2021

A Classification model leveraging Electronic Immunization Records to predict child immunization completion: case study - Mukono Health facility.

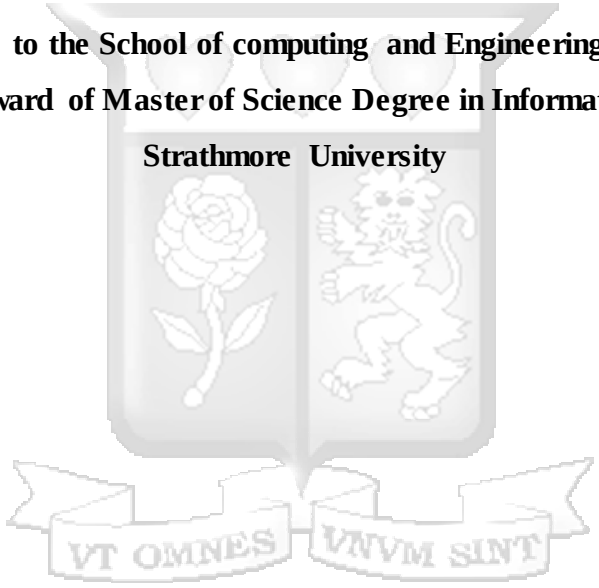
Kembabazi, Bertha
Faculty of Information Technology
Strathmore University

A classification model leveraging Electronic Immunization Records to predict Child Immunization completion.

Case study: Mukono Health facility

Bertha Kembabazi

A Thesis Submitted to the School of computing and Engineering Science in Partial Fulfilment of the Award of Master of Science Degree in Information Technology at Strathmore University



October 2021

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Bertha Kembabazi

10th October 2021

Approval

The thesis of Bertha Kembabazi was reviewed and approved by the following:

Dr. Henry Muchiri,
Lecturer, School of Computing and Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing and Engineering Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

Immunisation is one of the most cost-effective public health interventions, it prevents child deaths by strengthening immune response and preventing diseases that are not only deadly but also easily transmitted. However, countries like Uganda still face challenges that limit the attainment of immunisation completion targets like late and missed doses. It is noted that many children who start immunization do not follow through to the last dose which leads to incomplete doses hence no full protection and also missing some vaccinations which protect the child against other diseases, this in the long run exposes the child to the risk of contracting deadly diseases as well as spreading the same to others. There is potential to use data from electronic immunisation records systems to get projection insight to follow up on participants to increase access to immunisation. This study uses a random forest classification algorithm to develop a model to predict completion rates of infant immunisation to improve immunisation service delivery and utilization. This model predicts those likely to complete the recommended immunisation vaccines as per the schedule using DPT3 as an identifier classified into three categories. The categories were coded as 3 for those likely to miss, 1 for those who will receive on-time and 2 for those who are likely to receive the scheduled vaccine late. Using existing secondary electronic immunisation records data from the MyChild System implemented at Mukono district health facility, the data used was collected between 2015 and 2020. 75% of the data was used as training data while the other 25% was used as test data for the model. The predictors of this model include child dates of vaccine dose administration, the exposure to tetanus, whether a child was exposed to HIV, the date of birth and whether the caregiver was counselled. The model was tested and validated to give accurate predictions and the measure of accuracy as an output.

