DEVELOPMENT OF A PREDICTIVE MODEL FOR THE RISK OF TYPHOID USING DATA MINING TECHNIQUES

JACOB OLAMIDE ABOLAJI

MATRIC NUMBER: 15010301034

**BEING A PROJECT SUBMITTED IN THE DEPARTMENT OF COMPUTER SCIENCE
AND MATHEMATICS, COLLEGE OF BASIC AND APPLIED SCIENCES
IN PARTIAL FUFILLMENT OF THE REQUIREMENTS FOR THE
AWARD OF DEGREE OF BACHELOR OF SCIENCE
MOUNTAIN TOP UNIVERSITY, IBAFO,
OGUN STATE, NIGERIA**

2019

CERTIFICATION

This Project titled, **DEVELOPMENT OF A PREDICTIVE MODEL FOR THE RISK OF TYPHOID USING DATA MINING TECHNIQUES**, prepared and submitted by **JACOB OLAMIDE ABOLAJI** in partial fulfilment of the requirements of the degree of **BACHELOR OF SCIENCE** (Computer Science), is hereby accepted

____________________________ (Signature and Date)

DR. I. O. AKINYEMI

Supervisor

____________________________ (Signature and Date)

DR. I. O. AKINYEMI

Head of Department

Accepted as partial fulfilment of the requirements for the degree of BACHELOR OF SCIENCE (Computer Science)

____________________________ **(Signature and Date)**

Prof. A. I. AKINWANDE

Dean, College of Basic and Applied Sciences

DEDICATION

This project is dedicated to God almighty

ACKNOWLEDGEMENTS

I am most appreciative to almighty God for the gift of life and continuous strength for not only leading me through this project but also for the successful completion of my study. I also acknowledge the effort of my indefatigable supervisor, who is also my HOD, Dr. I.O. Akinyemi for his constructive suggestions, understanding, motivations and useful comments. Sincere to say without his dedication, this project work would not have become a reality. My profound gratitude to the Dean of College of Basic and Applied Sciences Prof. A. I. Akinwande for his immeasurable effort and support. My amiable lecturers, especially Dr. P.A Idowu, Dr. M.O. Oyetunji, Dr. O.B. Alaba, Dr. O.O Taiwo, Dr. F.A. Kasali, MR. O.J. Falana and Mrs Olutosin Taiwo and the other members of the department of computer science and mathematics for the seed you have sown in my life through your lectures and words of wisdom which has helped me so far in my journey in this institution.

I am deeply indebted to many people for their immense contributions in diverse ways towards the successful completion of this research work. My appreciation goes to my irreplaceable parents, Mr & Mrs Jacob, A.A., who sacrificed wealth and enjoyable moments of their lives for the sake of my success, I want to thank them for their relentless effort towards ensuring the best education for me and also for their moral, financial, spiritual, physical and psychological support in my life. May you live long to eat the fruit of your labour; and my siblings – Oluwatoyin and Eniola Jacob for their prayers, understanding and support. Then lastly I would like to thank my Mountain Top University colleague and friends especially Oladotun Pelumi, Ogeleka Chinedu, Ntuk Anderson, Lajide Gbotemi, Afolabi Omolade, Oye Jibola and Oki David for their prayers and support, and help in one way or the other. God bless them all greatly.

ABSTRACT

Typhoid fever, also called enteric fever, is caused by Salmonella enterica serovarTyphi, a gram-negative bacterium. Estimates for 2000 suggest that around 21.5 million infections and 200 thousand fatalities due to typhoid fever are reported worldwide every year. Typhoid fever and paratyphoid fever, especially Among Nigerian kids and adolescents, keep being significant causes of disease and death. The aim of this study is to use Techniques for Data Mining to develop a Typhoid risk predictive model in Nigerians using relevant risk factors.

Historical data on the distribution of typhoid risk among participants were gathered using questionnaires after medical professionals identified linked typhoid risk variables. The predictive model for typhoid risk was developed using the algorithm for decision trees to define and account for variables linked to typhoid risk The Waikato Environment for Knowledge Analysis (WEKA) software – a suite of machine learning algorithms was used to develop the predictive model as the simulation environment. Holdout and 10-fold cross-validation techniques were used to evaluate the performance of the algorithms. The data sets were therefore subject to 10-fold cross validation using the two (2) chosen decision trees learning algorithms, namely: C4.5 implemented as the WEKA J48 algorithm and the Naïve Bayes algorithm.

The result of the performance evaluation of the C4.5 and naïve Bayes' algorithms are presented in Table 4.3. The true positive rate which gave a description of the proportion of actual cases that was correctly predicted which showed values of 0.783, 0.519, 0.722 and 0.619 respectively for no, low, moderate and high risk cases by the C4.5 decision trees algorithm and 0.739, 0.556, 0.611 and 0.667 for the naïve Bayes classifier.

The study presented a predictive model of typhoid risk using relevant risk factors selected from a predefined set of typhoid risk factors in Nigerians using the C4.5 decision trees algorithm that outperformed the performance of the classification of the naïve Bayes. A stronger understanding of the connection between the characteristics appropriate to typhoid risk was suggested following the creation of the forecast model for typhoid risk classification. The model can also be incorporated into the current Health Information System (HIS) that captures and manages clinical data that can be supplied to the predictive model of typhoid risk classification, thus enhancing clinical choices influencing typhoid risk and evaluating clinical data that affects typhoid risk from distant places in real time.

Keywords: Typhoid, Classification, Data Mining, Machine Learning, Predictive Model, Naïve Bayes, Decision Tree

TABLE OF CONTENTS

CHAPTER ONE

CHAPTER TWO

Content Page

Content Page

CHAPTER FOUR

CHAPTER FIVE

LIST OF TABLES

Table Page

LIST OF FIGURES

APPENDIX

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Typhoid fever, also called enteric fever, is caused by Salmonella enterica serovarTyphi, a gram-negative bacterium. The disease is mainly correlated with low socio-economic status and bad hygiene, the only known natural hosts and infection reservoir for human beings. Estimates for 2000 suggest that around 21.5 million infections and 200 thousand fatalities due to typhoid fever are reported worldwide every year. It is therefore considered One of public health's most severe threats to infectious diseases on a global scale, with particular concern about the rapid and widespread Developing opposition to multiple antibiotics (Crump, Youssef & Luby, 2003 The global concern about typhoid reflects the perception that typhoid in Nigeria is a prevalent and severe disease in kids and adults, where highly publicized outbreaks among Professionals of the public and health have strengthened this view. Using the Widal test commonly to screen febrile Kids and adults in hospital and outpatient settings is one consequence, since few centers are capable of performing blood or bone marrow culture, accepted standard gold diagnostic tests as few centers are capable of performing blood or bone marrow culture (Chart, Cheesbrough & Waghorn, 2000).

The Global typhoid burden disease estimates was Based on a total of 22 incidence studies of 19 non-African continents and only three Africans. Based on these information and a forecast rule based on climatic and socio-economic features, continental estimates of the disease burden were obtained (Crump et al, 2003). These estimates proposed mild typhoid incidence in most African nations, with the greatest incidence in adolescence, of 10–100 cases/100,000 years for individuals. In East Africa, the estimated incidence was 39/100,000 years for people Increased information volumes have recently been reported Concerning the incidence of various Pathogens discovered in kids with disease at health facilities in Africa (Graham, Molyneux, Walsh, Cheesbrough, Molyneux & Hart, 2000). These data significantly heightened the profile of infections with non-typhoidal salmonella. However, to date, these information were not used to examine the burden of typhoid-related disease. It was hypothesized That understanding of the comparative burden of typhoid-related disease, but Not the precise burden, recent Information on invasive bacterial infection aetiology based on the facility could be obtained. If Typhoid is a significant pathogen, prevalent, it ought to be observed often in circumstances where other significant circumstances are pathogens in the childhood, Streptococcus and haemophilusinfluenzae, for example, they are often noted. From the view of the health care system and the health care providerWe also looked at relevant information on the usefulness of clinical case definitions and the Widal test for the diagnosis of typhoid in African children.

Clinical Decision Support Systems (CDSS) provide understanding and person-specific data to clinicians, employees, patients and other people, intelligently filtered and submitted at suitable moments, to improve health and health care (Osheroff et al., 2006). Medical errors have already become a global society's universal matter. In 1999, IOM (American Institute of Medicine) released a study entitled "It's human to err" (Kohn et al, 2000), which stated: First, the amount of medical errors is amazing, the medical errors have already become the fifth deadly; second, the majority of medical errors have happened due to the human factor that could be avoided through the computer system. Improving healthcare quality, decreasing medical mistakes and ensuring patient security are the hospital's most severe responsibility. The clinical guideline can improve the safety and quality of the diagnosis and therapy, which has already been widely approved for its significance (Miller and Kearney, 2004). In 1990, Clinical practice guidelines were described As "Systematically created declarations to help medical practitioners and patient choices for particular clinical conditions regarding suitable health care" (Field and Lohr, 2005). The Clinical Decision Support System (CDSS) is any software that requires data about a clinical condition as input and generates Inferences of the production that can be of help to professionals in making decisions and would be considered by users of the program as "smart" (Musen, 1997). Data mining is a method of discovery in big data repositories of significant helpful information.

Data mining is a method of discovery in big data repositories of significant helpful information.

Data mining can find useful but hidden knowledge from databases, particularly those used to store health-related data about patient-affected illnesses (Jing-Song et al., 2011). Clinical choices are often produced on the grounds of the intuition and experience of physicians rather than the knowledge-rich information concealed in the database. This practice results in unwanted biases, mistakes and excessive medical expenses that affect patients ' quality of service (Aqueel and Shaikh, 2012Integrating Clinical decision support with computer-based records of patients could decrease medical errors, improve patient safety, reduce unwanted variations in exercise, And enhance the outcome of patients (Chen and Greiner, 1999). This proposal is promising as instruments for data modeling and analysis, such as information mining, have the ability to create a knowledge-rich atmosphere that can assist enhance the quality of clinical choices considerably. This research is driven by the need to apply information mining methods in non-typhoid Nigerian people to develop a predictive model as a warning scheme for the risk assessment of typhoid.

1.2 Statement of the Problem

Typhoid fever and paratyphoid fever, especially among Nigerian kids and adolescents, keep being significant causes of disease and death. Even if the symptoms are gone, individuals can still carry typhoid bacteria, which means they can spread it to other people through their faeces. An estimated 11-20 million individuals become infected with typhoid and between 128 000 and 161 000 die each year. Poor communities and vulnerable groups are at the highest risk, including children. Developed nations throughout the world have systems in place that assist in the early detection of diseases, but developing nations such as Nigeria lack the availability of such capable systems. A model is needed to identify the risk of typhoid Nigerians in order to prompt early detection, hence this study.

1.3 Aim and Objectives

The aim of this study is to use Techniques for Data Mining to develop a Typhoid risk predictive model in Nigerians using relevant risk factors. The specific research objectives are to:

i. Elicit knowledge on the risk factors of typhoid;

ii. Formulate a predictive model from (i) above;

iii. Simulate the model in (ii); and

iv. Validate the model

1.4 Scope of the Study

The scope of this research is restricted to patients from kids to adults of all age groups. The data gathered for this research is restricted to information gathered from people in the western portion of Nigeria and the Lagos state is a key location with an adequate amount of patients that have encountered this disease one way or the other, the research uses primary data to distribute questionnaires.

1.5 Significance of the Study

Typhoid cases in Nigeria are growing and can become a severe burden if there is no way to curtail their impacts. Data mining methods can assist define interactions and events not observed by medical health professionals required to improve the decision influencing early detection of typhoid and thus reduce the probability of danger. Early detection of typhoid danger will assist to provide prospective people with alternative lifestyle patterns to prevent typhoid from starting. Nearly one new disease has emerged each year over the past few decades, with more than 75% of these diseases deriving from zoonotic origins. More investment and research is now needed to help us better manage these diseases.

This project aims at addressing the challenges posed by typhoid, also known as Enteric fever from the Gram-negative bacterium Salmonella entericaserovarTyphiin Offering monitoring, prevention and control of infectious diseases and assessing how it impacts public health in the 21st century. The project draws on the parallels between this disease, learning from existing challenges And the goal of connecting people to lay the foundations for a worldwide practice society. The aim is to renew and reinforce the development of scientific knowledge and to build human capital.

1.6 Arrangement of Work

Chapter one was presented in this section during the presentation of a description of other chapters.

Chapter two contains a review of the literature and its deploring effects on the subject of typhoid,

The importance of past patterns of typhoid in identifying future patterns in other people, applying machine learning algorithms in medicine, and developing models for accessing public health risk.

Chapter 3 includes the techniques of studies used to create the model from data identification and collection, model formulation and algorithms to be used in conjunction with the simulation environment and parameters for model validation.

Chapter four presents' results and the study findings are discussed.

Chapter fives specifies the overview of the work done, the study conclusion, and the recommendations based on the study result.

1.7 Definition of Terms

Salmonella- Infection (salmonellosis) is a prevalent bacterial disease that impacts the intestinal tract. Salmonella bacteria typically live in the intestines of animals and humans and are shed by feces. People are most frequently infected with contaminated water or food.

Bacterium- A member of a large group of unicellular microorganisms with cell walls but lacking in organelles and a structured nucleus, some of which may cause illness.

Synopsis- A short overview of something or a general survey.

Morbidity- The state of illness

Mortality- Death on a big scale in particular.

Virulence- Severe or harmful disease or toxicity

Serotypes- A microorganism strain that is serologically distinguishable.

Enteric- Concerning or happening in the intestines

Genomics- The molecular biology branch concerned with genome structure, function, development and mapping.

Predictive- Denoting or referring to a scheme for using information already stored on a computer or mobile phone to produce letters or phrases that are likely to be entered by the user next, based on those already entered.

Model- A three-dimensional representation, typically on a lower scale than the original, of an individual or thing or of a suggested framework.

Simulation- Imitation of a method or scenario.

Decision tree- A decision tree is a decision support instrument that utilizes a tree-like decision graph or model and its possible implications, including the results of chance events, resource expenses and usefulness.

Validation- Action to check or prove something's validity or precision.

Evaluation- Judging the quantity, number or value of something; evaluation.

Data set- A collection of associated information sets consisting of distinct components but which can be manipulated by a computer as a unit.

CHAPTER TWO

LITERATURE REVIEW

2.1 Typhoid

Typhoid fever is one of the preceding causes of morbidity and mortality in the world. Typhoid is caused by a bacterium of the Salmonella genus. Infection with salmonella in humans can be classified into two broad types caused by low-virulence Salmonella enterica serotypes causing food poisoning, And that caused by high-virulence Salmonella enterica typhi (S. typhi), which causes typhoid, And a serovar group known as S Paratyphi A, B and C that causes Paratyphoid (Kanungo, 2009). Human beings are the only host of this latter group of pathogens. S. Typhi is a highly adapted pathogen specific to humans (Nagashetty, 2010), and the disease caused by these bacteria is a serious concern for public health, especially in the developing world (Wang lx, 2012). Recent estimates have shown that 22 million later cases of typhoid occur in the world each year, with some 200,000 deaths (Karkey, 2010), Indicating that this disease's global burden has steadily increased from an estimated 16 million previously (Sharma PK, 2009) Case-fatality rates have, however, decreased significantly (Karkey, 2010). Southeast and Central Asia is believed to have Maximum number of instances (0.100 per 100,000 people per year) and consequent deaths (Crump JA, 2004). Typhoid is usually endemic in poor regions of the globe where the provision of secure drinking water and sanitation and bad quality of life are inadequate. Although contaminated food and water are the main risk factors for typhoid incidence, a number of other variables have been recorded in different endemic environments, such as bad sanitation, Closer contact with instances of typhoid or carriers, educational level, Larger household size, location closer of water bodies, flooding, personal hygiene, poor lifestyle, and endemic travel.

Furthermore, climate variables for example, rainfall, vapor pressure and temperature have a significant the impact of typhoid infection transmission and distribution in humans (Gasem MH, Dolamans, 2001). (Anna E. Newton 2014) Confusion may occur in some cases, Symptoms may last months without treatment. It's not common to have diarrhea. Other people may carry the bacterium without being affected, but they can still pass the disease on to others. Typhoid fever, together with paratyphoid fever, is an enteric form of fever. Poor hygiene and poor sanitation are some of the risk factors. (Wain, J, 2015) Those in emerging countries who are traveling are also at risk. (Milligan, R, 2018) Only humans can be infected. The symptoms Similar to many other infectious diseases. (Wain, J, 2015) Diagnosis is either by culturing the bacteria or by detecting DNA in your blood, Bone marrow or stool of the bacterium. Cultivating the bacterium can be difficult. The bone marrow's most precise test (Milligan, R, 2018). Over the first two years, a typhoid vaccine can prevent about 40% to 90% of cases. For up to seven years, the vaccine may have some effect. For people at high risk or Travel to fields where disease is common, it is recommended. Other efforts to prevent the disease include clean drinking water, good sanitation, and washing of hands. The individual should not prepare food for others until it has been confirmed that the infection of an individual is cleared. The Antibiotics like azithromycin, fluoroquinolones or cephalosporins are treated for the disease. Of the third generation (Wain, 2015) Resistance to these antibiotics has developed, making It's harder to treat the illness (Chatham-Stephens 2018). Worldwide, there were 12.5 million new cases in 2015. In India, the disease is most prevalent. Children are affected most frequently. In the developed world, disease rates decreased in the 1940s Increased sanitation and antibiotic use in the treatment of the disease. Approximately there are 400 recorded instances and approximately 6,000 people are estimated to have the disease. It resulted in around 149,000 deaths worldwide in 2015, above 181,000 in 1990 (approximately 0.3 percent of the total worldwide). Without treatment, the death risk may be as high as 20percent. It's From 1 to 4 percent with treatment. Typhus is another disease. However, because of the similarity in symptoms, the name typhoid means "like typhus."

