB detection performance?*

It is of importance to understand the extent to which each factor contributed to the prediction. The random forest algorithm built the predictive model and pinpointed the variable importance. Random forest variable importance function helped to output the predictor variables that are important in predicting the outcome based on the Mean Decrease in Gini (impurity). Mean Decrease in Gini is the average (mean) of a variable total decrease in the likelihood of incorrect classification of a new instance of a random variable from the data set. Thus, Table 18 shows the variables used in building the random forest model and their MeanDecreaseGini.

Table 18: Importance of variables generated by Random Forest algorithm.

Factor	MeanDecreaseGini
Age	1791.9167
Av_Weight_Per_Year	233.5753
Session_Start_Time	471.6647
Session_Completion_Time	1472.5424

From Table 18, a higher (1791.9167) and lower (233.5753) Mean Decrease in Gini portrays greater and less variable importance respectively. In other words, Age and Av_Weight_Per_Year are the most and least significant factors. However, for easy interpretation and visualization of these results, the variable importance function of the random

forest algorithm sorted and displayed the variables in a plot as reported in Figure 6 based on the prediction importance. And as such, the variable that contributed much in the prediction is at the top part of the plot with the highest Mean Decrease in Gini values, followed by the variables with less importance. Therefore, Figure 6 reported the predicted factors (variable importance) based on Mean Decrease in Gini using random forest algorithm.

Figure 6: Variable importance generated by Random Forest algorithm.

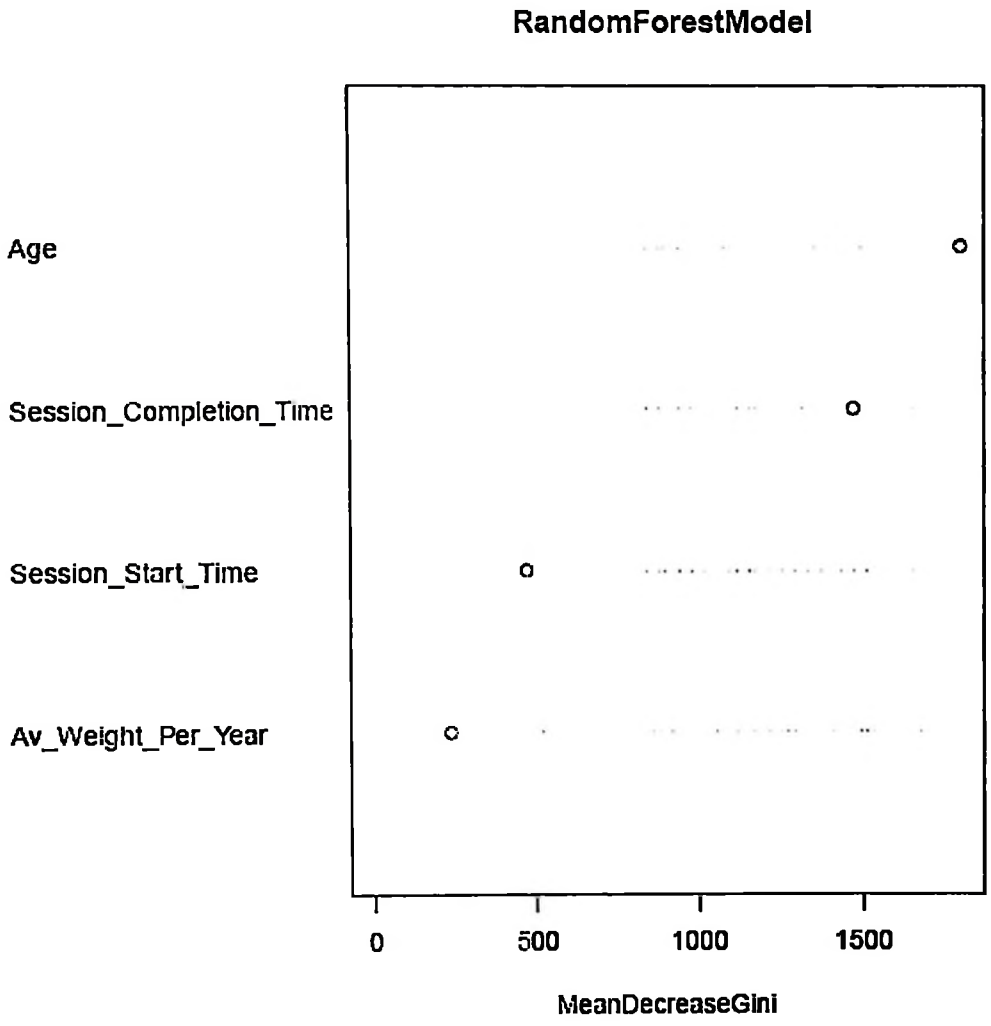


Figure 6 shows clearly that the random forest algorithm has predicted the significant factors with their mean prediction importance. However, it differs from the decision tree algorithm which has not displayed the weight of every manipulated variable. Thus, the results show that Age is the most significant factor since it has the highest mean decrease in Gini. In this sense,

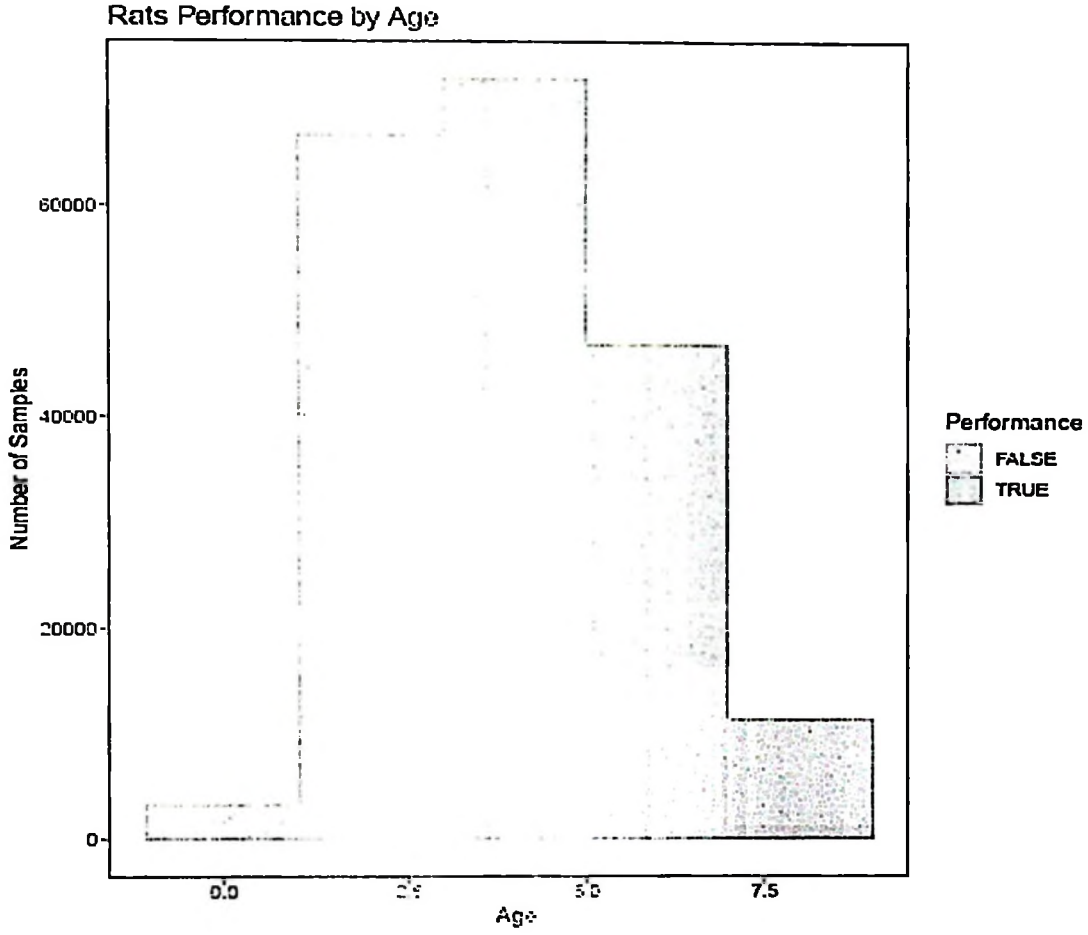
it is true that both decision tree and random forest have shown Age as the most significant factor.