Keywords: Random forests, classification, prediction, immunization

Table of Contents

Abstract iii

List of figures vii

List of tables ix

Abbreviations/Acronyms x

Chapter 1: Introduction 1

1.1 Background 1

1.2 Problem Statement 3

1.3 Objectives 3

1.3.1 General Objective 3

1.3.2 Specific Objectives 3

1.3.3 Research questions 4

1.4 Justification 4

1.5 Scope and limitations 5

2 Chapter 2: Literature review 6

2.1 Introduction 6

2.2 Empirical review 6

2.2.1 Factors affecting immunization completion 6

2.2.2 Techniques used in existing Electronic Immunization Records Systems use to support immunization completion 9

2.2.3 Existing prediction models that support immunization completion 11

2.2.4 Classification models that can be developed to predict immunization completion 14

2.3 Summary of the literature review 19

2.4 Conceptual Framework 20

3 Chapter 3: Methodology 22

3.1 Introduction 22

3.2 System methodology 22

3.2.1 *Feasibility study phase* 22

3.2.2 *Business study phase* 23
3.2.3 *Functional model iteration* 23
3.2.4 *Design and build phase* 23
3.2.5 *Implementation phase* 23
3.3 Research design 23
3.4 Target population 24
3.5 Data collection 24
3.6 Data Analysis and exploration 25
3.7 Model development Methodology 25
3.8 Testing and validation 26
3.9 Research Quality 26
3.10 Ethical considerations 26
4 Chapter 4: System Analysis, Design and Architecture 27
4.1 Introduction 27
4.2 System Analysis 27
4.3 System Architecture 28
4.4 System Design 29
4.4.1 Use Case diagrams 29
4.5 System sequence diagram 32
5 Chapter 5: Implementation and validation 34
5.1 Introduction 34
5.2 System implementation 34
5.2.1 Data Pre-processing, normalization, and loading 34
5.2.2 Development environment 35
5.2.3 Random Forest Model Components 36
5.3 System Implementation 42
5.4 Model Testing and results 42

6 Chapter 6: Results and Discussions 44
6.1 Introduction 44
6.2 Factors influencing immunization completion 44
6.3 Techniques used by existing Electronic Immunization Records Systems to support immunization completion 44
6.4 Models used to predict immunization completion 45
6.5 Developing a classification model to predict immunization completion 46
6.5.1 Chapter 7: Conclusion and Recommendation 47
6.6 Conclusions 47
6.7 Recommendations 47
6.8 Future Work 48
References 49
Appendices 60
Appendix A: Data cleaning, transformation and standardization 60
Appendix B: R implementation code 71
Appendix C: Strathmore University Ethics Review Approval Certificate 74
Appendix D: Originality Summary report 75

List of figures

Figure 2.1 ARIMA Model formula extracted from (Chase, 2013). 12
Figure 2.2 A sample of KNN pseudo code (Dey, 2016) 15
Figure 2.3 A decision tree (Sarker et al., 2020) 18
Figure 2.4A random forest classifier (Machine Learning for Subsurface Characterization | Siddharth Misra, Hao Li, Jiabo He | download, 2020) 19
Figure 2.5 Conceptual framework of the proposed classification model to predict immunization completion 21
Figure 3.1 DSDM process diagram (Abrahamsson et al., no date). 22
Figure 4.1 System Architecture 29
Figure 4.2 Use case diagram 32
Figure 4.3 Sequence Diagram 33
Figure 5.1A screenshot of code and output showing data structure and dimension 34
Figure 5.2 A screenshot of the code dropping null entries and the dimension of the new dataset 35
Figure 5.3A screenshot showing the conversion to factors and binding the public ID to create a new data frame to use in model development 35
Figure 5.4 A screenshot of the code to get optimal mtry parameter and random forest model training 37
Figure 5.5 A screenshot of the output mtry 37
Figure 5.6 A screenshot of the output of the trained model classifier 37
Figure 5.7The output of the trained model 38
Figure 5.8 Screenshot of the code for variable importance 38
Figure 5.9 A screenshot of the variable importance plot 38
Figure 5.10 A screenshot of variable importance details 39
Figure 5.11 A structure of a random forest (Random Forest - an Overview | ScienceDirect Topics, n.d.-b) 39
Figure 5.12 An illustration of bootstrapping process (Resampling Methods · UC Business Analytics R Programming Guide, n.d.) 40
Figure 5.13 A structure of a random Forest(Random Forest Algorithm- An Overview | Understanding Random Forest, n.d.) 41

Figure 5.14 Decision tree structure (Decision Tree Structure (Martínez et Al., 2009) | Download Scientific Diagram, n.d.) 41
Figure 5.15 A screenshot showing the code running the model on testing data 42
Figure 5.16 A screenshot of the output of prediction on test data. 42
Figure 5.17 A screenshot of the code and output of the confusion matrix. 43

List of tables

Table 3.1 A table showing the attributes of the data collected for the study 24
Table 2.1 Study performance metrics extracted from (Chandir et al., 2018). 13
Table 4.1 Use Case Scenario 1 29
Table 4.2 Use Case Scenario 2 30
Table 5.1 Packges and libraries used in developing the model……..306