2.1.1 Estimates of Disease Burden

Awareness of the burden of illness is crucial for several reasons: firstly, data on the impacts of the disease on human health and the local economy is essential to educate public health decision-makers; secondly, information on local trends is required to allocate funds; and thirdly, understanding local and regional disease patterns is necessary to provide timely advice to travelers. Economic Estimates of typhoid fever burden (defined as symptomatic S typhi infection) are frequently published (26·9 million instances of typhoid fever were recorded in 2010 (Buckle gc, 2010) and overall mortality information from worldwide and regional mortality surveys are accessible. (Lozano R, 2012) However, there are still bad comprehensive local surveillance information from endemic regions.

2.1.2 Diagnosis and Treatment of Typhoid Fever

Recent reviews of typhoid fever diagnosis and treatment show that laboratory diagnosis of typhoid fever depends mainly on PCR (the most suitable for epidemiological studies) or detection of bacteria in society in the blood (While sensitivity is still a constraint) (Khan kh, 2012). (Parry CM, 2011) The Widal test for Antibodies Production is unreliable and serology tests of the new generation such as typhidot and tubex in Africa or Asia have not proved reliable. The typhoid-paratyphoid diagnostic assay, which detects IgA, Is a fresh test format that is promising. This method has the specificity of detecting circulating IgA for typhoid fever diagnosis using ELISA (House D et al, 2001) and improves sensitivity (up to 100 %) by amplifying the signal by isolating and incubating peripheral blood lymphocytes. (Khanam f, 2013) The treatment of fluoroquinolones, azithromycin and cephalosporin drugs of the third generation Is the primary chloramphenicol therapy in areas with susceptible strains.

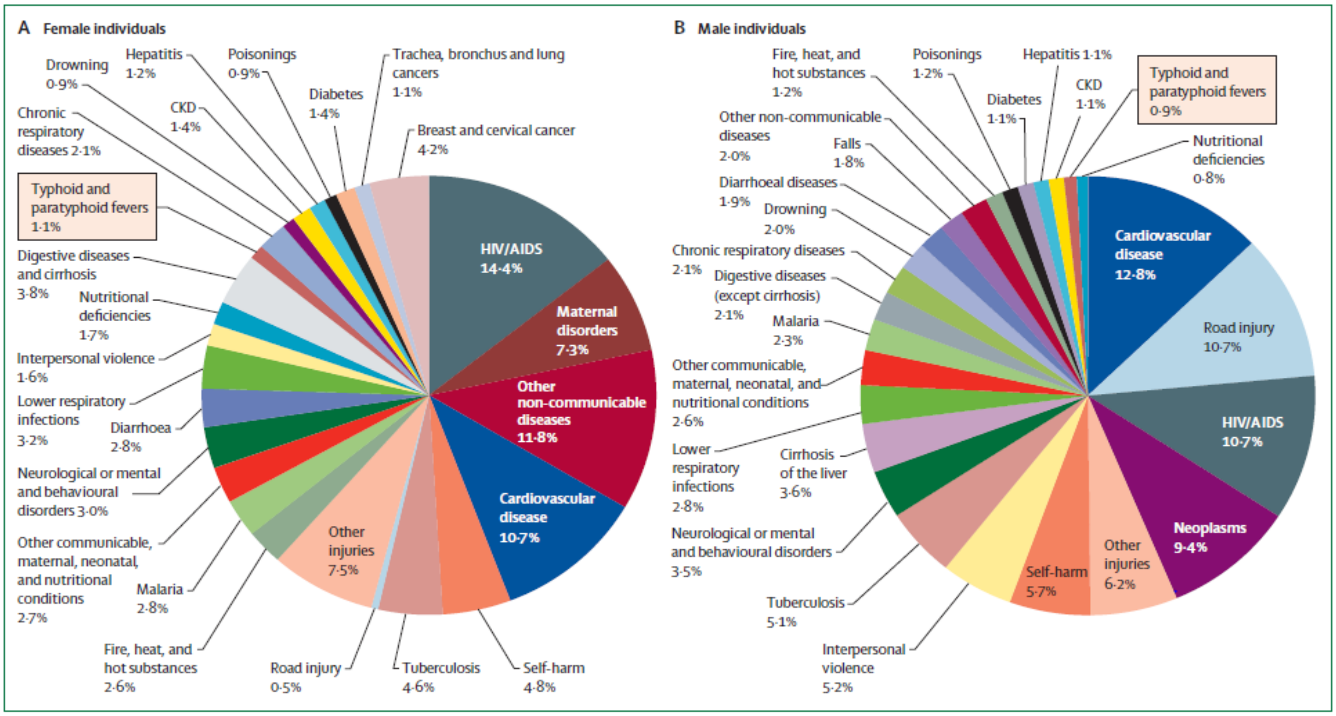


Figure 2.1: Mortality from enteric fever worldwide

Data from Jan 1, 2012, to Dec 31, 2012. 5 741 344 total deaths. CKD=chronic kidney disease. Used from Lozano and colleagues

2.1.3 Genomics of S Typhi

Despite the genetic resemblance of S Typhi and S entericaserovarTyphimurium (90% of genes are shared), there is little knowledge of the genetic variations underlying the capacity of S Typhi to trigger enteric fever, but not S Typhimurium. (Sabbagh Sc, 2010) Transposon knockouts ' large-scale libraries enable scientists to evaluate genome-level function and show distinctions between S typhi and S typhimurium. (Sabbagh SC, 2012) In S Typhi and S Typhimurium, the same genes could have distinguished regulatory pathways and potentially distinguished erent functions. (Barquist I et al, 2013) In endemic regions, this exciting new technology could provide objectives for the growth of S typhi vaccines and new antibacterial drugs. Investigate why S Typhi infects people, two mouse trials for typhoid fever have been created, but not humans. One is based on mice with haemopoietic stem cells in humans and the other is based on mice grafted with Toll-like knockouts with bacterial flagella recognition cells. (Mathur R, 2012)Research has also led to the description of gtgE, a virulence gene found in S Typhimurium but not in S Typhi that enables S Typhi to infect macrophages in the mouse. These models, although still to be verified, could allow pathogenesis and immunity to be investigated previously impossible for this human-restricted pathogen. By using the humanized mouse model to define city-specific receptor-binding and delivery systems for ctdB and pltA encoded typhoid toxin, the significance of these new models is proved, thus defining a future new destination for vaccines.

2.1.4 Typhoid Fever in Africa

Typhoid fever is even less well understood in Africa than in Asia; mostly due to poor laboratories diagnostic resources and inadequate infrastructure to support epidemiological and clinical studies. These problems reflect the difficulties facing a big, mainly poor continent with a heavy burden of HIV and unstable governments and healthcare priorities that overwhelm a country's capacity to provide secure food and drinking water. Access to safe water in Africa The access to piped water should not be confused because the age and resources of water treatment plants are reduced. (Starkl M, 2013)In Africa, Attempts to define typhoid fever burden clearly show a need for well-designed studies. Crump and colleagues tried to calculate the Burden on the basis of published research and found that too little data was available to predict anything better than a crude incidence rate—50 instances per 100,000 for a population of about 820 million.

(Lozano R, 2012) The 2010 Global Burden of Disease research estimated comparable rates, but one Kenyan publication added rates, These estimates were improved by Buckle and colleagues—724·6 instances per 100 000. Data from a region where no intervention was implemented to control typhoid fever drives the difference in estimates. Following two major placebo-controlled trials on the typhoid vaccine, previous studies as Crump and his peers used, assessed incidence rates of typhoid fever. In South Africa, More than 11 000 individuals and more than 32 000 individuals were engaged in these studies in Egypt, and so the protection of the herd (The incidence of typhoid fever in checks is therefore declining) could have lowered disease rates, where researchers identified the need to better estimate the disease burden. Using local resources could be one method of addressing this issue; several facilities in the least developed regions of Africa are capable of laboratory-based studies and have contributed to disease data. Comparison between Malawi and South Africa surveillance data shows variance in the demographics of typhoid fever patients. Typhoid fever in Malawi and South Africa continues a disease mainly in kids between 5 and 15 years of age, whereas in metropolitan regions of Kenya (where crude incidence levels of typhoid fever can be as high as 247 instances per 100 000), The disease is predominantly reported in children under 10 years of age. Incidence in metropolitan regions in rural Kenya is recorded 15 times. In Ghana, on the other hand, the disease appears to be more prevalent in kids below 5 years of age, and typhoid fever is commonly diagnosed in both adults and children in Moshi, Tanzania. These data are not comparable, as with early data from Asia, but the Demographic and geographic variations are evident among people prone to typhoid fever. Zimbabwe's position is particularly concerned with the 2013 cumulative formal figures by the end of April reporting Over 6800 suspected cases of disease And 142 patients with typhoid fever verified. A confounding factor in estimating Non-typhoid salmonellosis as an invasive and often deadly disease in Africa is the burden of typhoid fever.

2.1.5 Risk Factors

Typhoid fever continues to be a severe worldwide danger, particularly in the developing world, an estimated 26 million individuals or more are affected each year. In India, Southeast Asia, Africa, South America, etc., the disease is found (endemic).Worldwide, Children are at the highest danger of developing the disease, although they usually have milder symptoms than adults do.

You are at enhanced danger if you live in a nation where typhoid fever is uncommon:

- Work in or travel to areas where typhoid fever is established (endemic)
- Work as a clinical microbiologist handling Salmonella typhi bacteria
- Have close contact with someone who is infected or has recently been infected with typhoid fever
- Drink water contaminated by sewage that contains Salmonella typhi

2.1.6 Early Detection and Prevention

In many developing countries, it may be difficult to achieve the goals of public health that can help prevent and control typhoid fever— Safe drinking water, enhanced sanitation, and adequate medical care. Therefore, some specialists think that it is the best way to manage typhoid fever is to vaccinate high-risk populations. If you live or travel to areas where the risk of typhoid fever is high, a vaccine is recommended.

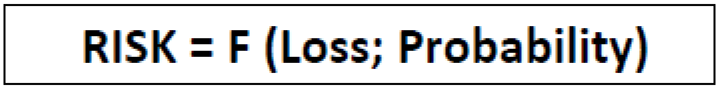
2.2 Machine Learning

Machine learning (ML) is an artificial intelligence branch that enables computers to use statistical and optimization techniques to learn from past examples (Quinlan, 1986; Cruz and Wishart, 2006). There are several machine learning applications, the most important being predictive modeling (Dimitoglou et al, 2012). The same set of features (attributes / independent variables) is used to represent each instance (records / sets of fields or attributes) in any dataset used by machine learning algorithms. The characteristics can be continuous, categorical or binary. If known labels are used for the instances (the corresponding target outputs), then, contrary to unsupervised learning, Learning is referred to as supervised where cases are not identified (Ashraf et al., 2013). Supervised one of the classifications tasks that Social Intelligent Systems most frequently perform. Thus, Artificial intelligence (logic-based methods) and statistics (Bayesian networks, instance-based networks) have developed a large number of techniques. The objective of supervised learning is to construct a concise model of class label distribution in terms of predictor characteristics. The resulting classification will then be used to assign class labels to test instances where the predictor value is known but the class label value is known (Gauda et al., 2013). Following is a discussion of some supervised algorithms of machine learning that will be considered for this study purpose.

2.2.1 Predictive Modeling

Without comprehending risk, you cannot Comprehension of risk adjustment or modeling.

• Risk is a mix of two variables at its most fundamental: loss and likelihood. We are defining a loss as occurring when the post-occurrence state of an individual is less advantageous than that of State of pre-occurrence. Financial risk depends on the quantity and probability of losses, but Risk and loss is not limited to financial quantities alone in healthcare. Therefore we use the definition of the following, more general:



RISK = F (Loss; Probability)

Predictive modeling is the method by which members are estimated, predicted or stratified Based on their comparative danger. For frequency (likely) and severity (loss), Predictions can be made separately. The idea of risk adjustment that is closely associated with predictive modeling. Their uses are one way to distinguish: Future-oriented predictive modeling; whereas Risk adjustment is often the case in the past.

2.2.2 Types of Predictive Models

Previously, machine learning has been used to predict business behavioral outcomes, such as identifying product consumer preferences based on prior history of purchase. There are a number of different techniques for developing predictive algorithms using a variety of predictive analytical tools / software and literature has described them in detail (Waijeeet al., 2010; Siegel et al., 2011). Examples include neural networks, vector machine support, decision trees, naïve Bayes, etc. For example, decision trees use techniques such as trees for classification and regression, boosting trees and random trees to predict different results.

Machine learning algorithms, such as random forest methods, have several benefits over traditional statistical explanatory modeling, like the absence of of a pre-defined hypothesis, making it less likely to ignore unexpected hypothesis (Liawet al., 2002). When there are many potential predictors available And when predictors interact with each other common in engineering, Processes of biological and social causes, it can be quite efficient to approach a predictive issue without a particular causal hypothesis. Therefore, Predictive models using algorithms for machine learning can facilitate recognition of significant variables that may not be originally recognized (Waijeeet al., 2010). Indeed, in the machine learning literature, there are many examples of discovering unexpected predictor variables (Singalet al., 2013).

2.2.3 Developing a Predictive Model

The first stage in the development of a predictive model is to use traditional regression analysis to select appropriate candidate predictor variables for possible incorporation in the model, but the best approach for this is not agreed (Royston et al., 2009). A backward elimination approach begins with all factors of candidates, hypothesis tests are applied sequentially to determine which variables from the final model should be removed, while all candidate variables are included in the full model strategy in order to avoid potential overfitting and selection bias. Irrespective of their statistics significance, significant predictor variables previously reported Usually should be included in the final model, However, the sample size of the dataset generally limits the amount of factors included (Greenland, 1989). Inappropriate variables selection is a prevalent and significant cause in this situation of poor model performance. Variables are less likely to be selected Using methods for machine learning as they are often not based solely on predefined hypotheses (Ibrahim et al., 2012). There are several other significant information problems when designing a predictive model. Management, such as addressing Missing information and conversion of variable (Kaambwaet al., 2012; Waijeeet al., 2013b).

2.2.4 Validating a Predictive Model

In order to be valuable a predictive model, not only must it have predictive capacity in the derivation cohort, But also in a validation cohort perform well (Hemingway et al., 2009). For several reasons, Model performance may vary significantly between derivation cohorts and validation cohorts, including model overfitting, missing significant predictor variables, and predictor inter-observer variability this leads to mistakes of measurement (Altman et al., 2009).

Therefore, model performance in the derivation dataset can be overly optimistic and is not a guarantee that the model will perform similarly well in a fresh dataset. There are a number of predictive studies released focusing solely on model derivation and very scarce validation studies (Waijeeet al., 2013b). Using internal and external validation, validation can be carried out. A common internal validation approach is to divide the data into two parts–a set of training and validation. If, given the limited data available, splitting of the dataset is not possible, for inner validation measures such as cross validation or bootstrapping (Steyerberg et al., 2010). Researchers are often tempted to dismiss the original model and use the validation dataset to create a fresh predictive model when a validation research demonstrates disappointing outcomes

2.2.5 Assessing the Performance of a Predictive Model

Remembering is essential when evaluating model performance that explanatory models are evaluated on the basis of associations ' strength, while predictive models are assessed exclusively on their capacity to predict correctly. The Predictive model performance is evaluated by means of several complementary tests evaluating overall performance; calibration, discrimination and reclassification (Steyerberget al., 2010). Datasets for derivation and validation, Determination and reporting of performance features.

Calibration is the difference between the data set observed and predicted event rates and is evaluated using the Hosmer-Lemeshow test (Hosmer et al., 1997). Discrimination is a model's ability to distinguish between records that do and do not have an interest outcome, and it is commonly evaluated using Operating Properties Receiver (ROC) curves (Haegrty et al., 2005). ROC analysis alone, however, is relatively insensitive to evaluate Differences between excellent models of prediction (Cook, 2007), so several relatively new performance measures have been proposed. The improvement of Net reclassification and inclusive discrimination are measures used to evaluate modifications in the predicted classification of outcomes between two models (Pencina et al., 2012).

2.3 Data Mining

Data Mining, Also popularly referred to as Database Knowledge Discovery (KDD), relates to the non-trivial extraction from database data of implicit, previously unknown and possibly helpful information.

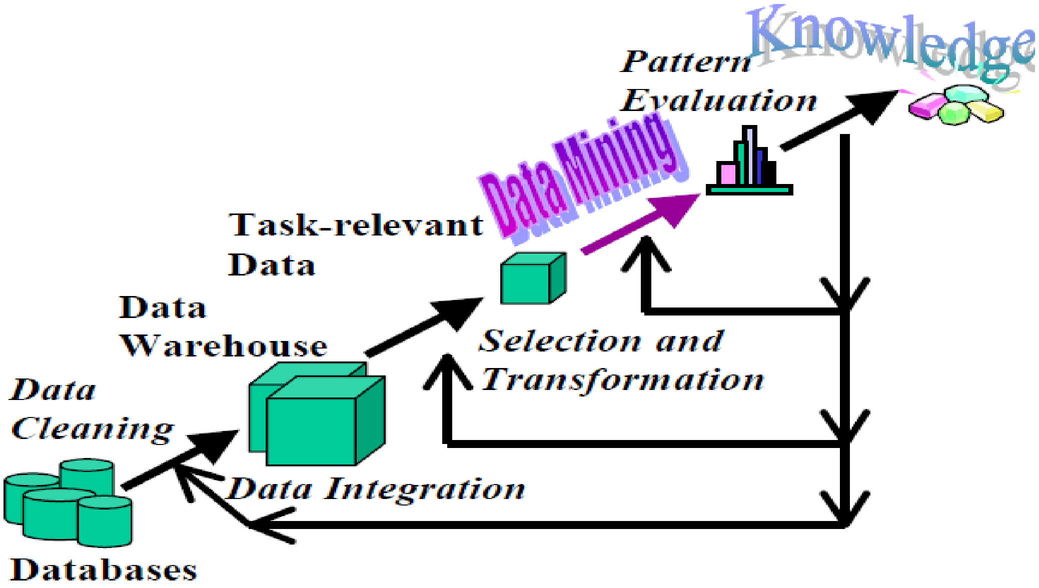


Figure 2.2 : Knowledge discovery in data base

Database knowledge detection is an accurate process consisting of several separate steps (Hemalatha and Megala, 2011). Data mining is the foundation step resulting from unknown discovery but useful knowledge from enormous databases. A formal concept of the discovery of knowledge in databases is given as a Computer assisted digging method and analyzing huge data sets and then the significance of the information is extracted. Data mining instruments predict future trends and behaviors, enabling businesses to make proactive decisions based on knowledge (Han and Kamber, 2001). Expertise in data mining provides a consumer-oriented approach to new and unknown data patterns. Healthcare administrators can use the exposed knowledge to advance service superiority. Data mining is gradually becoming more well-liked in healthcare, if not more essential. The use of data mining applications in health care has been motivated by several factors (Canlas, 2009). For example, the presence of fraud and abuse in medical insurance has resulted many insurers to try Reduce losses through the use of information mining instruments to assist them discover and monitor perpetrators (Christy, 1997). Fraud detection using In the business globe, for instance, data mining apps are common, when fraudulent credit card transactions are detected (Biafore, 1999). Recently, Successful application of information mining to detect fraud and abuse in healthcare have been reported (Wanqinget al., 2010).