Following this, the random forest algorithm and decision tree algorithm have predicted the factors that influence rats TB detection performance by using the classification technique. However, the naive Bayes algorithm was used to create the model and compare their classification accuracy. In other words, not for predicting the factors since it measures the probabilities of the variables and not their weights. The following part describes the range of age of the rats predicted having better performance.

- *If age is one of the significant factors, which one provides peak performance and why?*

From the given data and the aim of this thesis, the rats' performance implies their ability to detect a sample with either TB, TRUE (Sensitivity) or without TB, FALSE (Specificity). Table 5 manifests that the youngest and oldest rats had the ages of 0.79 and 7.95 years respectively with the median age of 3.71 years. And as such, rats with ages below and above the median refer to younger and older rats respectively. Since the given data had many numbers of observation with FALSE values than TRUE values as shown in Table 8, the rats' high performance in these data had a FALSE value. In this conception, rats' performance depended on the number of observations accomplished. Therefore, Figure 7 depicts the relationship between Age and Performance.

Figure 7: Rats performance by Age based on the number of detected samples



The results in Figure 7 pinpoints that rats with a range of ages from 0.79 to 1.2 years, detected a few observations and had low performance. Similarly, rule 2 reported in Table 17 from the decision tree shows that rats with ages of less than 0.88 years detected 1% of the whole observations. Nevertheless, one could argue whether these samples were present during that session or due to their low detection performance, they managed to complete only those few samples. Since the maximum age of the rat was 7.9 years, but the results in Figure 7 depicts that rats with a range of ages between 3.1 and 5, years had good performance. Comparably to rule 6 from Table 17 with many detected observations of 36% than others which states that rats within the range of 3.1 to 6 were potential. It is presumed as rats grow and probably getting qualified with the detection tasks their detection performance increases. Following this, the youngest and oldest rats revealed low performance.

Despite the high detection performance in the mentioned range of ages, the given data consisted of many values of FALSE than TRUE. Thus, rats out of this range of years portrayed low performance. In this conception, rats' performance decreases in youngest rats and when rats ages are getting either above 5 or 6 years. In other words, the rats' performance decreases when either they are too young or when their ages increase. Since the classification technique built three predictive models for predicting the factors that influence rats TB detection performance, it is of importance to compare their classification accuracy and find the best predictive model for the given data.

4.3 Comparison Analysis of Predictive Model Performance

It is useful to measure and compare the predictive model performance of any supervised learning algorithm for two reasons. First, it supports to estimate the future predictive accuracy of the predictive model and assure its ability to produce desired prediction results to the problem. Second, it can help to select the best classification model among the many generated from a given data. The predictive models' comparison often involves different metrics and evaluation criteria to assess their classification performance. These metrics include accuracy, sensitivity, specificity, and precision.

However, this comparison analysis used the accuracy metric to assess the classification accuracy of the generated predictive models. The accuracy is the metric which measures predictions that a model predicted correctly over the total number of predictions based on the testing data. In other words, it measures the extent to which the predictive model makes the correct prediction on the test data. It is applied most in classification problems where the output is of two or more types of classes (multi-class classification problems) (Sharda et al., 2014). In two-class classification problems, a confusion matrix defines the accuracy of the model based on the test data. The confusion matrix is a table that depends on its metrics such as accuracy, confidence interval (CI), sensitivity, and specificity to measure the classification accuracy of the predictive model (Chaurasia & Pal, 2013). Consider Table 18 which shows the description of the confusion matrix.

Table 19: Confusion Matrix for Multi-Class Classification Problem. (Sharda et al, 2015)

Prediction	Actual	
	Positive class	Negative class
Positive class	TRUE Positive (TP)	FALSE Positive (FP)
Negative class	FALSE Negative (FN)	TRUE Negative (TN)

Table 19 shows the confusion matrix with prediction and actual dimensions where in every dimension there is positive and negative classes. The right dimension prediction focused on the model while the actual dealt with the real data. Since the confusion matrix by itself is rather simple to understand, its related terminology may confuse. Based on the confusion matrix, TP and TN imply the number of positive and negative observations that are correctly classified. Furthermore, FP and FN reflect the miss-classified negative and positive which are incorrectly classified. Hence, the rating of the predictive model performance was measured based on the accuracy (acc) and error rate (err) values of the confusion matrix. Since the predictive models learned to classify the rats TB detection performance into either TRUE or FALSE, the positive class was a FALSE value since it had many observations. Therefore, this formula measured the predictive accuracy percentage for the positive class.

$$\text{Accuracy (acc)} = \frac{TP + TN}{TP + FP + TN + FN}$$

Moreover, the error rate (err) implies the fraction of the sum of FALSE positives and FALSE negatives and the sum of the total number of all the predictions made.

$$\text{Error rate (err)} = \frac{FP + FN}{TP + FP + TN + FN}$$

For achieving better accuracy in the generated predictive models, this analysis applied various data mining algorithms. These include the decision tree, random forest, and naive Bayes. The predictive models aimed to predict the factors that influence rats' TB detection performance and assign a class of every rat. Therefore, it is of importance to check their classification performance and usefulness of each algorithm applied. Thus, Table 20 shows a summary of classification performance based on the three applied algorithms.

Table 20: Comparison of predictive models' classification accuracy

Evaluation criteria	Predictive model		
	Decision tree	Random forest	Naive Bayes
Accuracy (%)	78.78%	78.82%	78.71%
Error rate (%)	21.22%	21.18%	21.29%
Correctly classified observations	51997	52019	51946
Incorrectly classified observations	14003	13981	14054
Speed to build a model (in sec)	3	154	1

Table 20 presents the comparison results of the predictive models' performance obtained after assessing their prediction accuracy of the same test data. It aimed to find the best model based on their classification accuracy. Since the predictive accuracy was the metric for predictive model performance, Table 20 depicts different evaluation criteria. These include the timing of building the model, correctly and incorrectly classified observations, and error rate.

Following this, Table 20 clearly shows that the overall performance results are closely related to all applied algorithms. In other words, the predictive models' performance has a very slightly accuracy rate and error rate differences. Since all the algorithms have the accuracy of greater than 70%, these results indicate moderate accuracy. As a result, it is the accepted accuracy in many cases including the medical field.

The random forest has the highest accuracy (78.82%) and lowest error rate (21.18%) followed by the decision tree (78.78%) where Naive Bayes (78.71%) is the least performance predictive model. The random forest and the decision tree algorithms have closely related accuracy. When comparing the accuracy of their predictive models based on training and test data, they both portray insignificant differences. In other words, the prediction accuracy of the training data is a bit higher than the prediction accuracy of testing data.

The naive Bayes shows the lowest accuracy and high error rate performance of 21.29% compared to the two algorithms. The nature of the given data might have led this since this algorithm fits well in normally distributed data. Moreover, this algorithm prevented the overfitting problem without removing values with less importance (pruning). As a result, the difference in prediction accuracy between training data and test data is also insignificant. However, naive Bayes assumed that the variables were independent which is incorrect as there were dependencies among the variables.

Based on the timing of building the model, the naive Bayes computation process used a short execution time and followed by the decision tree. The naive Bayes has a high score in

execution time because of its linear scaling and ability to process large dataset. Moreover, this algorithm assumed that the variables were independent for members of a given class and allow simplification in computing the likelihood. The decision tree and random forest used more time of execution when finding the best binary split from the random sampling of observations to create an assemblage of independent classification trees.

Additionally, data overlapping, and the random nature of the modeling algorithms presumed to affect the overall performance of the three predictive models. Following this, one can argue that the random sampling of observations could influence many FALSE values. As a result, it could prevent from yielding a desired estimate of the predictive model accuracy despite the larger random observations used. However, the given data had many FALSE values than TRUE values. Moreover, the independent variables were not highly correlated as shown in Figure 3 and since they were few could not affect the accuracy of the predictive model. Therefore, the classification accuracy increases as desired variables increases.