2.3.1 Significance of Data Mining in Healthcare

All healthcare organizations around the world have generally stored data on healthcare in electronic format. Healthcare data mainly contains all the patient information as well as the healthcare industry stakeholders. Such data type storage is increased very rapidly. There is a type of complexity in it due to the continuous increase in the size of electronic healthcare data. We can say that data on healthcare becomes very complex, in other words.

It becomes It's very hard to extract significant data it by using traditional methods. But the meaningful patterns can now be extracted from it due to advances Iin Statistics, mathematics and many other disciplines. In a scenario where there are big collections of information on healthcare, data mining is beneficial.

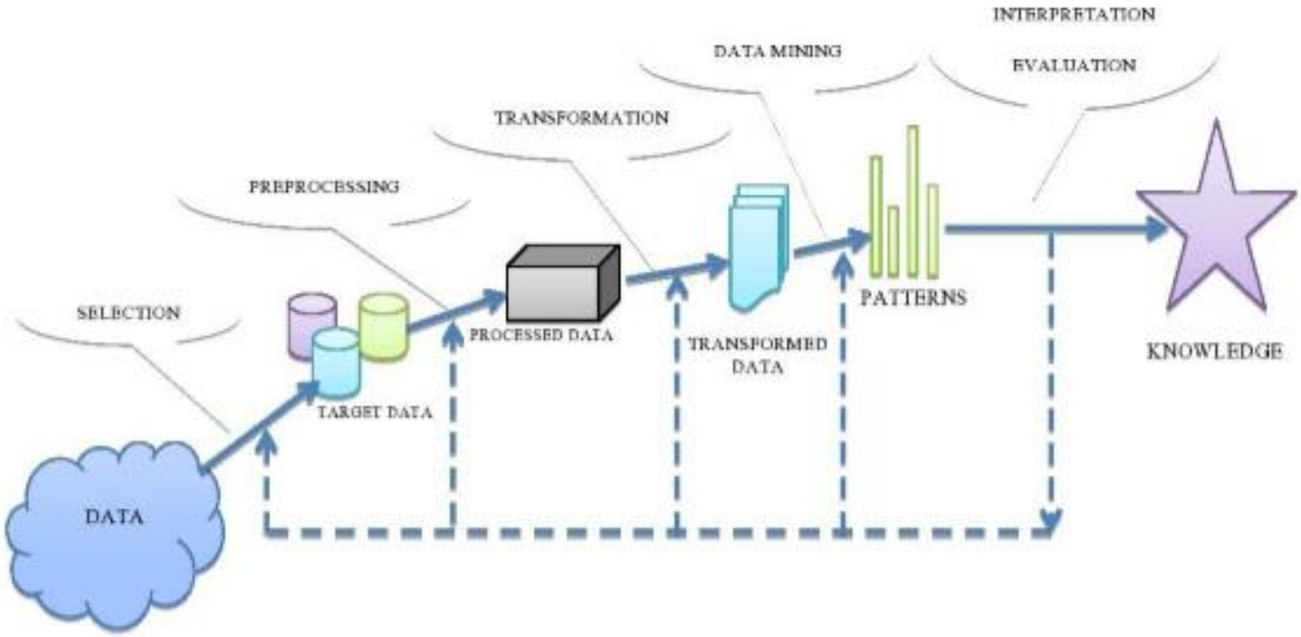


Figure 2.3 : Stages of Knowledge Discovery Process

Data Mining extracts mostly meaningful patterns that wasn't known before. Then you can integrate these patterns Knowledge and key decisions can be produced with the assistance of this understanding.

2.3.2 Data Mining Applications in Healthcare

Despite the differences and conflicts in approaches, today more data mining is needed in the health sector. There There are several arguments that might be put forward Promoting data mining use in the health industry, covering not only public health concerns But also the private health sector (including public health stakeholders, as can be shown later).

a. Overload of data. Computerized health records provide a wealth of knowledge. However, the overwhelming amount of data stored in these databases makes it extremely difficult, if not impossible, for people to search and discover knowledge (Cheng, et al, 2006). IndeedTo some specialists, medical breakthroughs have slowed down due To the extent and complexity of present medical data. For this purpose, Data mining and computers are best suited (Shillabeer and Roddick, 2007).

b. Medicine based on evidence and hospital error prevention. By applying information mining to their current information, medical organizations can find fresh, helpful and life-saving information that would otherwise remain Inert in the databases. For example, Continuing hospital and safety studies discovered that 87 percent of hospital deaths could have been avoided in the US had hospital staff (including physicians) be cautious to avoid mistakes (Health Grades Hospitals Study, 2007). Such security issues could be flagged and covered by leadership of hospitals and governmental regulators through mining records in hospitals.

c. Public health policy making. Lavrac et al. (2007), combined GIS and data mining with Weka and J48 (free, open source, information mining instruments based in Java) to evaluate similarities among, among others, community health centers in Slovenia. They found patterns among health facilities using information mining that gave their Institute of Public Health policy recommendations. They concluded that "data Methods of supporting mining and decision making, including new methods of visualization, Can contribute to an improvement decision-making performance.

d. Early detection and/or disease prevention. Cheng et al cited the use of classification algorithms to help detect heart disease early, a major concern for public health around the world. Cao et al (2008) Data mining described as a tool for monitoring trends in cancer vaccine clinical trials.

Medical specialists could discover better models and anomalies than just a set of tabulated information through Use of data extraction and visualization.

e. Early diagnosis and administration of pandemic diseases and formulation of public health policy. Health experts have also started to explore how you can use information mining to identify and handle pandemics early. Kellogg et al. (2006) Described methods combining spatial modeling, simulation and mining of spatial data to identify interesting disease outbreak characteristics. The analysis resulting from data mining in the simulated environment, more informed policy-making could be used to detect and manage disease outbreaks.

f. Support for non-invasive diagnosis and decision making. Invasive, expensive and painful for patients are some diagnostic and laboratory procedures. An example of this is performing a cervical cancer biopsy in women. The K-means algorithm for clustering was used by Thangavel et al (2006) to analyze patients with cervical cancer Clustering was discovered to have better predictive outcomes than current medical opinions.

2.3.3 Challenges to Data Mining in Healthcare

Because we know that various healthcare organizations generate and store a lot of healthcare data. But there are various health data challenges that can pose serious obstacles in making the right decisions. The first challenge with healthcare data is the data format that is stored in different healthcare organizations is different. There is no default format for data to be stored up to date. This lack of standard format in epidemic situations can make epidemic situations even worse. Suppose an epidemic disease spreads in its various geographical regions within a country. The country health ministry requires all health-care organizations to share their health-care data for analysis with their centralized data warehouse to take all the essential steps to resolve the epidemic situation. But since the data formats are different. Consequently, data analysis may take longer than usual. Because of this, the situation may become out of control. The data on healthcare is very useful for extracting significant data to improve patient healthcare services. It is very important to do this data quality because we are unable to extract meaningful information from data that is not of any quality. Therefore, data quality is another very important challenge. Data quality depends on different factors such as removing noisy data, free of data missing, etc. To maintain the quality of healthcare data, all necessary steps must be taken. Another major challenge is the sharing of data. Neither patients nor organizations of health care are interested in sharing their personal data. Because of this, the epidemic situations may get worse, it may not be possible to plan better treatment for a large population, and it may not be possible to detect fraud and abuse in health insurance companies, etc. Another challenge is that it is very expensive and time consuming process to build the data warehouse where all healthcare organizations within a country share their data.

2.4 Classification

Classification is one of Data Mining's most popular methods in the healthcare sector. It splits samples of data into target classes. The classification technique the target class is predicted for each data point by analyzing their disease patterns, Can be connected with a risk factor with patients with the assistance of the classification approach. It is a supervised approach to learning that has known class categories. The two methods of classification are binary and multilevel. Only two feasible classes in binary classification, such as "high" or "low" risk patient can be considered whereas, for example, the multi-class Approach has more than two objectives, "high," "medium" and "low" risk patient. The data set is partitioned as a data set for training and testing. It involves predicting an outcome based on a specific input. Training set is the algorithm of a set of characteristics to predict the outcome. To predict the result, it attempts to find the connection between characteristics. Their Target or forecast is their result. Another algorithm called the set of predictions is available. It is made up of the same set of attributes as the training set. But the prediction attribute is still to be known in the set of predictions. It mainly analyzes the input to process the prediction. The term defining the algorithm's "good" is its accuracy. Consider Pawti Medical Center's medical database, the training set consists of all the patient information previously recorded. Whether or not a patient was having a heart issue is the attribute of prediction. The different classification algorithms used in health care are as follows:

a. K-Nearest Neighbor (K-NN) K-Nearest Neighbor (K-NN) classifier is one of the simplest classifiers to detect unidentified data points using previously known data points (nearest neighbor) and classified voting data points (C. McGregor, 2012). Consider that there are different objects. It would be good for us to know the characteristics of one of these artifacts to predict for its closest neighbors because similar characteristics are found in the nearest neighboring objects. K-NN's majority votes Can play a very significant part in this classifying any new instance Where k is a favorable (tiny number) integer. It is one of the simplest techniques of data mining. It is primarily known as memory-based classification because there must always be examples of runtime training in memory (Alpaydin, E. 1997). In case of continuous attributes, the euclidean distance is calculated when we take the difference between the attributes. But when large values hold the lower ones down, it has a very severe issue. Continuous attributes need to be standardized to address this major problem in order to have the same influence on distance measurement between distances (Bramer, M., 2007).

K-NN Has a number of apps in different areas in various fields, including Health data, field of picture, analysis of clusters, recognition of patterns, online marketing, etc. KNN classifiers have different advantages. These are: ease, effectiveness, intuition, and performance in many areas of competitive classification. If the data on the training is large, it will be effective and noisy training data will be robust. The big Requirement for memory to store entire sample is a major disadvantage of KNN classifiers. If there is a big sample then on a sequential computer there will also be a big reaction time.

b. Decision Tree (DT) DT is considered one of the most famous of these classifier approaches. We can construct a decision-making tree using available data that can address the issues related to different areas of research. It is the same as the flowchart in which each non-leaf node Refers to a study on a specific attribute and each branch denotes the result of that test and each leaf node has a list of class. The node of root is a decision tree's highest node. For example, we can decide if or not a patient needs to be readmitted with the assistance of the medical readmission decision tree. Domain knowledge is not required to make a decision on any issue. Decision Tree's most common use is to calculate conditional probabilities in operational research analysis (Goharian & Grossman, 2003). Using Decision Tree, decision-makers can choose the The best root-to-leaf option shows a distinctive class separation based on maximum information gain (Apte S.M , 1997). Several advantages of the Tree of choice as follows: Trees of choice are self-explanatory and easy to follow when compacted.

c. It was developed at the beginning of the 20th century (Anderson, J. A, 1995). It was considered the best classification algorithm prior to the introduction of decision trees and the Support Vector Machine (SVM) (Obenshain, M. K, 2004). That was one of the reasons that encouraged NN in various fields of biomedicine and healthcare as the most widely used algorithm for classification (Bellazzi, R, 2008). NN, for instance, was commonly used as the algorithm that supports disease diagnosis including cancers (Romeo, M., 1998) and predicts results (Sharma, A., 1997). In NN, neurons or nodes are the basic elements. There are interconnections between these neurons and worked in parallel within the network to produce the functions of the output. They are able to make fresh findings from existing observations even in situations where certain neurons or nodes within the network fail or fall due to their ability to work in parallel. Each neuron is associated with an activation number and each edge is assigned a weight within a neural network. Neural network is mainly used to perform classification and pattern recognition tasks (M. H. Dunham, 2003). An NN's basic property is that through weight adjustment and structural modifications, it can minimize the error. Only because of its adaptive nature, it minimizes the error. NN is capable of producing more accurate predictions. One of NN's major advantages is that it can handle noisy training data properly and can reasonably be classified as fresh data types that differ from training data. NN also has different disadvantages. Firstly, it needs many parameters, including the optimal amount of parameters empirically determined hidden layer nodes, and its output in classification is highly sensitive to the selected parameters. Secondly, its process of training or learning is very slow and very expensive computationally. Another is that they do not provide any internal details about the phenomenon being investigated at the moment. It's like a "black-box" approach for us, therefore. Bayesian methods for probabilistic method of learning Bayesian classification is used. It can be easily obtained with the assistance of the classification algorithm. (Bayes statistics theorem plays a very significant part). While attributes such As signs of patients and their condition of health are correlated with one another in the medical domain, Classifier Naïve Bayes assumes all characteristics are independent. This is Naïve Bayes Classifier's major disadvantage. If attributes are independent, the classifier ‘Naïve Bayesian’ has shown great accuracy performance. They play very significant roles in the healthcare sector. There are therefore various advantages of BBN that Researchers worldwide have used them. One of them is that it helps make the process of computing very easy. Another is that it has better speed and accuracy for huge data sets (C. McGregor, 2012).

2.5 Related Works

Corner et al. Health Geographics International Journal 2013. Model typhoid danger in Bangladesh's Dhaka metropolitan region: the role of financial and environmental socio-economic variables: The three main factors used together describe 87 percent of the variance in the original applicant predictors, which eminently qualifies them for use as a collection of uncorrelated explanatory variables in a linear regression model. The initial outcome of regression using Ordinary Least Squares (OLS) was disappointing, which could be explained by analyzing the spatial autocorrelation intrinsic in the main variables. Based on these variables, the use of Geographically Weighted Regression caused a significant rise in the predictive power of regressions. The Determined analysis by the Akaike Information Criterion (AIC) analysis, the best forecast was discovered when the three variables were mixed into a quality of life index using a technique earlier released by others and had a determination coefficient of 73%. The typhoid occurrence / risk forecast equation was used to create the first risk map displaying Dhaka Metropolitan Area regions whose residents are at higher or lower risk of typhoid infection. This, combined with seasonal data on the incidence of typhoid also reported in this document, has the capacity to advise public health experts on creating approaches for prevention such as targeted vaccination.

Trans R Trop Med Hyg from Soc. 2006. An easy forecast principle for diagnosing typhoid fever in Turkey has been validated. The objective of this research was to create a straightforward rule of prediction for typhoid fever diagnosis. A model for predicting typhoid fever patients at hospital admission was obtained and validated by assigning weighted point values to autonomous predictive variables connected with a hospital admission diagnosis of typhoid fever. The use of demographic, clinical and experimental factors was used to match individuals with cell culture-confirmed typhoid fever with nurses with flu of unidentified source. The model was obtained in two distinct patient cohorts at Dicle University Hospital in Diyarbakir, Turkey. A maximum of 371 clients have been registered. A therapeutic coefficient rating was created using seven autonomous predictive variables at hospital admission connected with typhoid infection: Age < 30 years, abdominal distention, confusion, leukopenia, relative bradycardia, positive wide test and typhoid tongue. A clinical prediction concept assisted distinguish people with typhoid fever.

Vijayalaxmi V. Mogasale et al. (2018) Estimated risk of typhoid fever associated with lack of safe water access. Unsafe water is a well-known danger of typhoid fever, but a pooled estimate of the danger of typhoid fever at population level due to exposure to unsafe water has not been quantified. It will be helpful to accurately estimate the threat from unsafe water to demarcate high-risk populations, model the burden of typhoid disease, and target prevention and control operations. Methods. We conducted a comprehensive literature review and meta-analysis of observational studies that assessed the danger of typhoid fever connected with unimproved drinking water as described by WHO-UNICEF or microbiologically hazardous drinking water. A random effects model was used to calculate the mean value for the pooled odds ratio from case-control studies. We also mentioned categories of other risk variables from the chosen research as well as unimproved water and unsafe water.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This section shows the methodology design used to develop the predictive model for typhoid probability in a well narrated way. The methodology comprises of a series of methods / techniques that started with the identification of typhoid predictive factors alongside the information collection technique used to obtain the information needed to develop the model. In addition to the target variable (risk of typhoid) as output variable, the gathered historical information contained individual documents composed of their corresponding values for each recognized variables (risk factors) as inputs.

3.2 Data Identification and Collection

Following the evaluation of associated literature work in the typhoid body of understanding and risk-related factors, a number of variables (risk factors) were recognized. An expert physician with over 10 years of medical practice experience validated the recognized typhoid risk factors before the information collection tool was constructed alongside the identification of participants due to the absence of accessibility of information linked to typhoid risk but for those with the disease, the chosen information collection instrument for this research is the questionnaire. Appendix I shows the questionnaire provided to the chosen participants for this research.

3.2.1 Questionnaire Design

Before the questionnaire was constructed, the specialist physician supplied data on the related Typhoid risk variables. Risk factors connected with typhoid were categorized as:

a. Demographic – the demographic information used are: gender, age, marital status, ethnicity, occupation, religion and academic qualification;

b. Current and past illness – these factors were identified as those factors that describe the individual's social, habitual and behavioral patterns, e.g. and the way they relate with people on a day to day basis;

c. Dietary – these factors were identified as those associated with the individual's nutritional status, regularity of exercises etc;

d. Domestic – these factors were identified as those factors that describe the living condition of the individual i.e. the environment (either with the kids or with the spouse), local food exposure; and

e. Occupational – these factors were identified as those factors that describe the working condition of the individual i.e. occupation, Number of hours invested in work at the office, job position and office stress.

The constructed questionnaire consisted of three (3) sections, namely A, B and C sections. Section A of the questionnaire consisted of data relevant to the demographic information of the individual, namely: gender, age, education, occupation, marital status, employment, residential area and ethnicity. Section B of the questionnaire consisted of data about the threat variables of the danger typhoid from the individual respondent. Section C is the doctor's remarks; this room is left to the doctor free to comment on the related typhoid danger based on the data provided on each questionnaireIt is essential to state that the remarks of the physician are subjective to the knowledge of their own medical practice and may not be a real depiction of the generic danger of typhoid in Nigeria.

3.3 Formulation of the Predictive Model for Risk of Typhoid

The predictive model for typhoid risk was developed using the algorithm for decision trees to define and account for variables linked to typhoid risk and to gather historical explanations of the connection between the recognized risk factors and their corresponding risk for each record. Supervised machine learning algorithms were used to formulate predictive models as the pattern explaining the connection between the recognized factors (input variables) and the corresponding typhoid risk (target variable) was needed, the pattern can then be transformed into a set of rules that can help doctors make informed decisions about the risk of typhoid from Nigerians. For any supervised machine learning algorithm proposed to formulate a predictive model, a mapping function can be used to express the general expression of the predictive model for typhoid risk easily– resulting in most machine learning algorithms being black-box models using evaluators instead of power series / polynomial equations. Historical dataset S composed of records of people with areas representing the set of risk variables I amount of variable inputs for people ; Xij alongside the corresponding target variable (risk of typhoid) represented by the Yj variable – the risk of typhoid in the information gathered from the hospital chosen for the research for the jth person.

Equation 3.1 displays a mapping function that describes the risk factor-to-target class relationship – typhoid risk.

$$\varphi: X \rightarrow Y \qquad (3.1)$$

$$defined\ as: \varphi(X)=Y$$

The equation demonstrates the connection between risk variables set by a vector, X composed of I risk factor values, and Y label defining typhoid risk – low, moderate, and high typhoid risk expressed in equation 3.2. Assuming the risk factor set values for an individual are represented as X={X1,X2,X3, ... ,Xi } where Xi is the value of each risk factor, i=1 to I ; then the mapping π used to represent the typhoid risk predictive model maps each individual's risk factors to their respective typhoid risk according to equation 3.2.

Assuming that an individual's risk factor set values are represented as X={X1,X2,X3,...... Where Xi is the value of each risk factor, i=1 to I; then the map used to depict the predictive typhoid risk model maps the risk factors of each individual to their corresponding typhoid risk according to equation 3.2.

$$\varphi(X) = \begin{cases} Low\ risk \\ Mild\ risk \\ High\ risk \end{cases} \qquad (3.2)$$

The decision trees created for the danger of typhoid in people were used to suggest a set of guidelines that can be used directly to determine the risk risk of typhoid by observing the values of risk variables recognized by the model and the sequence of occurrences. Furthermore, the set of characteristics recognized for typhoid in the final decision trees model are the most appropriate risk factors for determining the risk of typhoid in people and it was suggested that significant consideration be provided to the physician during the typhoid risk assessment of an individual.

In the following chapter, you will find the decision trees algorithm used in people to formulate the predictive model of typhoid danger. Although it is essential to emphasize that the algorithm of decision trees produces a hierarchical tree structure with a top-down structure using a dividing criterion with an inherent conditional probability using the occurrence of risk variables. Thus, the dividing criteria used by each decision trees algorithm submitted were used to determine the range of tree nodes from the parent node (the most significant variable) to the subsequent nodes all the way to the leaf (target class representing typhoid danger).

3.3.1 Decision Tree (DT) Classifier

It was suggested that the predictive model for typhoid danger be formulated in people using a decision trees algorithm to classify the risk of typhoid as either low, moderate and high danger considering the values / labels of the defined risk variables (nodes) used in the growth of a hierarchical tree structure using a dividing criterion.

Each interior node of the choice tree (beginning from the root / parent node) reflects the characteristics (significant risk factors for typhoid danger) with edges corresponding to the values / labels of each attribute leading to a child node (another attribute dependent on the parent node value) at the bottom ; This method goes on until each successive attribute value reaches the leaf – the terminal node that also represents the target class (Typhoid risk – low, moderate or high).

By dividing the training data set into sub-sets based on an attribute value test for each input variable, the tree learns the pattern during the model development training process using the collected historical data set ; the process is repeated on each derived sub-set in a recursive manner called recursive partitioning. The recursion is finished when the sub-set target class value is the same at a node, or when splitting no longer adds value to the predictions. This is also called the induction of the top-down trees, an example of the greedy algorithm also known as divide-and-conquer.

3.3.2 Decision Trees (DT) Algorithms Used

The training data set, S is a set containing $S1$,$S2$,....,Sj of already classified samples of the records of individuals consisting of the values of their risk factors, X={$X1$,$X2$,....,Xi} alongside the risk of typhoid, Y={low,$moderate$,$high$} such that, S=(X,Y) for all individuals from 1 to j. The decision trees algorithm used to develop the predictive model for the risk of typhoid in individuals alongside their respective criteria was the C4.5 Decision trees algorithm (implemented as J48 algorithm). The theory of decision trees has the following parts: a root node which is the starting point of the trees with branches called edges connecting successive nodes showing the flow based on the values (edge for transition) of the attribute (node) and nodes that have child nodes are called interior nodes (parent nodes). Leaf or terminal nodes are those nodes that do not have child nodes and represent a possible value of the target variable (typhoid class) given the variables Represented by a root node path. Rules can then be induced by IF-THEN statements from the trees taking paths created from the root node all the way to their respective leaf.

The fundamental concept of any decision trees assessment is to divide the specified dataset into sub-sets by recursively partitioning the sibling nodes into child nodes depending on the homogeneity of the in-node cases or separating the inter-node cases with regard to their destination factors. Thus attributes are examined at each node and the splitter is selected as the attribute so that after dividing the nodes into child nodes according to the value of the attribute variable, the target is differentiated to the best use algorithm.

As a result of this, there is a need by the decision trees algorithm to distinguish between important variables attributes and attributes which contribute little to overall decision process which are based on the use of impurity measures. Following is the algorithm used by decision trees in growing their trees from a dataset containing a set of attributes. The algorithm is called TreeGrowth and takes in two arguments; which are the training records containing instances E and the attribute set (variables monitored) F which works by recursively splitting the data and expanding leaf nodes until a stopping criterion is met.

The motivation for using the selected decision trees algorithms are as follows:

3.3.3 Naïve Bayes Classifier

The Bayesian Classification reflects a monitored technique of teaching as well as a statistical method for model. It implies an inherent probabilistic model and enables us to catch confusion about the model in a principled manner by determining outcome probabilities. It can resolve diagnosis and predictive issues.

This ranking is named after Thomas Bayes (1702-1761), who proposed Bayes ' theorem. Bayesian classification offers practical learning algorithms and it is feasible to mix previous understanding with measured information. Bayesian Classification offers a helpful viewpoint for the comprehension and evaluation of many learning algorithms. It calculates specific hypothesis probabilities and is resistant to noise in output information.

The predictive model for the risk of typhoid was formulated using the naïve Bayes' classifier – a supervised machine learning algorithm that is based on the naïve Bayes' statistical theory of conditional probability shown in equation (1).

$$P(Class|Attributes) = \frac{P(Attributes|Class) * P(Class)}{P(Attributes)} \tag{3.3}$$

Where:

a. P(Class): Prior probability of class (risk of typhoid);

b. P(Attributes): Prior probability of training data attribute values;

c. P(Attributes|Class) : Probability of attribute values given the class; and

d. P(Attributes|Data) : Probability of Class given the attribute values.

Equation (3.3) is used to derive equation (3.4) and is used to estimate the probability of the record belonging to either of the classes (risk of typhoid) and the respondent is allocated to the class with maximum probability as shown in equation (3.5). Equations (3.4) and (3.5) are used by the naïve Bayes' classifier to formulate the prediction model for typhoid using the attributes, Xk representing the set of risk factors for the risk of typhoid and allocate a patient to either of class, Ci = {no, low, moderate, high}.

$$P(C_i|X_k) = \prod_{k=1}^{n} P(X_k|Ci) \cdot P(C_i) \quad (3.4)$$

$$Risk = MAX[P(no|X_k), P(low|X_k), P(moderate|X_k), P(high|X_k)] \quad (3.5)$$

3.4 Model Simulation Process and Environment

After identifying the decision tree (DT) learning algorithms required to formulate the predictive model for typhoid danger, the predictive model simulation was carried out using data gathered from individual documents containing risk factors information and their corresponding risk of typhoid gathered from hospitals in Western Nigeria. The Waikato Environment for Knowledge Analysis (WEKA) software – a suite of machine learning algorithms was used to develop the predictive model as the simulation environment.

The gathered information set was split into two sections: training and testing information–the training information was used to formulate the model while validating the model using the test information. The literature-based method of training and testing predictive model is a very challenging experience, particularly with the different validation methods available. It was natural to assess the efficiency of a classifier in terms of error rate for this classification issue. The classifier anticipated each instance's class–the record with values for each typhoid danger:

If it's right, it's considered a success; if not, it's a mistake. The error rate was the ratio of mistakes over a whole number of cases and thus measured the classifier's general efficiency the error level on the training data collection was unlikely to be a good measure of potential results; because the DT classifiers were obtained from the same training data.

To predict a classifier's performance on new data, there was a need to assess the predictive model's error rate on a dataset that played no role in the classifier's formation. This autonomous dataset was called the test dataset –a representative sample of the underlying issue, just like the training information. It was necessary that the test dataset was not used to produce the classifier in any manner since the classifiers for machine learning involve two phases: one to come up with a basic structure of the predictive model and the second to optimize parameters involved in that structure.

3.4.1 10-fold Cross Validation Technique

The process of leaving a portion of an entire dataset as test information while the remainder is used to train the model is called the method of holdout. The task here is to be able to discover a successful classifier by using as much historical information as possible to train; to achieve a decent estimate of error and to use it for model testing as much as possible. Holdouting one-third of the entire historical dataset for testing and the remaining two-thirds for practice is a prevalent practice.

The cross-validation method was used for this research, which involved splitting the entire datasets into a number of information folds (or partitions). Each partition was chosen for testing with the remaining k – 1 partitions used for training ; the next partition was used for testing with the remaining k – 1 partitions used for training (including the first partition used or testing) until all k partitions were chosen for testingThe error rate collected from each method was added to the mean recorded error rate. The process used in this research was the 10-fold stratified cross validation technique involving dividing the entire dataset into ten partitions. Figure 3.1 demonstrates a 10-fold cross validation process representation.

3.4.2 Simulation Environment

Weka is open source software under the GNU General Public License. The system was developed at the University of Waikato in New Zealand. Weka stands for the Waikato Environment for Knowledge Analysis. The software is freely available at http://www.cs.waikato.ac.nz/ml/weka. The system was written using object-oriented language, Java.

Weka can be used at several distinct levels. Weka offers state-of - the-art data mining and machine learning algorithms with implantations. Weka contains modules for data preprocessing, classification, clustering and association rule extraction for market basket analysis. The main features of Weka include:

a. 49 data preprocessing tools;

b. 76 classification/regression algorithms;

c. 8 clustering algorithms;

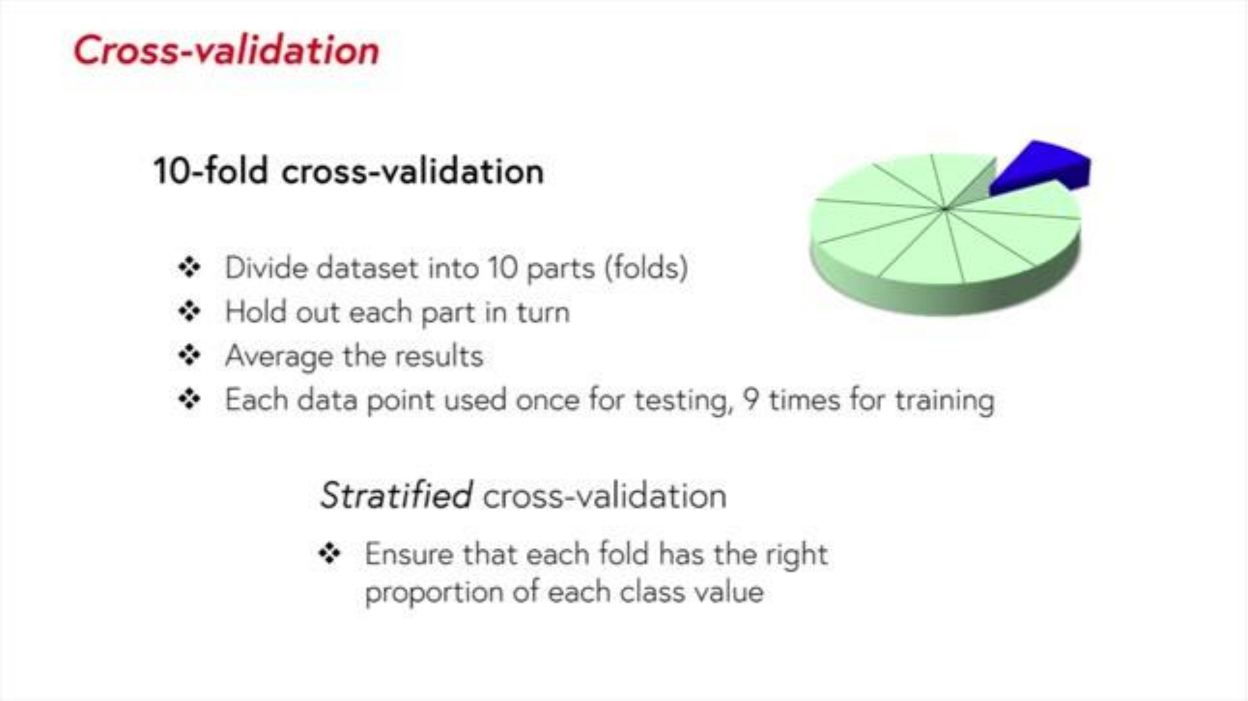
d. 15 attribute/subset evaluators + 10 search algorithms for feature selection;

e. 3 algorithms for finding association rules; and

f. 3 graphical user interfaces, namely:
i. The Explorer for exploratory data analysis;

ii. The Experimenter for experimental environment; and

iii. The Knowledge Flow, a new process model inspired interface.



Cross-validation

10-fold cross-validation

- Divide dataset into 10 parts (folds)
- Hold out each part in turn
- Average the results
- Each data point used once for testing, 9 times for training

Stratified cross-validation

- Ensure that each fold has the right proportion of each class value

Figure 3.1: 10-fold cross validation process

For the purpose of this research, the explorer used 3 distinct function choices each with its own distinctive search approach to perform the selection process. The training dataset comprising these cases was evaluated using the Weka environment experimenter interface following the identification of appropriate sensor-input factors. The data sets were therefore subject to 10-fold cross validation using the two (2) chosen decision trees learning algorithms, namely: C4.5 implemented as the WEKA J48 algorithm and the naïve bayes algorithm.

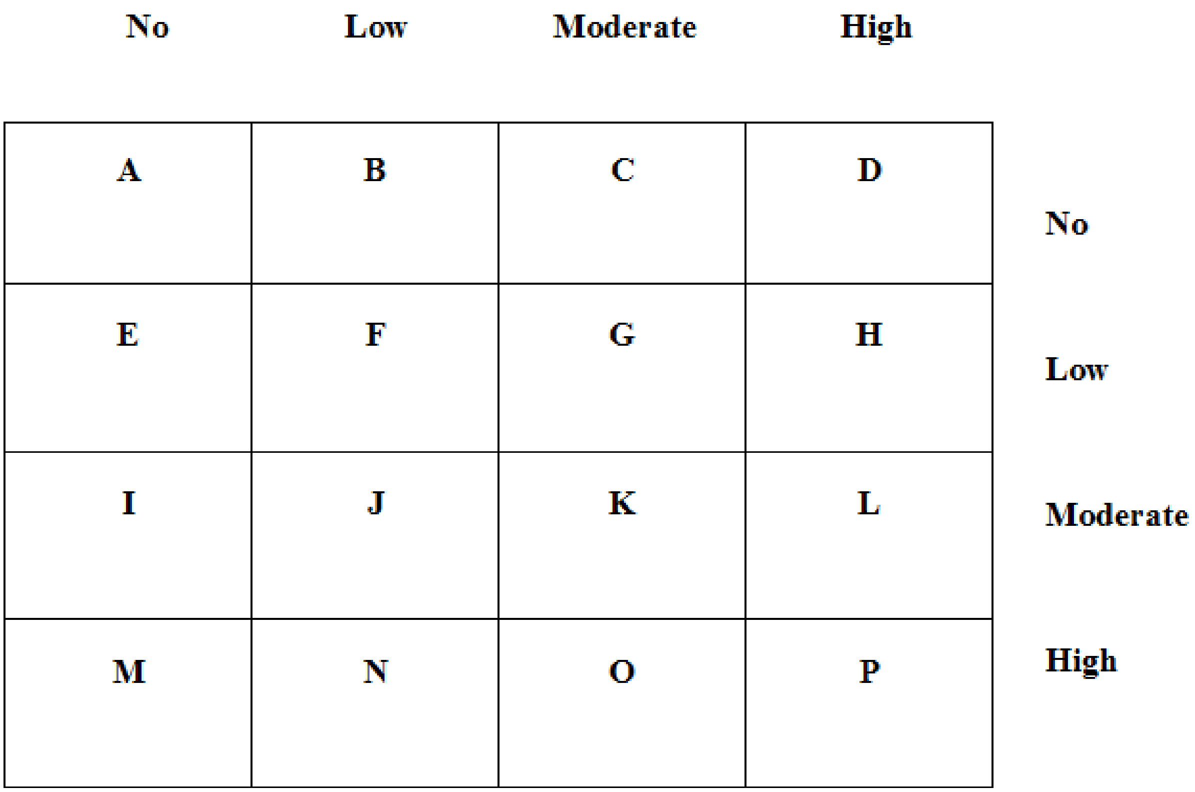
Before uploading the historical datasets comprising the risk factor values in addition to the risk of typhoid for each respondent's record in the initial dataset; storing the dataset according to the default information representation format required for information mining assignments in the Weka setting was necessary. The default form of file is called the file format attribute relationship (.arff). The arff file type stores three data categories: the first defines the relationship title, the second defines the relationship attributes alongside their respective labels and the third defines the relationship data followed by the values of each attribute for each record. You can also use Object-Database Connectivity (ODBC) to read information from comma-separated values (.csv) format and databases.