Based on the predictive model performance with accuracy metric, random forest and decision tree models fit the given data since they have predicted the factors that influence rats TB detection performance with the highest accuracies. Nevertheless, other metrics such as the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) would be useful to compare their performance. Despite the given data produced predictive model accuracy for every classification algorithm applied, the accuracies difference is too small or insignificant. On the other hand, when considering the predictive model accuracy and speed the results show that decision tree is the best predictive model since it has less building time (3 seconds) compared to the random forest (154 seconds) as reported in Table 20.

5 Discussion

This chapter discusses the results obtained from the data analysis. Initially, it presents the characteristics of the data, including any effects they might have on the results. Finally, it describes the strength of the results and the analysis based on the methodology and methods used.

5.1 Characteristics of Data

Data understanding phase stated that data used in this thesis (Rats dataset) resulted from the integration of the three given separated excel files with different worksheets. The given data consisted of many observations that may influence the desired results. Moreover, it contained many variables which helped to create other useful variables to answer the formulated research questions. However, it is vital to mention that the data given related only to four rats. In other words, the sample size for characterizing a TB rat is therefore only four. The number of rats given was the ones found with the requested data and was expected to address the aim of the thesis.

However, as it is a small sample size is presumed to have a slight possibility of affecting the results since the aim of the thesis was to find the influencing factors and not comparing the performance of every rat which would require large sample sizes. Based on the data preparation phase, this thesis aimed to use rat data from 2014 to 2018 years. This range of years used purposely to find the influence of the age of the rat in detection performance. Following this, it is useful to describe some of the variables used in the analysis such as Age, Gender, Av-Weight_Per_Year, and Performance and to check their implications on the results.

The analysed data consisted of only four named rats with different years of birth from 2011 to 2013. The ages of the youngest and oldest rat are 5.8 and 7.95 years respectively. This year calculation is only from 2011 to 2018 and not the current year 2019. The age differences were of importance to find the relationship between age and the detection performance. Following this, Mkuta started working with 0.79 years compared to others. However, there is no effect on the results since Mkuta is the youngest rat and detected a few numbers of observations.

Moreover, there was no gender equality in the given data since all rats were female. In this conception, this thesis dealt with only given four female rats. However, for the future, it is advantageous to analyze and investigate data with both male and female rats to understand which gender influences detection performance. Additionally, the analysis used rats average

weight per year to avoid bias within variables since most of the rats' daily detection performance data had no corresponding weights of the rats in each detection session day. Hence, the weight identified with the influence on detection performance is the rats average weight per year. Therefore, data with a consistent weight of rats in every detection task could increase results reliability and generalization.

Based on dependent variable Performance, data consisted of many FALSE values than TRUE values. Since it was the target class for classification, it is presumed to have an impact on the results. Thus, when one value has many samples than the other, its performance is also higher. It was advantageous if the data would have an estimation of about an equal number of values of the observations in the detection performance class. And as such, it would reduce the suspicion that the results might rely on one group of the data and limits generalization. Furthermore, Figure 3 shows that the given data had a weak positive and negative correlation between dependent and independent variables. However, many strongly correlated variables could increase accuracy and generalization of the results.

5.2 Factors Influencing Rats TB Detection Performance

The thesis aimed to predict the factors that influence rats' TB detection performance using data mining techniques. The classification technique was used to predict the influencing factors by building models that assigned a label to every manipulated instance (rats). Moreover, three algorithms include decision tree, random forest, and naïve Bayes were used to support this technique to build predictive models. Closely related to the studies conducted by Asha et al. (2011), Ameri et al. (2014), Suresh & Arulanandam (2018) which used the decision tree algorithm to learn past medical data and build a predictive model. Not only that but also Asha et al. (2011) and Ayas & Ekinci (2014) used random forest algorithm to learn patterns from the past data to discover and extract hidden interesting information by building the predictive model that categorized tuberculosis disease status.

However, the naive Bayes algorithm was used to build the predictive model that classify the instances and compare their classification accuracy. In other words, not for predicting the factors since it measured the probabilities of the variables and not their weights. Similarly, to the study of Maniya et al. (2011) which used a naive Bayes algorithm to classify patients affected by tuberculosis into two classes which are least probable and most probable. Therefore, the results show that the classification technique based on decision tree and random forest algorithms have predicted the factors that influence rats TB detection performance.

These factors are Age, Session_Completion_Time, Session_Start_Time, and Av_Weight_Per_Year. And as such, Age is the most significant factor. The information gain of each computed factor in the class decided this order of significance or variable importance.

The results depict the strength of the Age factor in the detection performance. Primarily, in the decision tree reported by Figure 5, the Age was split first due to the highest information gain ratio. As a result, it has appeared in all generated rules in Table 17. Contrary to the other variables that are shown only once in the generated rules. Moreover, in variable importance of random forest depicted in Table 18 and Figure 6, the decrease means Gini of Age was higher than the other variables. Therefore, the results manifested that rats between ages of 3.1 to 6 years positively affected the performance. On other hands, as rats grow to this range of years and getting qualified with detection tasks, their performance also increases. However, it may limit the generalization of the results since it referred to female rats. The study of Brushfield et al. (2008) proposes that detection performance may depend on rats' characteristics such as age. Nevertheless, successful training and growth progress might have led to this.

Furthermore, the results show that older rats portrayed a low detection performance. And as such, the olfactory deficit might have caused this since detection performance depends on the rats' olfactory sensitivity (Kraemer & Apfelbach, 2004). Similarly, the study of Brushfield et al. (2008) trained rats with 6 and 24-monthes to discriminate odors. As a result, the older rats needed many trials than the younger rats. Moreover, the study of Poling et al. (2011) agree with this result that a highly developed sense of smell increases their potential for use to detect TB bacteria in sputum samples.

Two groups of rats with a median age of 2.4 and 3.8. years respectively were involved in the study of Ellis et al. (2017). However, 3.8 years performed well and that rats with the age of fewer than 2.5 years experienced a low detection performance. Following this, younger and less old rats are good performers because older rats have low olfactory sensitivity and deficits in olfactory memory occurred in aged rats (Kraemer & Apfelbach, 2004).

Moreover, the results provide new insight into the relationship between time differences when the rat starts and ends detection tasks (Session_Completion_Time). From the generated rules, three rules out of seven mentioned this variable and, is the second in variable importance. Since these rats have a high-speed of detecting 100 samples in 20 minutes, good performers were the rats which completed the number of observations with less than 7 minutes. The study conducted by Mgode et al. (2018) pinpoints that during the training, rats learn to pause for

about 3 seconds to the sample hole with TB bacteria and take 1 second to the sample hole without bacteria. Thus, they use a short and long time to TB and non-TB sample respectively. Despite one could argue that the data also influenced the results since it contained many values of FALSE and implied the detected samples without TB. On other hands, one could question that the TB samples delivered to APOPO from 2014 to 2018 years for the second screening diagnosis from different DOTS centers consisted of many negative TB samples. As a result, the predicted session completion time is less than 7 minutes. Hence, the presence of most of the non-TB samples in the session might have an influence.

Furthermore, the time of day (Session_Start_Time) when the rats start detection tasks influence their performance. The given data depict that rats performed detection tasks in all time of the day, which includes morning, afternoon, and evening. However, rats which carried out detection tasks before 14:00 hours portrayed better performance than the ones performed after this time of the day. In other words, detection tasks carried out in the morning and afternoon times performed well. However, since rats are active most of the time one can question that presumed during that time of the day is where trainers presented many numbers of samples than the other times. Similarly, with the studies conducted by Mgode et al. (2018) and Ellis et al. (2017) which presume that time of day might influence rats' detection performance. In this conception, there is a possibility for the increase in rats' detection performance if only the responsible ones consider this time of the day. Therefore, understanding this time of the day would be of benefit for the center and rats' trainers to enhance better performance.