3.5 Performance Evaluation of Model Validation Process

During the course of evaluating the predictive model, a number of metrics were used to quantify the model’s performance. To determine these metrics, four parameters must be recognized from the outcomes of the classifier's projections during model experimentation. These are: real favorable (TP), real negative (TN), false-positive (FP) and false-negative (FP). In this research involving a binary classification, either endured or not existed can be regarded as favorable while the others are bad (e.g. if low is considered as a positive then moderate and high are negatives and vice versa).

True positives are the correct prediction of positive cases, true negatives are the correct prediction of negative cases, and false positives are the negative cases predicted as positives while false negatives are positive cases predicted as negativesThese results are presented on the confusion matrix – for this study the confusion matrix is a 4 x 4 matrix table due to the four (4) output class labels (see Figure 3.2).

Correct classifications were plotted along the diagonal from the north-west position for low cases predicted as No (A), low (F), followed by moderate predicted as moderate (K) and high predicted as high (P) at the south-east corner (also called true positive and negative). In the remaining cells of the confusion matrix (also known as false positives) the incorrect classifications were plotted. Also, the actual no cases are A+B+C+D, the actual low cases are E+F+G+H, the actual moderate cases.



No	Low	Moderate	High	
A	B	C	D	No
E	F	G	H	Low
I	J	K	L	Moderate
M	N	O	P	High

are I+K+K+L while actual high are M+N+O+P and the predicted no are A+E+I+M, low are B+F+J+N, predicted moderate are C+G+K+O and predicted high are D+H+L+P.

For each predictive system, the established system has validated a range of results metrics depending on the results of A – P in the nick equation. They're provided as follows.

a. Accuracy: The complete amount of right classifications

$$Accuracy = \frac{A + F + K + P}{total\_cases} \tag{3.6}$$

b. True positive rate (recall/sensitivity): The percentage of real instances properly categorized

$$TP_{no} = \frac{A}{A + B + C + D} \tag{3.7}$$

$$TP_{low} = \frac{F}{E + F + G + H} \tag{3.8}$$

$$TP_{moderate} = \frac{K}{I + J + K + L} \tag{3.9}$$

$$TP_{high} = \frac{P}{M + N + O + P} \tag{3.10}$$

c. False positive (false alarm/1-specificity): The percentage of negative instances wrongly categorized as positive

$$FP_{no} = \frac{E + I + M}{actual_{low} + actual_{moderate} + actual_{high}} \tag{3.11}$$

$$FP_{low} = \frac{B + J + N}{actual_{no} + actual_{moderate} + actual_{high}} \tag{3.12}$$

$$FP_{moderate} = \frac{C + G +}{actual_{no} + actual_{low} + actual_{high}} \tag{3.13}$$

$$FP_{high} = \frac{D + H + L}{actual_{no} + actual_{low} + actual_{moderate}} \tag{3.14}$$

d. Precision: Proportion of right projections

$$Precision_{no} = \frac{A}{A + E + I + M} \tag{3.15}$$

$$Precision_{low} = \frac{F}{B + F + J + N} \tag{3.16}$$

$$Precision_{moderate} = \frac{K}{C + G + K + O} \tag{3.17}$$

$$Precision_{high} = \frac{P}{D + H + L + P} \tag{3.18}$$

Using the aforementioned performance metrics, the performance of the predictive model for the risk of typhoid problem can be evaluated by validation using a historical dataset collected based on the information provided in the questionnaire. The TP rate and precision lie within the interval [0, 1], accuracy within the interval of [0, 100] % while the FP rate lies within an interval of [0, 1]. The closer the accuracy is to 100% the better the model, the closer the value of the TP rate and precision is to 1 the better while the closer the value of FP rate is to 0 the better. Therefore, the evaluation of an effective model has a high TP/Precision rates and a low FP rates.

CHAPTER FOUR

SIMULATIONS RESULTS AND DISCUSSION

4.1 Introduction

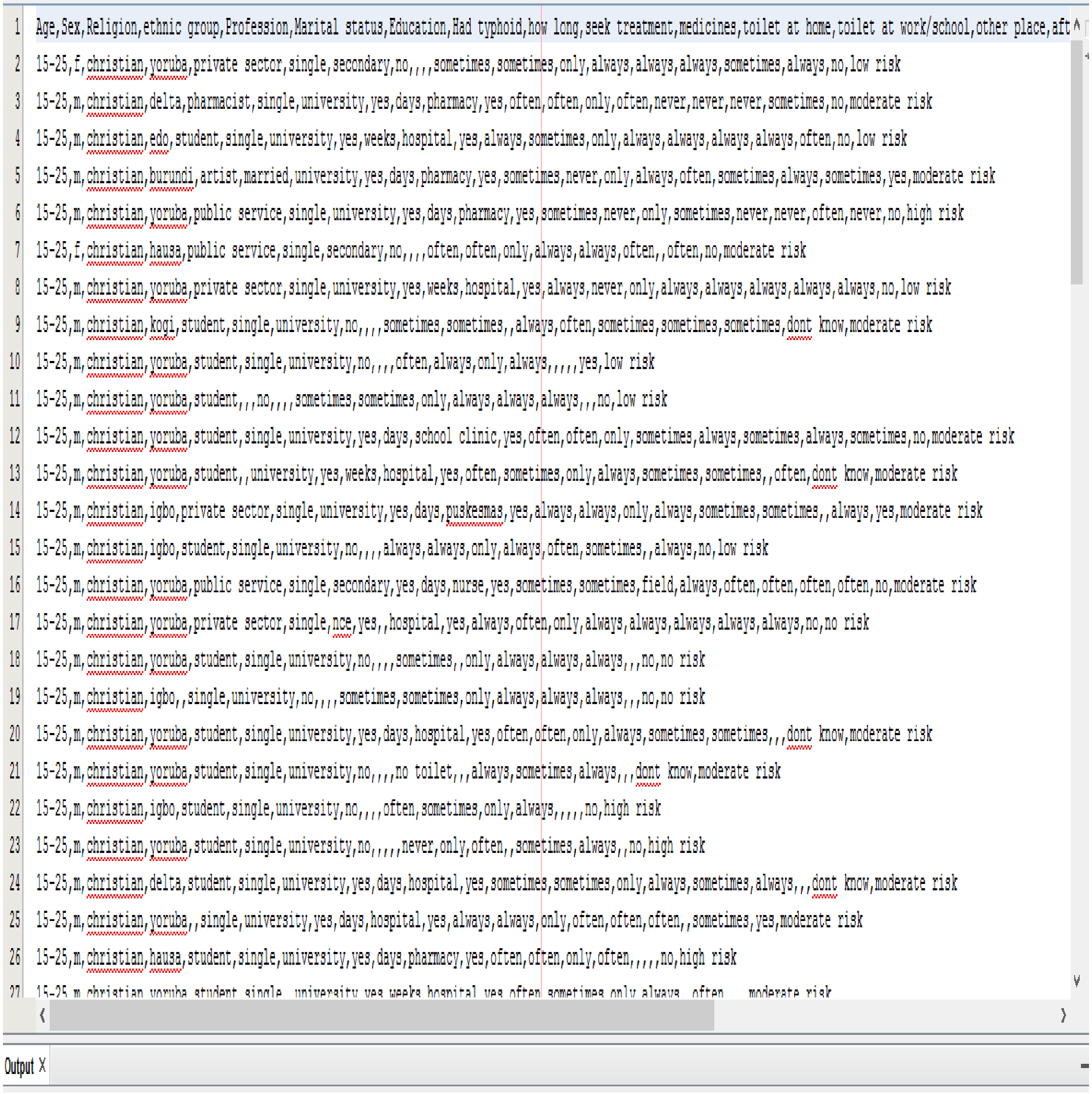
The findings of the methodological method outlined above are discussed in this chapter of the research. To know the distribution of the values of each risk factor of typhoid among the participants chosen for this research using the minimum and maximum values and the mean and standard deviation of the data distribution, a thorough inquiry was originally carried out into the evaluation of the description of the collected data set.

The findings of the model formulation and simulation method were subsequently provided for the creation of the predictive model for typhoid danger.

To define the most effective and efficient predictive model for typhoid risk, the performance of the predictive models for typhoid risk was created using the decision trees algorithms assessed. Thus, among the patients chosen for this research, the variables recognized by the decision trees algorithm were suggested as the most significant and meaningful indicator of typhoid danger.

4.2 Results and Discussion of Data Summarization of Historical Dataset

For this study, data was collected from 89 patients using the questionnaires constructed for this study among which; the moderate risk of typhoid was identified for 29 patients, the low risk of typhoid was identified 24 patients, no risk of typhoid was identified 21 patients, high risk of typhoid was identified for 13 patients. Figure 4.1 shows a screenshot of the data collected from the 89 respondents selected for this study. The data was placed in the attribute relation file format (.arff), which is the appropriate standard for the data mining test setting chosen for this research.



```
Age,Sex,Religion,ethnic group,Profession,Marital status,Education,Had typhoid,how long,seek treatment,medicines,toilet at home,toilet at work/school,other place,aft
15-25,f,christian,yoruba,private sector,single,secondary,no,,,,sometimes,sometimes,only,always,always,always,sometimes,always,no,low risk
15-25,m,christian,delta,pharmacist,single,university,yes,days,pharmacy,yes,often,often,only,often,never,never,never,sometimes,no,moderate risk
15-25,m,christian,edo,student,single,university,yes,weeks,hospital,yes,always,sometimes,only,always,always,always,always,often,no,low risk
15-25,m,christian,burundi,artist,married,university,yes,days,pharmacy,yes,sometimes,never,only,always,often,sometimes,always,sometimes,yes,moderate risk
15-25,m,christian,yoruba,public service,single,university,yes,days,pharmacy,yes,sometimes,never,only,sometimes,never,never,often,never,no,high risk
15-25,f,christian,hausa,public service,single,secondary,no,,,,often,often,only,always,always,often,,often,no,moderate risk
15-25,m,christian,yoruba,private sector,single,university,yes,weeks,hospital,yes,always,never,only,always,always,always,always,always,no,low risk
15-25,m,christian,kogi,student,single,university,no,,,,sometimes,sometimes,,always,often,sometimes,sometimes,sometimes,dont know,moderate risk
15-25,m,christian,yoruba,student,single,university,no,,,,often,always,only,always,,,,,yes,low risk
15-25,m,christian,yoruba,student,,,no,,,,sometimes,sometimes,only,always,always,always,,,no,low risk
15-25,m,christian,yoruba,student,single,university,yes,days,school clinic,yes,often,often,only,sometimes,always,sometimes,always,sometimes,no,moderate risk
15-25,m,christian,yoruba,student,,university,yes,weeks,hospital,yes,often,sometimes,only,always,sometimes,sometimes,,often,dont know,moderate risk
15-25,m,christian,igbo,private sector,single,university,yes,days,puskesmas,yes,always,always,only,always,sometimes,sometimes,,always,yes,moderate risk
15-25,m,christian,igbo,student,single,university,no,,,,always,always,only,always,often,sometimes,,always,no,low risk
15-25,m,christian,yoruba,public service,single,secondary,yes,days,nurse,yes,sometimes,sometimes,field,always,often,often,often,often,no,moderate risk
15-25,m,christian,yoruba,private sector,single,nce,yes,,hospital,yes,always,often,only,always,always,always,always,always,no,no risk
15-25,m,christian,yoruba,student,single,university,no,,,,sometimes,,only,always,always,always,,,no,no risk
15-25,m,christian,igbo,,single,university,no,,,,sometimes,sometimes,only,always,always,always,,,no,no risk
15-25,m,christian,yoruba,student,single,university,yes,days,hospital,yes,often,often,only,always,sometimes,sometimes,,,dont know,moderate risk
15-25,m,christian,yoruba,student,single,university,no,,,,no toilet,,,always,sometimes,always,,,dont know,moderate risk
15-25,m,christian,igbo,student,single,university,no,,,,often,sometimes,only,always,,,,,no,high risk
15-25,m,christian,yoruba,student,single,university,no,,,,,never,only,often,,sometimes,always,,no,high risk
15-25,m,christian,delta,student,single,university,yes,days,hospital,yes,sometimes,sometimes,only,always,sometimes,always,,,dont know,moderate risk
15-25,m,christian,yoruba,,single,university,yes,days,hospital,yes,always,always,only,often,often,often,,sometimes,yes,moderate risk
15-25,m,christian,hausa,student,single,university,yes,days,pharmacy,yes,often,often,only,often,,,,,no,high risk
15-25,m,christian,yoruba,student,single,,university,yes,weeks,hospital,yes,often,sometimes,only,always,,often,,,moderate risk
```

Figure 4.1: Screenshot of the dataset collected from the respondents

The format required the identification of three (3) parts of the dataset, namely:

a. The relation section: The name of the recognized file, in this case typhoid-model-training, was used to define the information comprising all 89 patients chosen for training and model testing afterwards. Use the name @relation before the name of the relationship to identify the relationship tag;

b. The attribute section: Used to define fields / attributes (risk factors) recognized as typhoid risk input variables where the latest characteristics describe typhoid risk. There are 21 attributes identified in the file with the first 20 identifying the input variables (risk factors of the risk of typhoid) while the last variable is the risk of typhoid. Each attribute has its own label, showing the feasible values that each attribute specified in the dataset can indicate. The attribute tag is recognized for each attribute using the @attribute name before each name of the attribute; and

c. The data section: For each respondent collected in the same order as the attributes were listed, the data set values were used. The data record of each respondent is depicted as the set of values on each line with the typhoid danger shown on each line's last part. The data containing the values of the attributes for each respondent is listed on the line following the name tag identified as @data.

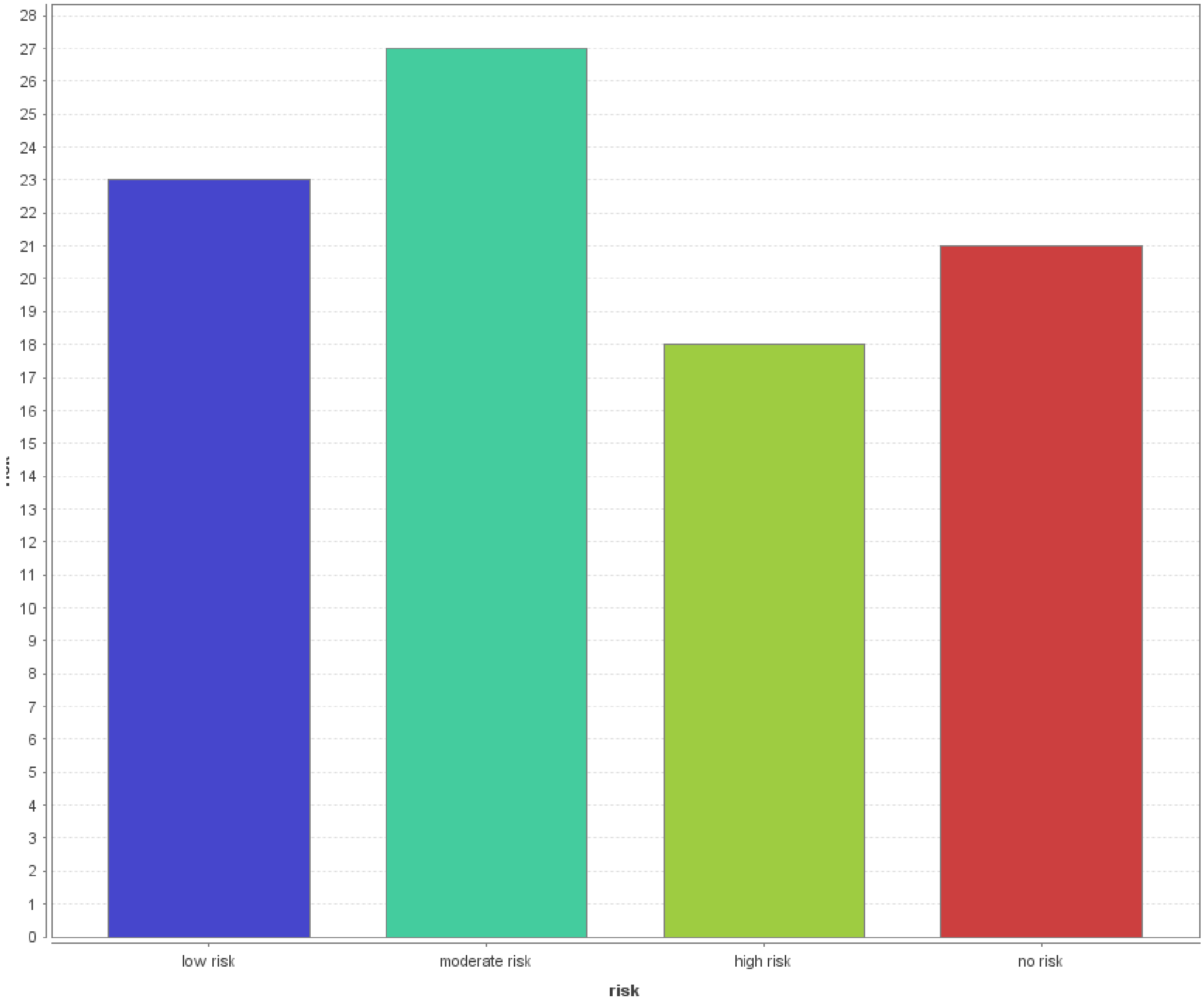


Figure 4.2: Gives a description of the number of patients with their respective risk of typhoid from the 89 patient records selected for model formulation and validation which were stored in the typhoid-training.arff file.

Table 4.1: Description of Historical Data of all 89 respondents

Variables	Labels	Frequency	Percentage
Age	0-15	8	65.2
	15-25	58	19.1
	26-40	17	9
	41-60	6	6.7
Sex	Male	52	58.4
	Female	37	41.6
Religion	Christian	69	79.3
	Muslim	18	20.7
Ethnic group	Yoruba	53	61.6
	Igbo	16	18.6
	Hausa	8	9.3
	Delta	3	3.5
	Edo	1	1.2
	Efik	1	1.2
	Burundi	1	1.2
	Idoma	1	1.2
	Urhobo	1	1.2
	Kogi	1	1.2
Profession	Student	30	41.1
	Public service	23	31.5
	Private sector	13	17.8
	Trader	5	6.8
	Artist	1	1.4
	Pharmacist	1	1.4

Marital status	Single	63	74.1
	Married	22	25.9
Education	University	67	76.1
	Secondary	18	20.5
	Nce	1	1
	Polytechnic	1	1
	Postgraduate	1	1
Had typhoid	Yes	56	63.6
	No	32	36.4
How long	Days	39	67.2
	weeks	19	32.8
Seek treatment	Hospital	44	69.8
	Pharmacy	8	12.7
	Puskesmas	4	6.3
	Nurse	3	4.8
	Private gp	3	4.8
	School clinic	1	1.6
Medicines	Yes	57	87.7
	No	8	12.3
Toilet at home	Always	34	41.0
	Often	25	30.1
	Sometimes	20	24.1
	Never	3	3.6
	No toilet	1	1.2

Toilet at work/school	Always	13	15.5
	Often	14	16.7
	Sometimes	15	53.6
	Never	12	14.3
Other place	Only	79	95.2
	Field	3	3.6
	Pond	1	1.2
After toilet	Always	63	74.1
	Often	13	15.3
	Sometimes	8	9.4
	Never	1	1.2
Before eating	Always	50	51.7
	Often	15	18.5
	Sometimes	13	16.0
	Never	3	3.7
After eating	Always	47	58.0
	Often	9	11.1
	Sometimes	23	28.4
	Never	2	2.5
Changing diaper	Always	39	73.6
	Often	7	13.2
	Sometimes	5	9.4
	Never	2	3.8
Coming home	Always	29	46.8

	Often	8	12.9
	Sometimes	21	33.9
	Never	4	6.5
Contact	Yes	31	36.9
	No	42	50.0
	Don’t know	11	13.1
Risk	No risk	21	23.6
	Low risk	23	25.8
	Moderate risk	27	30.3
	High risk	18	20.2

Majority of the respondents also lies in the age group of above 15-25 years (65.2%), 19.1% of the respondents within the age group of 61-40 years of age and 0-15 years having (9%) while the remaining 6.7% belonged to the age group of 41-60 years. Also, 76.1% of the respondents had university qualifications, while 18.8% of secondary school qualifications. 41.1% of the respondents were students, 31.5% were public servants while 27.4% were private sector workers. 74.1% of the respondents were single while the remaining percentage was single. Majority of the respondents were male 58.4% while 41.6% were female with most of the respondents been Yoruba (61.6%).

Based on the clinical information inferred from the respondents, the following were identified from the information provided. 63.6% of respondents had typhoid at least once in their lifetime. 74.1% of the respondents always wash their hands after using the toilets. 61.7% of respondents always wash their hands before eating. 58.0% wash their hands after eating. 73.6% of the respondents always wash their hands after changing diapers. 46.8 always wash their hands when coming home. About 50% of the respondents never had contact with a typhoid patient, 36.9% had contact with typhoid patients while 13.1% don’t know.

4.3 Results of Model Formulation and Simulation

The next stage is model formulation using the above-mentioned decision trees algorithms accessible in the Weka software after identifying the risk factors connected with pediatric SCA survival. Using the randomly selected test samples from the historical test used to train the model, the 10-fold cross validation technique was used to evaluate the performance of the developed predictive model for SCA survival. Compared to the most efficient, this method was carried out for both decision trees algorithm used with their corresponding results.

4.3.1 Results of Model Formulation and Simulation Using the C4.5 Decision Trees Algorithm

The training data were used from the statistics gathered from the participants to formulate the predictive model required to predict typhoid risk. Using the simulation area, the J4.8 decision trees algorithm was used to introduce the C4.5 decision trees algorithm to formulate the predictive model. The results of the formulation of the predictive model for the risk of typhoid using the C4.5 decision trees algorithm showed that four (4) variables were the most important risk factors of typhoid and were used by the algorithm to develop the tree that was used in formulating the predictive model for risk of typhoid using the C4.5 decision trees algorithm. The variables identified in the order of their importance were:

a. Age
b. Does the respondent wash his/her hands after using the toilet
c. Does the respondent wash his/her hands before eating
d. Contact with a typhoid patient

Based on the four (4) variables identified by the C4.5 decision tees algorithm, The predictive model for typhoid risk was developed on the basis of simulation outcomes using the WEKA simulation environment J48 decision trees algorithm. Figure 4.3 shows the decision trees that was formulated based on the four (4) variables that were proposed by the algorithm. Based on the values of the four factors recognized by the algorithm, the tree was used to deduce the set of rules suggested to determine the danger of typhoid. The rules extracted from the tree are as follows:

1. If (after toilet = always) and (before eating= always) and (age 15-25), (0-15), (26-40), (41-50) then (risk of typhoid = low);
2. If (after toilet = always) and (before eating = never) then (risk of typhoid = moderate risk);
3. If (after toilet = always) and (before eating = often) then (risk of typhoid = moderate risk);
4. If (after toilet = always) and (before eating = sometimes) then (risk of typhoid = moderate risk);
5. If (after toilet = often) and (contact = no) then (risk of typhoid = high risk);
6. If (after toilet = often) and (contact = yes) then (risk of typhoid = high risk);
7. If (after toilet = often) and (contact = don't know) then (risk of typhoid = high risk);
8. If (after toilet = sometimes) then (risk of typhoid = high risk);
9. If (after toilet = never) then (risk of typhoid = high risk);

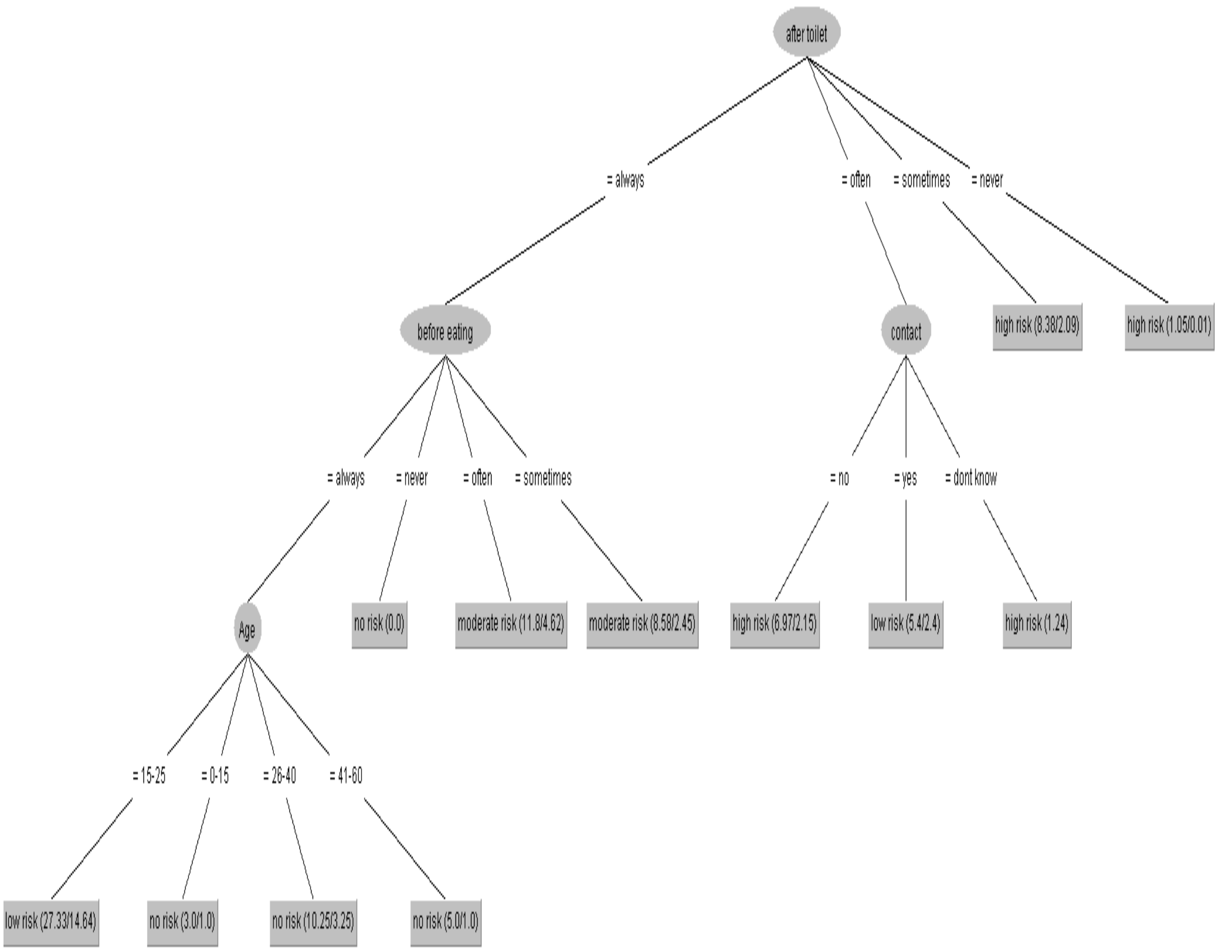
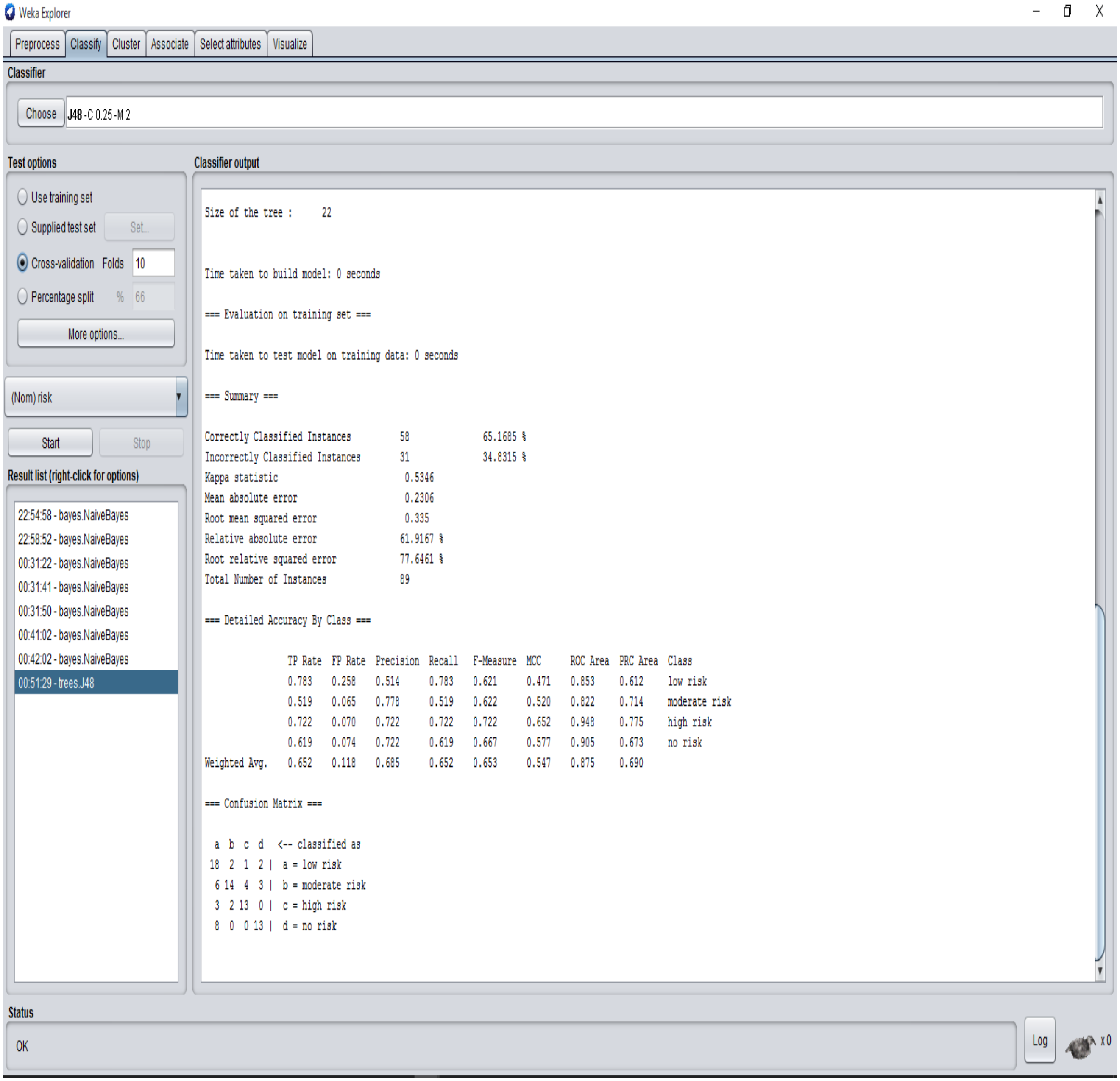


Figure 4.3: Decision Tree formulated using C4.5 for Risk of typhoid



J48 -C 0.25 -M 2

● Cross-validation Folds 10

% 66

(Nom) risk

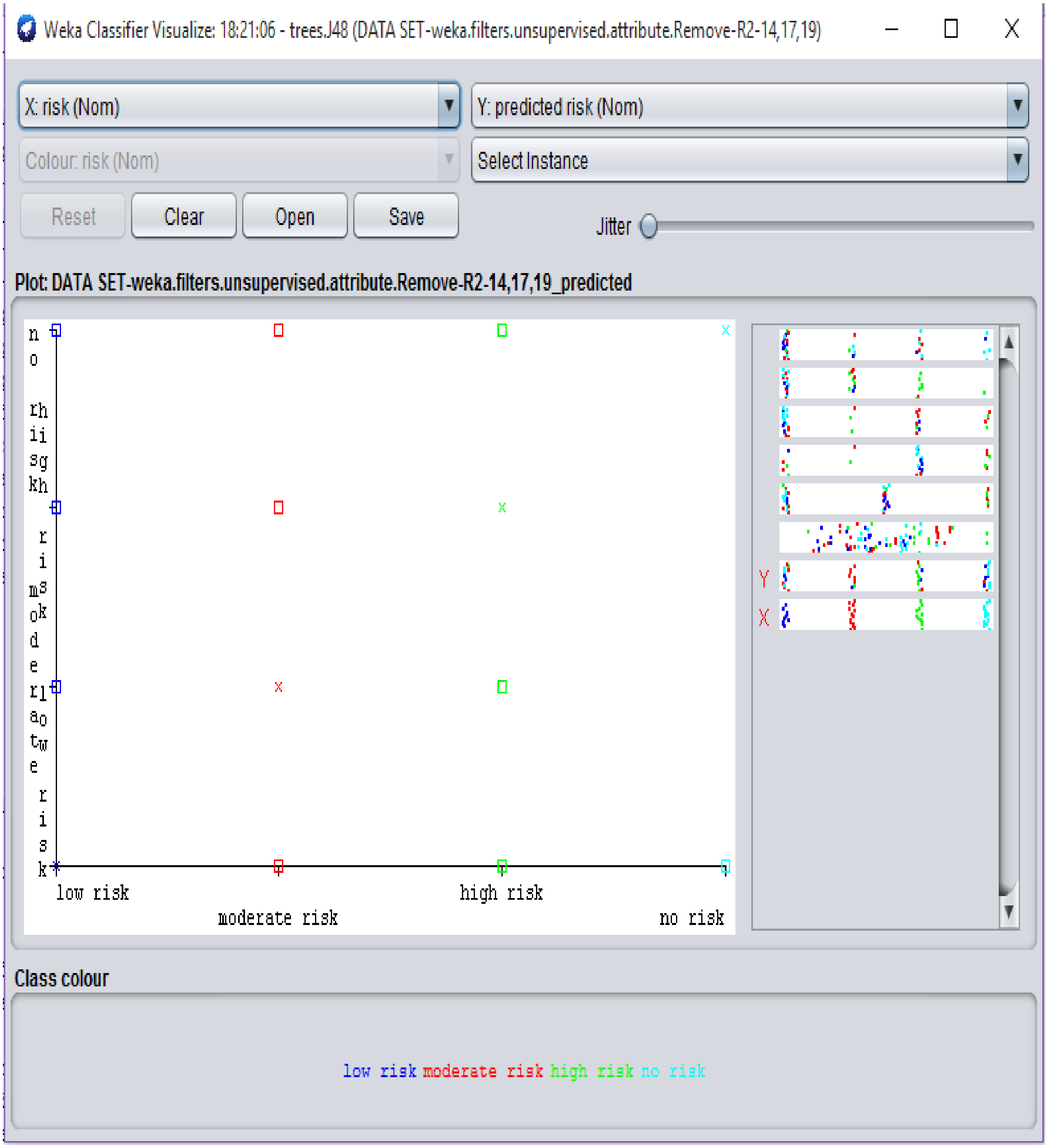
Result list (right-click for options)

- 22:54:58 - bayes.NaiveBayes
- 22:58:52 - bayes.NaiveBayes
- 00:31:22 - bayes.NaiveBayes
- 00:31:41 - bayes.NaiveBayes
- 00:31:50 - bayes.NaiveBayes
- 00:41:02 - bayes.NaiveBayes
- 00:42:02 - bayes.NaiveBayes
- 00:51:29 - trees.J48

```
Size of the tree :      22


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          58               65.1685 %
Incorrectly Classified Instances        31               34.8315 %
Kappa statistic                          0.5346
Mean absolute error                      0.2306
Root mean squared error                  0.335
Relative absolute error                 61.9167 %
Root relative squared error             77.6461 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.783    0.258    0.514      0.783    0.621      0.471    0.853     0.612     low risk
                 0.519    0.065    0.778      0.519    0.622      0.520    0.822     0.714     moderate risk
                 0.722    0.070    0.722      0.722    0.722      0.652    0.948     0.775     high risk
                 0.619    0.074    0.722      0.619    0.667      0.577    0.905     0.673     no risk
Weighted Avg.    0.652    0.118    0.685      0.652    0.653      0.547    0.875     0.690

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 18  2  1  2 |  a = low risk
  6 14  4  3 |  b = moderate risk
  3  2 13  0 |  c = high risk
  8  0  0 13 |  d = no risk
```

OK

Figure 4.4: Screenshot of C4.5 decision trees results on dataset



Weka Classifier Visualize: 18:21:06 - trees.J48 (DATA SET-weka.filters.unsupervised.attribute.Remove-R2-14,17,19)

X: risk (Nom)

Y: predicted risk (Nom)

Colour: risk (Nom)

Select Instance

Plot: DATA SET-weka.filters.unsupervised.attribute.Remove-R2-14,17,19_predicted

Class colour

low risk moderate risk high risk no risk

Figure 4.5: Screenshot of correct and incorrect classifications made by C4.5

Following the simulation of the predictive model for typhoid danger using the C4.5 decision trees algorithm, the performance assessment of the model after inner validation was registered using the 10-fold cross validation technique. Figure 4.6 shows The confusion matrix used to represent the outcomes of the true positive and negative as well as the false positive and negative validation outcomes. The confusion matrix shown in figure 4.6 was used to evaluate the performance of the predictive model for risk of typhoid.