Not only that but also, the results contribute a clearer understanding of the influence of average weight per year (Av_Weight_Per_Year) on detection performance. However, it is the least factor of importance with the least decrease mean in Gini. The absence of corresponding individual rat weight for analysis might have led this. The results demonstrate that rats with an average weight of greater or equal to 845g performed better. According to the study conducted by Beyene et al. (2012), the weight range of adult rats' females ranges from 1 to 1.5kg. Also, Table 5 shows that the greatest rats' weight of the given data was 1.05kg. Therefore, one can argue that young rats with at most 1.05 kg were the most performer. Nevertheless, presumed the reliability of the results could increase with the corresponding weight rather the applied average.

Therefore, these results should be considered by the APOPO center to utilize the usefulness of this technology. In other words, they must understand that the detection

performance depends on rats' factors and thus must maintain them for sustainability. In other hands, from the analysed data it is shown that the observations presented for the second screening were from 81 MOTS centers. However, Mwananyamala consisted of many observations. Following this, the population of the area might have led to the high transmission of the disease (Poling et al., 2011).

Finally, the results reveal that for the three different algorithms used, the classification accuracy was much more in the random forest (78.82%) than decision tree (78.78%) and naive Bayes (78.71%). However, the accuracies difference is too small or insignificant. On other hands, when considering the predictive model accuracy and modelling speed the decision tree is the best, since it had less building time (3 seconds) compared to the random forest (154 seconds). The nature of data and algorithms used might have caused this in the sense that random forest and decision tree algorithms fit in skewed data different from naive Bayes which do better in normally distributed data. As a result, the random forest and decision tree had a high classification accuracy than naive Bayes. Thus, the nature of the data often affects the classification accuracy (Sharda et al., 2014). Moreover, in the random forest, the ability to assembly several trees and make the final decision from several trees might influence this highest classification accuracy (Osha et al., 2011).

Despite the found results based on the dependent and independent variables given from the data, other factors presumed the influence on these results. These factors may include training procedures, trainers or recorders (data recording), experimental setup, and laboratory technicians (quality control). The study conducted by Reither et al. (2015) argue that since rats are trained based on the conditioning techniques which support to change their behavior such as learning to recognize sound during the training, it is useful to have the justifiable rules to avoid incorrect results. Likewise, Mgode et al. (2018) demonstrate that rats successful and consistent training procedures are most important in TB healthcare centers that apply rat as odor-detection technology. Following this, there is a possibility that rats from the given data succeeded in the training procedures and thus manifested better performance.

Moreover, observing precision in data recording during detection tasks is highly emphasized to avoid false results. Since rats trainers and recorders are the ones performing data recording, they should have skills in getting consistent records. In the study of Poling et al. (2011), there were two trainers of rats with the different clinical status of the sample. And as such, there was no agreement on data recording during that session and presumably affects the results recording. Furthermore, as mentioned above, detection training depends on

conditioning techniques such as a reward or punishment to change rats' behavior and since rats often perform detection tasks in cage holes and when a rat detects a positive sample is getting food from a syringe as a reward. Therefore, a well-organized experiment setup may facilitate rats to portray better performance.

Additionally, before presenting the sample in a cage for detection, a standard heat is applied into it to kill infectious microorganisms and enhance quality control. In this conception, quality control may determine the effectiveness of rats in detection performance. Therefore, one can argue that despite the outcome of the dependent and independent variables, the mentioned confounding variables might influence the results. Following this, rats' detection performance depends on the main and confounding factors.

6 Conclusion

This chapter provides the main conclusions of the thesis based on the aim and research questions. Initially, it describes its implications and limitations. Finally, it presents recommendations for future research in the area.

6.1 Implications of the Thesis

It is considered useful to apply data mining techniques to solve multi-classification healthcare problems. These techniques are of importance to discover valuable information from complex data produced by healthcare sectors to improve their operations and ease the decision-making process. This thesis has focused on the prediction of factors influencing rats TB detection performance using data mining techniques through understanding the relationship of the manipulated variables. Moreover, building predictive models for predicting the class for every new instance (rat).

Following this, the results indicate that the factors predicted for rats' TB performance include Age, Session_Completion_Time, Session_Start_Time, and Av_Weight_Per_Year. However, the age of the rat found as the most influencing factor. Nevertheless, the information gain and variable importance determined the order of these factors. Besides, the results pinpoint that rats with the following factors conditions were the best performers. These include rats with ages of 3.1 to 6 years, at least 845g of the average weight per year, less than 14:00 hours as the session start time, and less than 7 minutes as the session completion time. However, these results limit generalization since they refer to female rats. Therefore, the center can utilize rats with these factors for better detection performance.

On other hands, the random forest predictive model found as the most suitable model for predicting and assigning a class for every rat. And as such, it has the highest classification performance accuracy of 78.82%. Followed by the decision tree with 78.78% and naive Bayes is the last model with 78.71%. However, their accuracy differences are too small or insignificant. Therefore, when considering both the predictive model accuracy and speed the decision tree is the best since it had less building time (3 seconds) compared to the random forest (154 seconds).

Therefore, these results are valuable as a reference for several groups. Initially, rats' trainers and decision-makers are encouraged to consider the influencing factors to maintain their usefulness such as predicting the performance of the formerly and newly trained rats. Following this, they can prevent any risk related to involving rats with a low performance by

taking several actions. These actions can be taken care of and support TB detection factors to increase rats 'detection performance throughout their lifetime. Moreover, the result of this thesis is of importance to either TB specialists or any other medical specialists. Additionally, the results may help researchers in healthcare. Since this thesis implemented data mining techniques in a social setting by predicting factors that influence rats in detecting TB disease, it is also helpful to the academic society of Information System (IS).

6.2 Limitations of the Thesis

Despite the given data contained many observations this thesis encountered several limitations. First, the sample size was small with only four rats. And as such is presumed to have a slight possibility on affecting the results because the aim was to find the factors and not comparing the performance of every rat which could require large sample size. Second, the lack of the corresponding weight of rats in every detection session. As a result, the average weight per year was solely used and not the weight of rats which was more advantageous. The motives behind the usage of the average weight were to avoid bias in the variables. However, the average weight may limit the generalization of the results. Moreover, all rats are female as shown in Table 6. Nevertheless, for increasing the number of known factors that influence rats TB detection performance, it would also be interesting to discover the hidden patterns on the gender factor. Not only that but also the dependent variable consisted of many values of FALSE (78.8%) than TRUE (21.2%) as shown in Table 8 and presumed that the data were limited to a FALSE group of performance. Additionally, confounding factors such as training procedures, trainers or recorders (data recording), experimental setup, and laboratory technicians (quality control) might have an impact on the results. Lastly, it was useful to get a comprehensive understanding of data transfer ethical approval and considerations at the beginning of the thesis.

6.3 Future Research

Given the limited number of studies has been conducted at APOPO TB center on the prediction of factors influencing rats' TB detection performance using data mining techniques. Therefore, to maximize the effectiveness and efficiency of these results, several criteria for future research will have to be optimized. First, a dataset with large sample size and many desirable variables for rats TB detection performance is valuable to increase the number of known factors. Moreover, to enhance the generalization of the results, the dataset should include the weight of rats and not their averages. Not only that but also to predict significant sex differences, the dataset should balance gender distribution. Furthermore, the dependent variable should contain

classes with an approximately equal number of values. Besides, the applicability of other classification algorithms and data mining or machine learning tools such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression, Weka and Rapid Miner and select the best one. Since rats also detect land mines, it is of importance to predict their significant factors using data mining techniques.

7 References

Alvesson, M. & Sandberg, J. (2011). Generating Research Questions through Problematization, Academy of Management Review 36(2): 247–271.

Ameri, H., Alizadeh, S. & Hadizadeh, M. (2014) Assessing the Effects of Infertility Treatment Drugs Using Clustering Algorithms and Data Mining Techniques. Journal of Mazandaran University of Medical Sciences, 24, 26-35. (Persian)

Asha, T., Natarajan,S., Murthy, K.N.B., (2011). Effective Classification Algorithms to Predict the Accuracy of Tuberculosis-A Machine Learning Approach.