From the confusion matrix shown in figure 4.6, the following sections present the results of the model's performance. Out of the 23 low cases, there were 18 correct classifications with 2 misclassified as moderate risk, 1 for high risk and 2 for no risk; out of the 27 moderate risk cases, there were 14 correct classifications with 6 misclassified as low risks, 4 misclassified for high risk, and 3 misclassified for no risk, out of the 18 high risk cases, there were 13 correct classifications with 3 misclassified as low risks, 2 misclassified for moderate risk, out of the 21 no risk cases, there were 13 correct classifications with 8 misclassified as low risks,. Therefore, there were 58 correct classifications out of the 89 records considered for the model development owing for an accuracy of 65.168%. Table 4.2 shows the summary of the evaluation results.

Low risk	Moderate risk	High risk	No risk	
18	2	1	2	Low risk
6	14	4	3	Moderate risk
3	2	13	0	High risk
8	0	0	13	No risk

Figure 4.6: confusion matrix of performance evaluation using C4.5

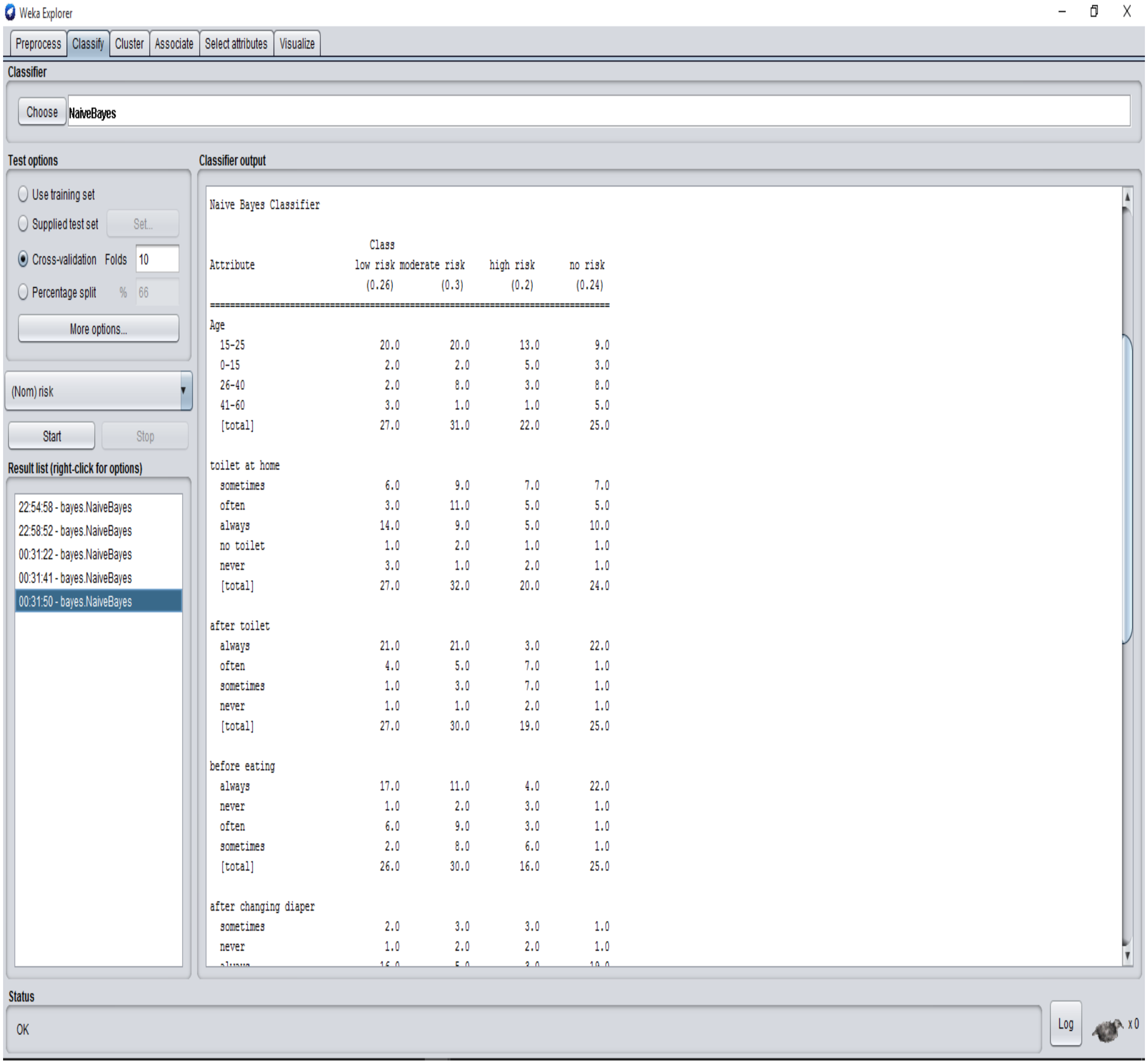
Table 4.2: Summary of the results of performance evaluation using C4.5

Performance Metrics	**Risk Labels**	**Values**	**Average**
TP rate (sensitivity/recall)	Low risk Moderate risk High risk No risk	0.783 0.519 0.722 0.619	0.652
FP rate (false alarm rate)	Low risk Moderate risk High risk No risk	0.258 0.065 0.070 0.074	0.118
Precision	Low risk Moderate risk High risk No risk	0.514 0.778 0.722 0.722	0.685
Roc Area	Low risk Moderate risk High risk No risk	0.853 0.822 0.948 0.905	0.690

4.4 Results of Model Formulation and Simulation of Naïve Bayes' Classifier

Following the formulation of the predictive model for the risk of typhoid, the next phase was model formulation using naïve Bayes' classifier algorithm available in the Weka software. The 10-fold cross validation technique was used in evaluating the performance of the developed predictive model for typhoid risk using the historical dataset used for training the model. This process was performed and compared with the performance of the predictive model developed using the C4.5 decision trees algorithm for the most effective. From the dataset collected from the respondents, the training data was used for the formulation of the predictive model needed for the prediction of the risk of typhoid. The naïve Bayes' classifier was used for the formulation of the predictive model using the simulation environment

Following the simulation of the predictive model for risk of typhoid using the naïve Bayes' classifier, the evaluation of the performance of the model following validation using the 10-fold cross validation method was recorded. Figure 4.7 shows a screenshot of the outcomes of Naive Bayes classification algorithm projections for 89 cases of information gathered from clients deemed for this research. The figures shows the correct and incorrect classifications made by the algorithm while Figure 4.8 Shows a graphical display of the Naive Bayes classifier model projections on the dataset. In figure 4.8, each typhoid class is shown using a specific color and each correct classification is shown with a star while each misclassification is shown as a square. The results presented in figure 4.8 the performance of the Naive Bayes classifier algorithm was evaluated and thus the confusion matrix was determined.



NaiveBayes

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

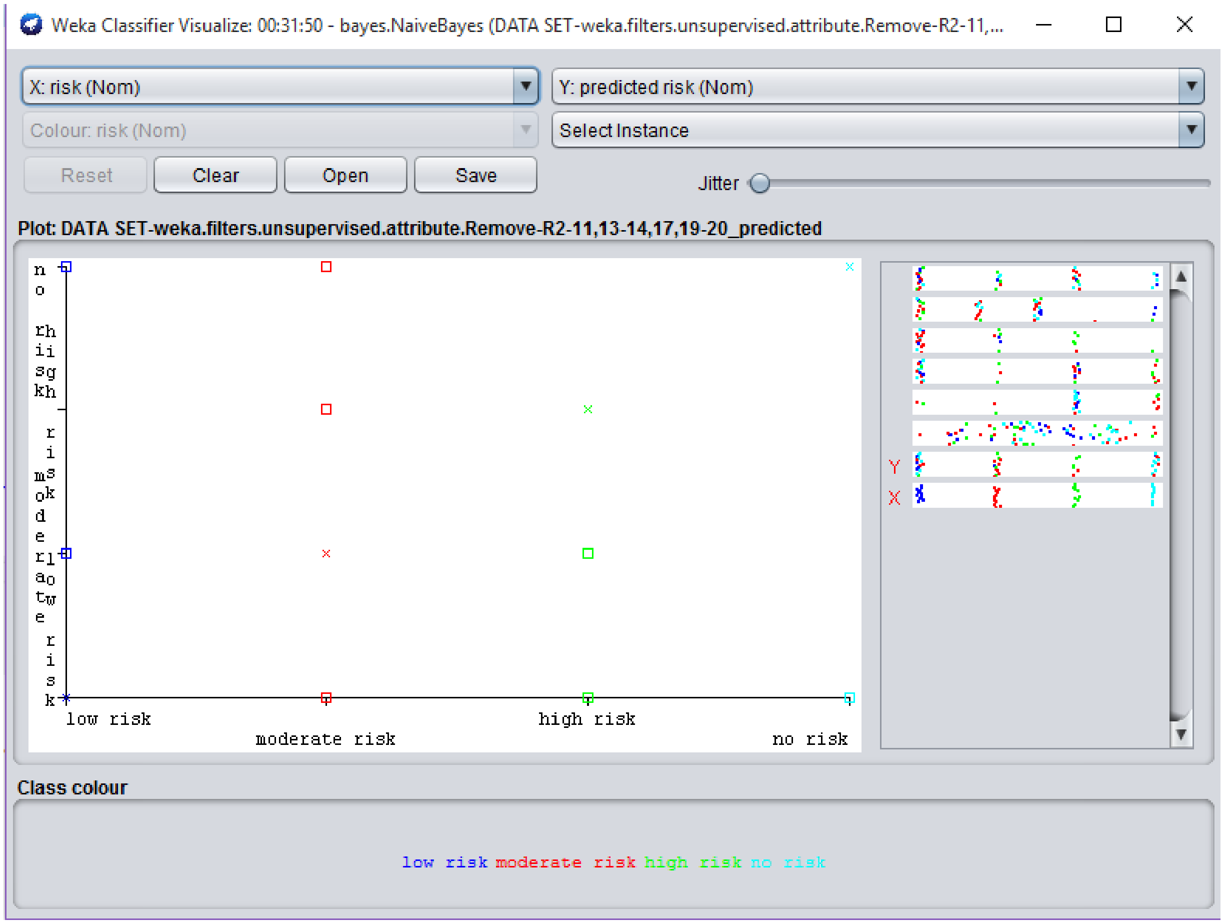
(Nom) risk

22:54:58 - bayes.NaiveBayes
22:58:52 - bayes.NaiveBayes
00:31:22 - bayes.NaiveBayes
00:31:41 - bayes.NaiveBayes
00:31:50 - bayes.NaiveBayes

```
Naive Bayes Classifier

                              Class
Attribute                  low risk moderate risk    high risk      no risk
                             (0.26)         (0.3)        (0.2)       (0.24)
============================================================================
Age
  15-25                        20.0          20.0         13.0          9.0
  0-15                          2.0           2.0          5.0          3.0
  26-40                         2.0           8.0          3.0          8.0
  41-60                         3.0           1.0          1.0          5.0
  [total]                      27.0          31.0         22.0         25.0

toilet at home
  sometimes                     6.0           9.0          7.0          7.0
  often                         3.0          11.0          5.0          5.0
  always                       14.0           9.0          5.0         10.0
  no toilet                     1.0           2.0          1.0          1.0
  never                         3.0           1.0          2.0          1.0
  [total]                      27.0          32.0         20.0         24.0

after toilet
  always                       21.0          21.0          3.0         22.0
  often                         4.0           5.0          7.0          1.0
  sometimes                     1.0           3.0          7.0          1.0
  never                         1.0           1.0          2.0          1.0
  [total]                      27.0          30.0         19.0         25.0

before eating
  always                       17.0          11.0          4.0         22.0
  never                         1.0           2.0          3.0          1.0
  often                         6.0           9.0          3.0          1.0
  sometimes                     2.0           8.0          6.0          1.0
  [total]                      26.0          30.0         16.0         25.0

after changing diaper
  sometimes                     2.0           3.0          3.0          1.0
  never                         1.0           2.0          2.0          1.0
  always                       16.0           5.0          3.0         10.0
```

OK

Figure 4.7: Screenshot of naïve Bayes' classifier results on dataset



Weka Classifier Visualize: 00:31:50 - bayes.NaiveBayes (DATA SET-weka.filters.unsupervised.attribute.Remove-R2-11,...

X: risk (Nom)

Y: predicted risk (Nom)

Colour: risk (Nom)

Select Instance

Plot: DATA SET-weka.filters.unsupervised.attribute.Remove-R2-11,13-14,17,19-20_predicted

Figure 4.8: Screenshot of correct and incorrect classifications made by naïve Bayes' classifier

Figure 4.9 Shows the confusion matrix used to interpret the results of the true positive and negative alongside the false positives and negatives of the validation results. The confusion matrix shown in figure 4.9 was used to evaluate the performance of the predictive model for risk of typhoid. From the confusion matrix shown in figure 4.9, the following sections present the results of the model's performance. Out of the 23 low cases, there were 17 correct classifications with 3 misclassified as moderate risk, 0 for high risk and 3 for no risk; out of the 27 moderate risk cases, there were 15 correct classifications with 5 misclassified as low risks, 2 misclassified for high risk, and 5 misclassified for no risk, out of the 18 high risk cases, there were 11 correct classifications with 2 misclassified as low risks, 5 misclassified for moderate risk, out of the 21 no risk cases, there were 14 correct classifications with 7 misclassified as low risks,. Therefore, there were 57 correct classifications out of the 89 records considered for the model development owing for an accuracy of 64.0449%.

4.5 Discussion of Results

The result of the performance evaluation of the C4.5 and naïve Bayes' algorithms are presented in Table 4.3. The true positive rate which gave a description of the proportion of actual cases that was correctly predicted which showed values of 0.783, 0.519, 0.722 and 0.619 respectively for no, low, moderate and high risk cases by the C4.5 decision trees algorithm and 0.739, 0.556, 0.611 and 0.667 for the naïve Bayes classifier. Thus, the decision trees algorithm showed equal capacity to correctly classify the actual no and high cases of typhoid better than the moderate and low risk cases while the naïve Bayes' classifier had the ability to correctly classify the no risk cases better than the C4.5 decision trees but not as good as C4.5 for the other risks of typhoid.

Figure 4.9: confusion matrix of performance evaluation using naïve Bayes

Low risk	Moderate risk	High risk	No risk	
17	3	0	3	Low risk
5	15	2	5	Moderate risk
2	5	11	0	High risk
7	0	0	14	No risk

Table 4.3: Summary of the results of performance evaluation using C4.5 and naïve Bayes' classifiers

Performance Metrics	**Risk Labels**	**C4.5 DT**	**Naïve Bayes**
TP rate (sensitivity/recall)	Low risk Moderate risk High risk No risk	0.783 0.519 0.722 0.619	0.739 0.556 0.611 0.667
FP rate (false alarm rate)	Low risk Moderate risk High risk No risk	0.258 0.065 0.070 0.074	0.212 0.129 0.028 0.118
Precision	Low risk Moderate risk High risk No risk	0.514 0.778 0.722 0.722	0.548 0.652 0.846 0.636
Roc Area	Low risk Moderate risk High risk No risk	0.853 0.822 0.948 0.905	0.886 0.849 0.940 0.930

The false positive rate which gave a description of the proportion of predicted cases that was incorrectly classified showed values of 0.258, 0.065, 0.070 and 0.074 for the no, low, moderate and high risk cases respectively for the C4.5 decision trees algorithm while the naïve Bayes' classifier had values of 0.212, 0.129, 0.28 and 0.0118 respectively for no, low, moderate and high risk cases. The naïve Bayes' classifier was able to correctly distinguish moderate cases better than the C4.5 decision trees algorithm and showed equal capacity to distinguish no and high cases as equal as the C4.5 decision trees algorithm.

The precision which gave a description of the proportion of the predicted cases that was correctly classified showed values of 0.514, 0.778, 0.722 and 0.722 for no, low, moderate and high cases respectively while the naïve Bayes' classifier showed values of 0.548, 0.652, 0.846 and 0.636respectively for no, low, moderate and high risk cases. The naïve Bayes' classifier was able to provide better prediction results of the risk of typhoid compared to the C4.5 decision trees algorithm based on the precision.

In general, the C4.5 decision trees algorithm was able to predict the risk of typhoid than the naïve Bayes' classifier in addition to the identification of the relevant variables that can be used for the early detection of typhoid. The number of variables identified by the C4.5 decision trees algorithm was three (4) which shows more capacity of determining the risk of typhoid.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

In order to classify the risk of typhoid in chosen respondents for this research, this research concentrated on developing a prediction model using recognized risk variables. Historical data on the distribution of typhoid risk among participants were gathered using questionnaires after professional medical professionals identified linked typhoid risk variables.