Beyene, N., Mahoney, A., Coxi, C., Weetjens, B., Makingi, G., Mgode, G, et al. (2012). APOPO's tuberculosis research agenda: achievements, challenges and prospects. *Tanzania Journal of Health Research.* doi: 10.4314/thrb. v14i2.5

Boell, S.K & Cecez-Kecmanovic,D. (2015). On being 'systematic' in literature reviews in IS. Journal of Information Technology, 30, 161–173, *JIT Palgrave Macmillan*

Brushfield, A., Luu T., Callahan, B., & Gilbert, P. (2008). A comparison of discrimination and reversal learning for olfactory and visual stimuli in aged rats. Behav Neurosci, 122(1):54–62.

Chaurasia, V and Pal, S (2013). Data Mining Approach to Detect Heart Disease. International Journal of Advanced Computer Science and Information Technology Volume 2, Issue 4, ISSN: 2296-1739.

Ellis, H., Mulder, C., Valverde, E., Poling, A., & Edward, T. (2017). Reproducibility of African giant pouched rats detecting Mycobacterium tuberculosis. BMC Infectious Diseases, 17:298. doi 10.1186/s12879-017-2347-3

Finfgeld-Connett, D. & Johnson, E.D. (2013). Literature Search Strategies for Conducting Knowledge-Building and Theory-Generating Qualitative Systematic Reviews, Journal of Advanced Nursing 69(1): 194–204.

Green, D.M. and Swets, J.A. (1966) Signal Detection Theory and Psychophysics. Wiley, New York.

Kraemer, S & Apfelbach, R. (2014). Olfactory sensitivity, learning and cognition in young adult and aged male Wistar rats. *Physiol Behav.*

Kolk, A. H. J., van Berkel, J. J. B. N., Claassens, M. M., Walters, E., Kuijper, S., Dallinga, J.W., & van Schooten, F.J. (2012). Breath analysis as a potential diagnostic tool for tuberculosis. *The International Journal of Tuberculosis and Lung Disease.* doi: 10.5588/ijtld.11.0576

Mahoney A, Edwards TL, Weetjens BJ, Cox C, Beyene N, Jubitana, M, et al. (2013). Giant African pouched rats (Cricetomys Gambianus) as detectors of Tuberculosis in human sputum: Two operational improvements. The Psychological Record, 63, 583–594.

Malik, M, Abdallah, S, Ala'raj, M. (2016). Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review.

Maniya, H., Hasan, M.I. & Patel, P.K. (2011). Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis. *International Journal of Computer Applications (IJCA).*

Mgode, G.F., Cox, C.L., Mwinzi, S., & Mulder, C. (2018). Pediatric tuberculosis detection using trained African giant pouched rats. Pediatric Research, 84(1). doi:10.1038/pr.2018.30

Morrell, K. (2008). The Narrative of 'Evidence Based' Management: A polemic, Journal of Management Studies 45(3): 613–635

Mulder, C., Mgode, G.F. & Reid, S.E. (2017). Tuberculosis diagnostic technology: an African solution ... think rats. African Journal of Laboratory Medicine, Vol.6 No.2

Nagabhushanam, D., Naresh, N., Raghunath, A. and Praveen Kumar, K. (2013) Prediction of Tuberculosis Using Data Mining Techniques on Indian Patient's Data. *IJCST, 4, 262-265.*

Oates, B.J., Edwards, H.M. and Wainwright, D.W. (2012). A Model-Driven Method for the Systematic Literature Review of Qualitative Empirical Research, in ICIS 2012 Proceedings, Shanghai, China, 1–18.

Okoli, C. and Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research, Sprouts: Working Papers on Information Systems 10(26).

Poling, A., Weetjens, B., Cox, C., Beyene, N., Durgin, A., & Mahoney, A. (2011). Tuberculosis Detection by Giant African Pouched Rats. The Behavior Analyst, 34(1), 47–54.

Poling, A., Valverde, E., Peyene, N., Mulder, C., Cox, C., & Mgode, G.F. (2016). Active Tuberculosis detection by pouched rats in 2014: More than 2,000 new patients found in two countries. Journal of Applied Behavior Analysis

PrasannaDesikan, Kuo-Wei Hsu, Srivastava.J. (2011). Data Mining for Healthcare Management. SIAM International Conference on Data Mining.

Reither, K., Jugheli, L., Gloss, T.R., Sasamalo, M., Mhimbira, F.A., Weetjens, B.J, et al. (2015). Evaluation of Giant African Pouched Rats for Detection of Pulmonary Tuberculosis in Patients from a High-Endemic Setting.

Samuel, E.W & Snodgrass, M. (2015). Signal Detection Theory. The Oxford Handbook of Philosophy of Perception

Sharda, Delen & Turban (2014). Business Intelligence and Analytics (Tenth edition).

Suresh, N. & Arulanandam, N. (2018). A Mining Approach for Detection and Classification Techniques of Tuberculosis Diseases.

Torraco, R.J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. University of Nebraska–Lincoln.Human Resource Development Review Vol. 4, No. 3 September 2005 356-367 DOI: 10.1177/1534484305278283

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, 26 (2), xiii-xxiii.

Weetjens, B. J., Mgode, G. F., Machang'u, R. S., Kazwala, R., Mfinanga, G., Lwilla, F., et al. (2009). African pouched rats for the detection of pulmonary tuberculosis in sputum samples. International Journal of Tuberculosis and Lung Disease, 13, 737–743.

World Health Organization. END TB strategy. Geneva, Switzerland: WHO; 2014.

World Health Organization. Global tuberculosis report 2018. New York, United States of America: WHO; 2018.

8 Appendices

Appendix A.

RAT_WEIGHT Dataset Example

Number	ID_RAT	RAT_NAME	WEIGHT_DATE	WEIGHT
1	20	[illegible]	11/23/2013	900
2	34	[illegible]	7/3/2014	302
3	46	[illegible]	3/6/2012	820
4	11	[illegible]	12/30/2013	675
5	55	[illegible]	9/18/2015	786
6	20	[illegible]	2/29/2016	1002
7	34	[illegible]	5/19/2017	346
8	46	[illegible]	7/14/2014	298
9	11	[illegible]	10/23/2016	675
10	55	[illegible]	11/5/2014	230

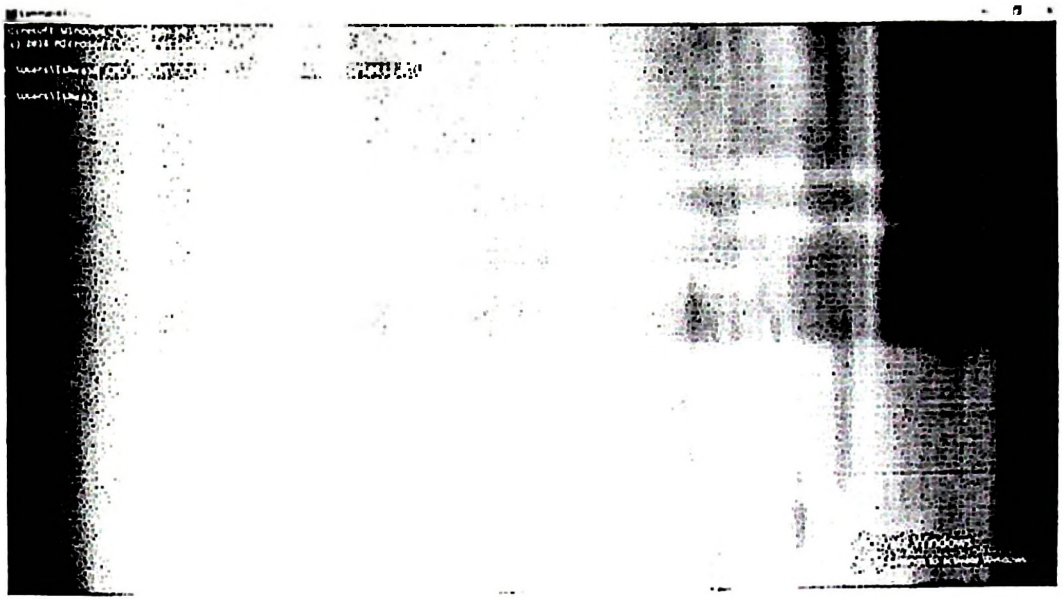
Appendix B.

Preprocessed Rats data[illegible] Example from DetectionRatsData Dataset

Number	DOTS	Rat[illegible]	Av_Weight_Per_Year	Session_Start_Time	Session_Completion_Time	Performance
1	Amana	Sc[illegible]	855	12[illegible]	12	FALSE
2	Ukonga	Mk[illegible]	902	13.[illegible]	5	TRUE
3	St.Gemma	Ha[illegible]	899	11.[illegible]	7	FALSE
4	Morogoro	Mk[illegible]	904	17.[illegible]	13	TRUE
5	Isanga	Sc[illegible]	1001	13.[illegible]	20	FALSE
6	Tumbi	Ha[illegible]	934	10.[illegible]	1	FALSE
8	Ilala	Ca[illegible]	345	10[illegible]	11	FALSE
9	Sabasaba	Sc[illegible]	837	11[illegible]	2	TRUE
10	Tandukaleli	Ha[illegible]	678	1[illegible]	7	FALSE
11	Mzinga	Mk[illegible]	954	1[illegible]	13	TRUE

Appendix C.

CMD BATCH helped to run the scripts (Scripts2.R) non-interactively on the command line of R and sent the output to another file (Scripts2.Rout) placed in the same location as Scripts2.R file. Meanwhile, located the Rplots in the working directory path.



Appendix I .

R scripts output [illegible]

```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "C:/Users/Ishva"
> setwd("C:/Users/Ishva/Documents/Revised rats data")
> #Creating the object rats
> rats<-read.csv("Rats.csv")
> #Calling the package dplyr for sampling
> library(dplyr)
> #Random sampling 200,000 observation
> rats<-sample_n(rats,200000)
> #Removing some columns in the data filter
> rats= select (rats, -c (CUSTOMER_AgeOrg,Start_Time,End_Time))
```

```
> #Displaying data in filter bar
> View(rats)
> #Displaying data structure
> str(rats)
'data.frame':    200000 obs. of  7 variables:
> #Displaying the summary of data
> summary(rats)
>
> #Data Transformation
> rats$Session_Start_Time= as.numeric(rats$Session_Start_Time)
> rats$Performance= as.factor(rats$Performance)
> rats$DOTS= as.factor(rats$DOTS)
> rats$Rat= as.factor(rats$Rat)
> rats$Age= as.numeric(rats$Age)
> rats$Av_Weight_Per_Year= as.numeric(rats$Av_Weight_Per_Year)
> rats$Session_Completion_Time= as.numeric(rats$Session_Completion_Time)
>
> #Data split
> training<-sample_frac(rats,.67)
> test<-sample_frac(rats,.33)
>
> # Calling the ggplot2 package
> library(ggplot2)
> #Plotting a histogram of rat age against performance
> ggplot(rats,aes(x = Age, fill = Performance)) + geom_histogram(binwidth = 2)+theme_bw()+ labs(y =
"Number of Samples", x = "Age",  title = "Rats Performance by Age")
>
> #Calling the packages for building a decision tree predictive model
> library(rpart)
```

```
> library(caret)
> library(rattle)
> library (RColorBrewer)
>
> #Decision tree modelling with the rpart package
> tree1<-
rpart(Performance~Age+Av_Weight_Per_Year+Session_Start_Time+Session_Completion_Time,data=
training, control=rpart.control(minicriterion=0.9,minsplit=15000, minbucket=2, cp=-0.4,maxdepth = 4))
> #Calling rpart plot package for plotting the decision tree
> library (rpart.plot)
Warning message:
package 'rpart.plot' was built under R version 3.5.3
> #Plotting the decision tree
> rpart.plot(tree1,extra = 3, type = 3,fallen.leaves = FALSE,tweak = 1)
>
> #Generating decision tree rules
> rpart.rules(tree1, cover = TRUE)
>
> #Decision tree model prediction accuracy with training data
> tree1.Performance.predTest< predict (tree1, training,type="class")
> #Creating a confusion matrix
> confusionMatrix (tree1.Performance.predTest,training$Performance)
>
> #Decision tree model prediction accuracy with test data
> tree1.Performance.predTest<-predict (tree1, test,type="class")
> #Creating a confusion matrix
> confusionMatrix(tree1.Performance.predTest,test$Performance)
>
> #Calling the package for building a random forest predictive model
```

```
>library(randomForest)
>#Creating a random forest model
>RandomForestModel<-
randomForest(Performance~Age+Av_Weight_Per_Year+Session_Start_Time+Session_Completion_Time,
data= training)
>#Plotting the random forest model
>plot (RandomForestModel)
>#Creating the random forest variables importance
>RandomForestModel$importance
>
>#Random forest model prediction accuracy with training data
>RandomForestModel.Performance.predicted<-predict (RandomForestModel,training)
>#Creating a confusion matrix
>confusionMatrix (RandomForestModel.Performance.predicted,training$Performance)
>
>#Random forest model prediction accuracy with test data
>RandomForestModel.Performance.predicted<-predict (RandomForestModel,test)
>#Creating a confusion matrix
>confusionMatrix(RandomForestModel.Performance.predicted,test$Performance)
>
>#Calling the packages for building the naïve Bayes model
>library(naivebayes)
>library(e1071)
>library(rminer)
>#Building the naïve Bayes model
>NaiveBayesModel<-
naiveBayes(Performance~Age+Av_Weight_Per_Year+Session_Start_Time+Session_Completion_Time,
data= training)
>#Calling the generated naïve Bayes model
>NaiveBayesModel
```

```
>
> #Naive Bayes model prediction accuracy with training data
> NaiveBayesModel.Performance.predicted<-predict (NaiveBayesModel,training)
> #Creating a confusion matrix
> confusionMatrix (NaiveBayesModel.Performance.predicted,training$Performance)
>
> #Naive Bayes model prediction accuracy with test data
> NaiveBayesModel.Performance.predicted<-predict (NaiveBayesModel,test)
> #Creating a confusion matrix
> confusionMatrix (NaiveBayesModel.Performance.predicted,test$Performance)
```

Appendix E.

Rats TB dataset application letter to Data Provider

Joan Jonathan Minyambo,
Ruckahergsgatan 46,
P.O Box 752 32.
Uppsala, Sweden.
21/1/2019.

Program manager tb tanzania.
APOPO TB Tanzania,
Sokoine University of Agriculture
P.O. Box 3078.
Morogoro.

Dear Sir,

RE: REQUEST FOR THE TUBERCULOSIS DETECTION RATS DATASET FOR MSC. RESEARCH USE.

Refer to the heading above.

I am a second-year student pursuing master program in Information Systems at Uppsala University Sweden. As part of my master studies, I am really interested to do my master thesis with APOPO TB centre using TB detection rats data. I would like to request for provision of dataset preferable ranging between 2014-2018. However, it is not necessary to separate these data based on years but as a collection in a single excel file since the study will not compare trends in years. The columns may consist seven variables which include age, sex, weight, clinic identity, bacterial count, time of day, and detection (target class). However, the rows may have numerical data and not exceeding 500 rows. The detection class column may consist of categorical data in rows (either low or high) which helps to show the class of every rat.

These data will be strictly used for this MSc study only. My supervisor who is a Professor in Artificial Intelligence will guide me throughout the study. Based on his research and ethics experience we will do everything we can do to protect its privacy. Also, hopefully the study will add scientific knowledge to the centre on factors that influence rats TB detection performance. With this letter, I have enclosed a concept note of the proposed master thesis, confirmation letter from my supervisor, and a draft of an excel sheet for capturing the requested data.

I hope my request shall meet your positive consideration.

Yours Sincerely,

Joan Jonathan Minyambo

Appendix F.