Using C4.5 decision trees and naïve Bayes ' classifier algorithm, the dataset comprising data on the risk variables recognized and gathered from the participants was used to formulate predictive models for typhoid risk. Using the WEKA software, the predictive model design was developed and simulated using the algorithms.

The research findings disclosed the factors that were recognized as appropriate to assessing the risk of typhoid in participants by the decision trees algorithm. The algorithm for decision trees was observed to show a better accuracy compared to the classification of the naïve Bayes using the training data described in the research.

5.2 Conclusion

The study presented a predictive model of typhoid risk using relevant risk factors selected from a predefined set of typhoid risk factors in Nigerians using the C4.5 decision trees algorithm that outperformed the performance of the classification of the naïve Bayes. Four (4) risk factors were recognized in the C4.5 decision trees algorithms. The predictive model was formulated using the variables identified by the C4.5 decision trees and compared to the classification proposed by naïve Bayes for this study and the performance evaluation showed that the model developed by the C4.5 decision trees was able to predict the risk of typhoid with an accuracy of 65.168 percent compared to 64.0449 percent of the classification of naïve Bayes.

5.3 Recommendations

A stronger understanding of the connection between the characteristics appropriate to typhoid risk was suggested following the creation of the forecast model for typhoid risk classification. The model can also be incorporated into the current Health Information System (HIS) that captures and manages clinical data that can be supplied to the predictive model of typhoid risk classification, thus enhancing clinical choices influencing typhoid risk and evaluating clinical data that affects typhoid risk from distant places in real time. It is recommended that a continuous evaluation of factors monitored for typhoid risk be conducted to improve the amount of data appropriate to the creation of an enhanced forecast model for typhoid risk classification using the model suggested in this research.

5.4 Limitations of the Study

The data gathered for this research is restricted to information gathered from people in the western portion of Nigeria. The Study assessed serological trials against PCR for typhoid fever (Polymerase chain reaction is a method widely used in molecular biology to make numerous copies of a particular segment of DNA) as a study of reference. Although laboratory tests are vital to confirm this severe disease, Low bacterial load and low concentrations of specific antibodies in the blood of typhoid patients coupled with their acute personality make the interpretation of laboratory testing complicated. For this study only the classification algorithm was used of which only two classification algorithm were used: the c4.5 decision tree algorithm and the naïve bayes algorithm respectively.

REFRENCES

Alpaydin, E. (1997), Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, p. 115–132.

Altman, D. G., Vergouwe, Y and Royston, P. (2009). Prognosis and Prognostic research: validating a prognostic model. *BMJ 338*: 605.

Anderson, J. A., and Davis, J., An introduction to neural networks. MIT, Cambride, 1995.

Anna E. Newton (2014). "3 Infectious Diseases Related To Travel". CDC health information for international travel 2014 : the yellow book. ISBN 9780199948499. Archived from the original on 2015-07-02.

Apte S. M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. (1997) Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsaptewe issue_with_cover.pdf,.

Aqueel, A. and Shaikh, A. H. (2012). Data Mining Techniques to find out Heart Diseases: An Overview. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 1(4): 1 – 6.

Ashraf, M., Chetty, G. and Tran, D. (2013). Feature selection techniques on thyroid, hepatitis and breast cancer datasets. *International Journal on data mining and intelligent information technology 3*(1): 1 -8.

Bangladesh: A Spatial and Time-Series Approach. PLoS Negl Trop Dis 7(1): e1998. doi:10.1371/journal.pntd.0001998

Barquist L, Langridge G. C, Turner D. J., et al. (2013). A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res*; **41:** 4549–64.

Bellazzi, R., and Zupan, B. (2008), Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inform. 77:81–97, Bramer, M. (2007). Principles of data mining: Springer

Buckle G. C., Walker CL, Black RE. (2010). Typhoid fever and paratyphoid fever: systematic review to estimate global morbidity and mortalityfor. *J Glob Health* 2012; **2:** 10401.

Cao, X., Maloney, K.B. and Brusic, V. (2008). Data mining of cancer vaccine trials: a bird's-eye view. *Immunome Research 4*:7 - 11. DOI:10.1186/1745-7580-4-7.

Chart H, Cheesbrough J & Waghorn D. (2000). The serodiagnosis of infection with Salmonella typhi. Journal of Clinical Pathology 53, 851–853.

Chatham-Stephens, K; Medalla, F; Hughes, M; Appiah, GD; Aubert, RD; Caidi, H; Angelo, KM; Walker, AT; Hatley, N; Masani, S; Nash, J; Belko, J; Ryan, ET; Mintz, E; Friedman, CR (11 January 2019). "Emergence of Extensively Drug-Resistant Salmonella Typhi Infections among Travelers to or from Pakistan - United States, 2016-2018". MMWR. Morbidity and mortality weekly report. 68 (1): 11–13. doi:10.15585/mmwr.mm6801a3. PMID 30629573.

Chen, J., Greiner, R. (1999). Comparing Bayesian Network Classifiers. In Proceedings of UAI-99: 101–108.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristics curve in risk prediction. *Circulation 115*: 928 – 935.

Corner R, Hashizume M, Ongee E.T. (2013) Typhoid Fever and Its Association with Environmental Factors in the Dhaka Metropolitan Area of

Crump J, Luby S, Mintz E. (2004) The global burden of typhoid fever. Bulletin of the World Health Organization. ;82:346–353. [PMC free article] [PubMed] [Google Scholar]

Crump J, Youssef F, Luby S, et al. (2003) Estimating the incidence of typhoid fever and other febrile illnesses in developing countries. Emerging Infectious Diseases. ;9:539–544. [PMC free article] [PubMed] [Google Scholar]

Crump J, Youssef F, Luby S. (2003) Estimating the incidence of typhoid fever and other febrile illnesses in developing countries. Emerging Infectious Diseases 9, 539–544.

Dimitoglou, G., Adams, J.A. and Jim, C.M. (2012). Comparison of the C4.5 and a naïve bayes classifier for the prediction of lung cancer survivability. *Journal of Computing 4*(8): 1 – 13. N.B:

Field, M. J. and Lohr, K. N. (2005). Clinical Practice Guidelines: Direction for a New Program. Institute of Medicine, Committee on Clinical Practice Guidelines. Washington, DC. National Academy Press.

Gasem M. H., Dolamans W. M., Keuter M.M. , Djokomoeljanto R. R. (2001). Poor food hygiene and housing as risk factors for typhoid fever in Semarang, Indonesia. Trop Med Intl Health. ;6(6):484–490. doi: 10.1046/j.1365-3156.2001.00734.x. [PubMed] [CrossRef]

Gauda, R. and Chahar, V. (2013). A comparative study on feature selection using data mining tools. *International Journal of advanced research in computer science and software engineering 3*(9): 26 – 33.

Goharian & Grossman. (2003). Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DMClassification.pdf,.

Graham S, Molyneux E, Walsh A, Cheesbrough J, Molyneux M&Hart C. (2000). Nontyphoidal Salmonella infections of children in tropical Africa. The Pediatric Infectious Diseases Journal 19, 1189–1196.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *Am J Public Health 79*: 340 – 349.

Hagerty, R. G., Butow, P.N. and Ellis, P.M. (2005). Communicating prognosis in cancer care: a systematic review of the literature. *Ann Oncol 16*: 1005 – 1053.

Health Grades, Inc. (2007). The Fourth Annual HealthGrades Patient Safety in American Hospitals Study.

Hemingway, H., Riley, R. D. and Altman, D.G. (2009). Ten steps towards improving prognostic research. *BMJ 339*: 4184.

Hosmer, D. W., Hosmer, T. and Le Cessie, S. (1997). A comparison of goodness-offit tests for the logistic regression model. *Stat Med 16*: 965 – 980

House D, Wain J, Ho VA, et al. (2001)Serology of typhoid fever in an area of endemicity and its relevance to diagnosis. *J Clin Microbiol*;**39:** 1002–07.

Ibrahim, J. G., Chu, M. and Chen, M.H. (2012). Missing data in clinical studies: issues and methods. *J Clin. Oncol. 30*: 3297 – 3303.

Jing-Song, L., Hai-Yan, Y. and Xiao-Guang, Z. (2011). Data Mining in Hospital Information System. In New Fundamental Technologies in Data Mining, Prof. Kimito Funatsu (Ed.). ISBN: 978-953-307-547-1. Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/data-miningin-hospital-information-system.

Kaambwa, B., Bryan, S., Bilingham, L. (2012). Do the methods used to analyze missing data really matter? An examination of data from an observational study of intermediate care patients. *BMC Res Notes 5*: 330.

Kanungo S, Dutta S, Sur D. Epidemiology of typhoid and paratyphoid fever in India. J Infect Dev Count. 2008;2(6):454–460. [PubMed]

Karkey A, Arjyal A, Anders KL, Boni MF, Sabina D, Koirala S, My PVT, Nga TVT, Clements ACA, Holt K, Duy PT, Day JN, Campbell JI, Dougan G, Dolecek C, Farrar J, Basnyat B, Baker S. (2010). The burden and characteristics of enteric fever at a healthcare facility in a densely populated area of Kathmandu. PloSOne. ;5(11):e13988. [PMC free article] [PubMed]

Khan K. H. (2012). Recent trends in typhoid research—A Review. *Int J Biosci*; **2:** 110–20.

Khanam F, Sheikh A, Sayeed MA, et al. (2013). Evaluation of a typhoid/paratyphoid diagnostic assay (TPTest) detecting anti-Salmonella IgA in secretions of peripheral blood lymphocytes in patients in Dhaka, Bangladesh. *PLoS Negl Trop Dis*; **7:** e2316.

Kohn, L. T., Corrigan, J. M. and Donaldson, M. S. (2000). To err is human: building a safer health system. Washington, D.C.: National Academy Press.

Liaw, A. and Weiner, M. (2012). Classification and Regression Trees by random forest. *R. News 2*: 18 – 22.

Lozano R, Naghavi M, Foreman K, et al. (2012). Global and regionalmortality from 235 causes of death for 20 age groups in 1990 and2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012; **380:** 2095–128.

Dunham M. H. (2003). "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc.

Mathur R, Oh H, Zhang D, et al. (2012) A mouse model of *Salmonella* typhi infection. *Cell*; **151:** 590–602.

McGregor, C. Christina and J. Andrew. (2012). "A process mining driven framework for clinical guideline improvement in critical care", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). http://ceur-ws. org, vol. 765.

Miller, M. and Kearney, N. (2004). Guidelines for Clinical Practice: Development, Dissemination and Implementation. International Journal of Nursing Studies, 41 (7), 813 - 827.

Milligan, R; Paul, M; Richardson, M; Neuberger, A (2018). "Vaccines for preventing typhoid fever". The Cochrane Database of Systematic Reviews. 5: CD001261. doi:10.1002/14651858.CD001261.pub4. PMID 29851031.

Musen, M. A. (1997). Modeling of Decision Support. Handbook of medical informatics. Bemmel, J.H.V. and Musen, M.A. (Eds.) Houten: Bohn Stafleu Van Loghum.

Nagashetty K, Channappa ST, Gaddad SM. (2010). Antimicrobial susceptibility of Salmonella Typhi in India. J Infect Dev Count. ;4(2):070–073.

Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. Infect. Control Hosp. Epidemiol. 25(8):690–695,.

Osheroff, J. A., Teich, J.M. and Middleton, B.F. (2006). A Roadmap for National Action on Clinical Decision Support. American Medical Informatics Association. Retrieved from http://www.amia.org/inside/initiatives/cds/ on June 23, 2016.

Parry C. M., Wijedoru L, Arjyal A, Baker S. (2011). The utility of diagnostictests for enteric fever in endemic locations. *Expert Rev Anti Infect Ther*; **9:** 711–25.

Pencina, M. J., D‘Agostino, R. B. and Demier, O.O. (2012). Novel metrics for evaluating improvement in discrimination: net classification and integrated discrimination improvement for normal variables and nested models. *Stat Med 31*: 101 – 113.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning 1*: 81-106

Romeo, M., Burden, F., Quinn, M., Wood, B., and McNaughton, D. (1998). Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. Cell. Mol. Biol. (Noisy-le-Grand, France) 44(1):179,.

Royston, P., Moons, K.G. and Altman, D.G. (2009). Prognosis and prognostic research: developing a prognostic model. *BMJ 338*: 604.

Sabbagh S. C., Forest C. G, Lepage C, Leclerc J. M, Daigle F. (2010). So similar, yet so different: uncovering distinctive features in the genomes of *Salmonella enterica* serovars Typhimurium and Typhi. *FEMS Microbiol Lett*; **305:** 1–13.

Sabbagh S. C., Lepage C, McClelland M, Daigle F. (2012). Selection of*Salmonella enterica* serovar Typhi genes involved during interactionwith human macrophages by screening of a transposon mutant library. *PLoS One*; **7:** e36643.

Sharma P. K., Ramakrishnan R, Hutin Y, Manickam P, Gupte M. D. (2009). Risk factors for typhoid in Darjeeling, West Bengal, India: evidence for practical action. Trop Med Intl Health.;14(6):696–702. [PubMed]

Sharma, A., and Roy, R. J. (1997). Design of a recognition system to predict movement during anesthesia. IEEE Trans. Biomed. Eng.44(6):505–511,.

Shillabeer, A. and Roddick, J (2007). Establishing a Lineage for Medical Knowledge Discovery. *ACM International Conference Proceeding Series 311*(70): 29-37.

Singal, A. G., Mukherjee, A. and Higgins, P. D. (2013). Machine Learning Algorithms outperform conventional regression models in identifying risk factors for hepatocellular carcinoma in patients with cirrhosis. *Am J. Gastroenterol 108*: 1124 – 1130.

Starkl M, Brunner N, Stenstrom T. A. (2013) Why do water and sanitation systems for the poor still fail? Policy analysis in economically advanced developing countries. *Environ Sci Technol*; **47:** 6102–10

Steyerberg, E. W., Vickers, A. J. and Cook, N. R. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology 21*: 128 – 132.

Waijee, A. K., Joyce, J. C. and Wang, S. J. (2010). Algorithms outperform metabolite tests in predicting response of patients with inflammatory bone disease to thiopurines. *Clin Gastroenterol Hepatol 8*: 143 – 150

Wain, J; Hendriksen, R. S; Mikoleit, M. L; Keddy, K. H; Ochiai, R.L (2015). "Typhoid fever". Lancet. 385 (9973):1136–45. doi:10.1016/s0140-6736(13)62708-7. PMID 25458731..

Wang L. X., Li X.J, Fang L.Q., Wang D.C., Cao W.C., Kan B. (2012). Association between the incidence of typhoid and paratyphoid fever and meteorological variables in Guizhou, China. Chinese Med J. ;125(3):455–460. [PubMed]

Whitaker J. A., Franco-Pardes C, del Rio C, Edupuganti C. (2009). Rethinking typhoid fever vaccines: implications for travellers and people living in highly endemic areas. J Travel Med. ;16(1):46–52. doi: 10.1111/j.1708-8305.2008.00273.x. [PubMed] [CrossRef]

APPENDIX I

Questionnaire administered to respondents for the risk of Typhoid

MOUNTAIN TOP UNIVERSITY
COLLEGE OF BASIC AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

QUESTIONNAIRE ON PREDICTIVE MODEL FOR THE RISK OF TYPHOID

Dear Respondent,
This questionnaire is designed to assist in the development of a predictive model for the risk of typhoid. Your cooperation in responding to the items will be highly appreciated. All information you supply will be used for this study only and treated with utmost confidentiality.
Thank you for your valuable time, attention and cooperation.

Yours Faithfully,
Researchers

A. *Personal Details*

1. Age: Below 15years □ 15–25 years □ 26–40 years □ 41-60 years □ above 60 years □
2. Sex

Male □

Female □

3. Religion: Christian □ Muslim □ others: ________________

    a. Ethnic group: Yoruba □ Hausa □ Igbo □others: ___________________

4. Profession: trader □ private sector □ public service □ artisan □ others _______
5. Marital Status: Single □ Married □
6. Level of education:

None □

Primary □

Secondary □

University □

Others ________________

B. Current & past illness

7. Ever had typhoid: yes □ no□
   a. How long did your typhoid last : ___days□ ___ weeks□)

8. Where did you seek treatment?

Hospital □

Puskesmas (Primary Health Centre) □
Private GP □
Nurse / Paramedic □
Shaman □
Pharmacy □
Other: ______________________________

9. Did you take any medicines?

Yes □

No □

C. Sanitation

10. How often do you use the latrine/toilet at home?

Always □

Often □

Sometimes □

Never □

We don't have a latrine □

11. How often do you use the latrine/toilet at school/work?

Always □

Often □

Sometimes □

Never □

I don’t go to school / work or there is no latrine at school/work □

12. Which other places do you use for defecating?

I only use latrine or toilet □

Field □

Pond/river/canal □

13. When do you wash your hand during the day?

After toilet □ How often: Always □

Often □

Sometimes □

Never □

Before eating □ How often: Always □

Often □

Sometimes □

Never □

After eating □ How often: Always □

Often □

Sometimes □

After changing diaper□ How often: Always □

Often □

Sometimes □

Never □

When coming home □ How often: Always □

Often □

Sometimes □

Never □

D. Contact with typhoid patients

14. Did you have contact with another typhoid patient?

Yes □ What is your relation to this patients? __________________

No □

I don’t know □

E. Knowledge of typhoid fever

15. Have you ever heard of typhoid fever?

Yes □

No □

I don’t know □

16. Why does someone get typhoid fever? You can mention all the causes you know.

Don’t know □

1. ___________________________________

2. ___________________________________

3. ___________________________________

4. ___________________________________

17. What helps to prevent typhoid fever? You can mention all causes you know

Don’t know □

1. ___________________________________

2. ___________________________________

3. ___________________________________

4. ______________________________

----------------------------- FOR OFFICIAL USE ONLY ------------------------------

SECTION F – DOCTOR’S COMMENT -Risk of Typhoid (Based on the information provided in this questionnaire by the respondent, what is your inference about the risk of Typhoid?)

No Risk □ Low Risk □Moderate Risk □ High Risk □