Introduction letter from main Supervisor to Data Provider

UPPSALA UNIVERSITET

2018-12-29

PROGRAM MANAGER TB TANZANIA,
APOPO TB RESEARCH AND TRAINING CENTRE.
P.O BOX 3078,
MOROGORO

To whom it may concern

I hereby confirm that Joan Jonathan Mnyambo will be writing her Msc thesis under my supervition during the spring semester 2019 and that we both will be committed to protecting the privacy of the data set she has requested from your institution.

Yours sincerely

Andreas Hamfelt
professor

Institutionen för Informatik och Media

Andreas Hamfelt
professor

Besöksadress:
Ekonomikum
Kyrkogårdsgatan 10
753 13 Uppsala

Postadress
Box 513
751 20 Uppsala

Telefon:
018 - 471 1037
073 - 425 0254

Telefax
018 - 471 71 43

Hemsida:
www.im.uu.se

E-post:
Andreas.Hamfelt@im.uu.se

Department of Informatics and Media

Andreas Hamfelt
Professor

Visiting address:
Kyrkogårdsgatan 10
Ekonomikum
SE-753 13 Uppsala

Postal address:
Box 513
SE-751 20 Uppsala
SWEDEN

Telephone:
+46 18 - 471 1037
+46 73 - 425 0254

Telefax:
+46 18 - 471 71 43

Web site:
www.im.uu.se

E-Mail:
Andreas.Hamfelt@im.uu.se

Appendix G.

Introduction letter from co-Supervisor to Data Provider

UPPSALA
UNIVERSITET

Institutionen för informatik och medie

Department of Informatics and Media

Box 513, 751 20 Uppsala

Besöksadress:

Ekonomikum 10

Telefon: 018 471 3321

Program Manager TB Tanzania
Ifakara TB Research and Training Centre
P.O. Box 3078
Morogoro
TANZANIA

To whom it may concern:

I hereby confirm that Joan Jonathan Mnyambe will be writing her MSc thesis under my supervision during the spring semester 2019 and that we both will be committed to protecting the privacy of the dataset she has requested from your institution.

Yours faithfully,

Dr David Johnson
Senior Lecturer
Department of Informatics and Media
Box 513
SE-751 20 Uppsala
SWEDEN
E-mail: david.johnson@im.uu.se

Appendix II.

Data Transfer Agreement (DTA) for ethical approval from Data Provider

NATIONAL INSTITUTE FOR MEDICAL RESEARCH

DATA TRANSFER AGREEMENT FOR RESEARCHERS/ORGANIZATIONS

(FOR RESEARCH USE ONLY)

THIS DATA TRANSFER AGREEMENT FOR Researchers/Organizations (here-in-after referred to as *the "Agreement"*) is made this 20. ..... Day of Feb ..... , 2019.

Between

SUA-APOPO PROJECT ..... of P O Box 3078 MOROGORO .....

(here-in-after referred to as the *"PROVIDER"*);

and

....of P.O.B. .....

(here-in-after referred to as a "party" or the *"RECIPIENT"*).

PROVIDER and RECIPIENT may each be referred to as a "Party" or collectively as "Parties" to this Agreement.

This preamble shall be a definitive part of this Agreement

WHEREAS under this Agreement it is agreed that DATA of medical research may be transferred between Parties to this Agreement only through the conditions stipulated in this Agreement;

WHEREAS the PROVIDER retains all ownership rights of DATA procured from the study;

WHEREAS under this Agreement it is agreed that the DATA to be transferred pursuant to this Agreement are only those to be used for academic or research purposes;

WHEREAS it is hereby agreed that no transfer to third parties is allowed, except for academic or research purposes where RECIPIENT has seeked the written consent of the PROVIDER;

WHEREAS it is hereby agreed that the RECIPIENT shall cooperate with the PROVIDER to facilitate capacity building in DATA management and its use;

AND WHEREAS the parties to the Agreement undertake to be bound by any lawful order or instruction, as they will be from time to time be advised by the Ethics-Issuing Organization.

NOW THEREFORE in consideration of the mutual benefits to be derived and the representations, conditions and promises herein contained,

the PARTIES HEREBY AGREE AS FOLLOWS:

ARTICLE 1

DEFINITIONS AND RULES OF INTERPRETATION

1.1 Definitions

"Agreement" means the "DATA Transfer Agreement for Researchers/Organizations" between the Parties.

"DATA" in this context refers to facts, observations, or any information generated and documented (numerical, descriptive or visual) as specified in *Annex 1*, which forms part of this agreement.

"Medical Research Coordinating Committee" means a committee of the NIMR Council which reviews, monitors and [illegible] Research in the United Republic of Tanzania.

"Permit-Issuing Organization" means the entity with the legal authority under the law to issue permits and/or to conduct scientific research or to do any activity collateral to that scientific research or matters connected thereto.

"Permit" means all consents, approvals, authorization, notifications, concessions, acknowledgements, licenses, permits or similar terms required to be obtained from any Permit-Issuing Organization.

"PROVIDER" means a person or organization providing the original DATA.

"RECIPIENT" [illegible] to which the original DATA is transferred.

"The Law" means [illegible] applicable laws of the United Republic of Tanzania or the RECIPIENT country when there is a [illegible] in the laws of Tanzania.

CONFIDENTIAL MATTER means information that is PROVIDER's proprietary and confidential information. Such CONFIDENTIAL MATTER shall not include any item of information, data, that [illegible] the public [illegible] prior to the time of the disclosure by the PROVIDER to the Receiving Party or thereafter [illegible] within the public domain other than as a result of disclosure by the RECIPIENT or any of its representatives in violation of this Agreement; (b) was [illegible] to the date of [illegible] in the possession of the RECIPIENT; (c) is acquired by the RECIPIENT from [illegible] party [illegible] obligation of confidentiality; (d) is hereafter independently developed by the RECIPIENT without reference to the information received from the PROVIDER; [illegible] PROVIDER expressly [illegible] allows the RECIPIENT to disclose.

1.2 Rules of Interpretation

In this Agreement:

a) The headings are for convenience only and shall not be considered in interpreting this

Agreement.

b) The [illegible] the plural and the vice versa;

c) The obligations on part of the PROVIDER or RECIPIENT shall be interpreted to apply to the co[illegible] of the PROVIDER Investigator or RECIPIENT Investigator, respectively.

ARTICLE II

GUIDING PRINCIPLES FOR DATA TRANSFER AGREEMENTS

1. This Agreement shall be linked to a project that has received ethical clearance from the MRCC under the National Institute for Medical Research. The need to transfer DATA shall be stipulated in an approved proposal or subsequent amendment. Any proposal that has received clearance from a Medical Research Coordinating Committee will require the Agreement to be processed through the National Institute for Medical Research.

2. Signing of this Agreement shall be mandatory for all research involving foreign researchers, and this shall be declared in a research application for a research permit.

3. This Agreement shall also be mandatory for local researchers collaborating with foreigners, before sending [illegible] DATA. Such research [illegible] Agreement applies also to local researchers when using DATA [illegible].

4. Make [illegible] and provide all necessary [illegible] applications for the consents to the Permit-Issuing [illegible] obligation to make all such applications and shall use all reasonable [illegible] obtained and,

5. In the case [illegible] counterpart, before signing the Implementing Letter of Agreement [illegible] collaborating institutions in the PROVIDER country, in this case, the United Republic of Tanzania should access information from the *National Research Registry* formed under the Tanzania Commission for Science and Technology (COSTECH) Act No. [illegible] of 2008 and published in 2012 [illegible] to determine whether the foreign researcher had obtained [illegible].

ARTICLE III

TRANSFER OF THE DATA

3.1 *DATA to be t[illegible]*

Subject to the [illegible] Agreement, the PROVIDER agrees to transfer the DATA and the [illegible] the DATA as identified in *Annex I*.

3.2 *Obligation of the PROVIDER*

It is hereby [illegible] that the following conditions to the Agreement shall be binding on the RECIPIENT:

(a) The RECIPIENT agrees to use, store or dispose of the DATA in compliance with all applicable [illegible] relating to research involving the use of human and animal subjects.

(b) The DATA [illegible] the property of the PROVIDER and PROVIDER hereby consents to the DATA [illegible] service to the research community.

(c) The RECIPIENT shall use the DATA for teaching or academic research purposes only.

(d) Except [illegible] approved by the Permit-Issuing Organization, and with the written consent [illegible] PROVIDER, the RECIPIENT shall not transfer or distribute the DATA to a third party.

(e) The RECIPIENT will acknowledge the source of the DATA in any publications reporting use of it.

(f) Subject [illegible] Article V of this agreement, the RECIPIENT shall be liable for damages which may arise [illegible] use and disposal of the DATA.

(g) The RECIPIENT [illegible] shall sign two copies of this Agreement and return [illegible] to the PROVIDER. The PROVIDER shall then transfer the DATA.

(h) The RECIPIENT shall provide the PROVIDER with a manuscript of any proposed publication [illegible] from the study using the DATA at least thirty (30) days prior to [illegible] publication or presentation. The PROVIDER reserves the right to [illegible] manuscript and to require the removal of CONFIDENTIAL MATTER [illegible] proprietary rights and interests. PROVIDER shall notify RECIPIENT [illegible] thirty (30) day period concerning the removal of CONFIDENTIAL [illegible] suggested modifications in the manuscript.

3.3 *Obligation of PROVIDER*

It is hereby [illegible] that the following conditions to the Agreement shall be binding on the PROVIDER:

(a) The PROVIDER agrees to use, store or dispose of the DATA in compliance with all applicable [illegible]

(b) The PROVIDER shall transfer timely the DATA upon receipt of one of the two copies duly signed [illegible]

(c) Subject to [illegible] the PROVIDER may [illegible] to make the DATA available under a separate [illegible] to [illegible] at nonprofit organizations or governmental agencies [illegible] to replicate the RECIPIENT Investigator's scientific research.

(d) Subject to Article V of this Agreement, the PROVIDER shall be liable all liabilities for damages [illegible] arising out of RECIPIENT's use, storage and disposal of the DATA.

ARTICLE IV

COSTS AND PAYMENT ARRANGEMENTS

The DATA shall be [illegible]

ARTICLE V

WARRANTIES

Any DATA transferred pursuant to this Agreement is understood to be experimental in nature. The PROVIDER and RECIPIENT MAKE NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESS OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE DATA WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS.

ARTICLE VI

LEGAL TITLE TO THE DATA TRANSFERRED AND BENEFIT SHARING

Legal title to the DATA [illegible] vested by this Agreement in the transfer of any Material hereunder. (i) As between the PROVIDER and the RECIPIENT, the PROVIDER shall be the sole owner of all right, title and interest in and to any inventions and intellectual property rights. (ii). The PROVIDER and RECIPIENT [illegible] sharing of benefits arising from use of the DATA in accordance with this [illegible]

ARTICLE VII

PERMITS, LICENCES AND APPROVALS

Prior to commencement of this Agreement, PROVIDER and RECIPIENT shall, at their own expense:

a) Make or cause to be made all necessary participate applications for the consents to the Permit-Issuing [illegible] and [illegible] diligently pursue all such applications and shall use all reasonable efforts to maintain the consents in effect once obtained and;

b) Give all required notices and allow all required inspections under all consents obtained in connection with this contract. The information supplied in the applications shall be complete and accurate and shall satisfy the substantive and procedural requirements of the applicable laws of the United Republic of Tanzania or of the other country where the DATA is transferred.

ARTICLE VIII

NON-EXCLUSIVE LICENSE

The transfer of the DATA will entitle a non-exclusive license to use the DATA solely for academic and research purposes only. The transfer of DATA does not grant the RECIPIENT any additional rights in the DATA other than specifically set forth in this Agreement.

ARTICLE IX

AMENDMENTS

This Agreement may be amended by mutual written Agreement of the Parties, which shall enter into force on the date agreed by the Parties.

ARTICLE X

TERMINATION

Termination of this Agreement may occur by:

a) Immediately upon mutual written consent of both Parties;
b) Either of the Parties with sixty (60) days written notice to the other Party; or
c) Upon immediate notice if a Party's contravention of law, and
d) As stated in Article XI

ARTICLE XI

APPLICABLE LAW, SEVERABILITY

The Parties recognize that this Agreement is a contract and not an international agreement, that

International Law is not applicable to this Agreement, and that International Law does not govern the interpretation of [illegible] of this Agreement. Any dispute arising under this Agreement which is not disposed of by [illegible] between the Investigators shall be submitted jointly to the Authorized signatories of this Agreement. A joint decision of the Authorized signatories or their designees shall be the [illegible] of such dispute. If the Parties cannot reach a joint decision, either Party may terminate this Agreement immediately.

The Parties hereby submit to the jurisdiction of the Courts of the United Republic of Tanzania for any action, suit or proceeding arising out of or relating to this letter agreement brought against the United Republic of Tanzania or NIMR; and to the jurisdiction of the courts of the RECIPIENT Government for any action brought against the RECIPIENT Government or any of its agencies.

This Agreement is effective when duly executed by all Parties and countersigned by the Chair of the Medical Research Coordinating Committee (MRCC) of the Government of United Republic of Tanzania. The Authorized officials who sign this Agreement certify that they are the legal representatives of the respective organizations authorized to sign on behalf of their respective organizations for the purpose of binding their organizations to the terms of this Agreement, for the transfer specified above.

ARTICLE XII

NOTICE

All notices pertaining to or required by this Agreement shall be in writing, shall be signed by an authorized representative and shall be addressed to the addresses indicated on the signature page for each Party.

ARTICLE XIII

NONAPPLICABILITY OF THIS AGREEMENT TO EXISTING OR FUTURE AGREEMENTS

The terms of this Agreement are not subject to and do not affect any other existing or future agreements between the Parties.

IN WITNESS WHEREOF the PARTIES have signed this Agreement in the presence of the witnesses and at the dates set out opposite their respective signatures.

SIGNATURE PAGE

FOR RECIPIENT

RECIPIENT'S A... ...

Signature

Printed Name and T...

Mailing Address f... MATERIAL

Tel: ... Fax ...

E...

... Investigator: I acknowledge ... this Agreement

... Jonathan Mnyambo

... Title

... Notices

... 513 SE-751 20 Uppsala SWEDEN

... Fax

... jonathan@sua.ac.tz

FOR PROVIDER

PROVIDER'S Au... Signatory

Signature

Georgies Ma... Program Manag...

Printed Name and ...

Mailing Address f... MATERIAL

SUA APOPO, ... University, P.O. Box 30... Morogoro Tanzania

Tel: 077738...

Email: georg... @apopo.org

... Principal Investigator: I acknowledge ... this Agreement.

... Mgode Program Manager

... Title

... Address ... Notices

SUA APOPO Sokoine University, ... Box 30... Morogoro Tanzania

...077738... Fax

...orgies.mgode@apopo.org

6/5/2019

CERTIFICATION

Authorized Officer CHAIRMAN ...

Signature

Printed Name and Title

Mailing Address

Tel:

Email:

Annex I

Description of Information to be [illegible] under this Agreement: (DTA)

1 1.Trained rats performance [illegible] on human sputum samples evaluated to find TB
1 2. Weight of rats
1 3. Age of rats
1 4. Hits on individual samples
1 5. Duration of performance

Was the DATA described above collected under (study Title) :

Training African giant pouched rats [illegible] for TB diagnosis

Research protocol Approved by Tanzania Authorities:
☑ Yes Certificate Number [illegible]
☐ No

Research protocol related Grant or Contact from RECIPIENT's Government or Organization
☑ Yes Number [illegible]
☐ No

Provider Investigator : I declare that the above mentioned type(s) and format of Dataset is/are only the one to be transferred here :

Name: Dr. George Mtinde

Signature: [illegible] 20 February 2019

Official Stamp

Authorized Official: CHAIRMAN [illegible] type (s) and format of Dataset mentioned here above to be transferred from Tanzania.

Name: [illegible]

Signature: [illegible] 2019

Official Stamp

SUA-APIPO