Abbreviations/Acronyms

ARIMA : Autoregressive Integrated Moving Average

BID : Better Immunization Data

DHIS District Health Information System

DPT : Diptheria, Tetanus tocoids and Pertusis vaccine

EHR : Electronic Health Records

EIRS : Electronic Immunization Records Systems

EPI: Extended Programme for Immunization

HMIS : Hospital Management Information Systems

KNN : K Nearest Neighbors

OOB: Out of Bag

OPV: Oral Polio Vaccine

PCV : Pneumococal Conjugate Vaccine

SVM : Support vector machine

TIMR : Tanzania Immunization Registry

UHC: Universal Health Coverage

UNEPI : Uganda National Expanded Programme on Immunization

UNICEF : United Nations Children's Fund

WHO : World Health Organization

Chapter 1: Introduction

1.1 Background

Complete and timely immunization is globally recognized as essential for public health as it prevents killer diseases (Mutua et al., 2011). Vaccination is estimated to avert an estimated two to three million deaths every year from polio, tuberculosis, meningococcal, pneumonia, diphtheria, tetanus, pertussis and measles among other diseases (Machingaidze et al., 2013). In Uganda, basing on the timelines of the World Health Organization with guidance from Uganda National Expanded Programme on Immunization, a child is considered to have received a full dosage if they have received BCG and polio vaccines at birth and then three doses of OPV, DPT, PCV, RotaVirus, Hepatitis B and Influenza type B vaccines at 6, 10 and 14 weeks respectively and a dose of measles vaccine at 9 months (Babirye et al., 2012a; Bbaale, 2013a). Whereas vaccination is one of the most cost-effective public health interventions (Tamirat & Sisay, 2019; Zida-Compaore et al., 2019), immunisation completion remains suboptimal across the population especially in low income countries (Wariri et al., 2019). Over the years it has not improved despite the strategies like Reach Every District, over twenty million children do not receive immunization services (WHO, 2018; World Health Organization, 2019). This immunization gap is due to different causes from the unavailability of resources, myths and beliefs and general lack of information (Malande et al., 2019b).

Electronic Immunization Records Systems like any other E-health Records solutions can improve quality of health care especially in low income countries. These systems use information and communication technologies for systematic collection, storage, retrieval, analysis and dissemination of health information (Howarth et al., 2018). To achieve information-based decisions, the highest level of process automation, reduced medical errors, cost reduction, data mining and real-time information access, information systems must be able to link patient-specific information with evidence-based knowledge to come up with fitting deductions from algorithms and models embedded in the system software (Frøen et al., 2016). According to UNICEF, health systems strengthening strategy 2016-2031, digital interventions should cater for the patient, health professionals, healthcare organizations and policymakers, that is by availing information, communication technology like SMS, patient record portals, decision support, eLearning tools, predictive analytics, shared EHRs and real time data. (UNICEF, 2015)

In the context of African countries which have disproportionately lower immunisation completion rates (Hosseinpoor et al., 2016), there is potential to leverage the full functionality of existing Electronic Immunization Records Systems. The current outstanding challenges are resource wastage, little or no follow up, non-timely vaccination and ineffective targeting strategies, limited information to the (Babirye et al., 2012b; Nakatudde et al., 2019; Nawaz et al., 2015; Ndiritu et al., 2006; Oryema et al., 2017; Rahman & Obaida-Nasrin, 2010). This result in missed doses, funding fatigue by governments and donors since immunization is not accurately presented hence doubt in the accountability.

While Uganda has recorded some high rates of completion, however, the intervention strategies, timeliness of vaccine administration and scheduling of some immunization outreach programmes have not met their desired goals (Babirye et al., 2012b; Nsubuga et al., 2019). There is still a lot of disparity of immunization completion rates in rural vs urban areas due to socio-economic factors that need to be factored in while developing intervention programmes and systems (Babirye et al., 2011b; Malande et al., 2019b; Nakatudde et al., 2019; Oryema et al., 2017) as of 2016, the missed opportunity of vaccination in rural areas was 55.5% (Adamu et al., 2019). Therefore, this study seeks to use existing EIR data to develop a classification model to predict the likelihood of completion so this can inform stakeholder decisions.

This study uses a novel model to address some of the current gaps in using Electronic Immunization Records Systems to improve immunisation coverage. The study uses a classification random forest algorithm to develop a model to predict immunization completion rates be used to boost decision support for policymakers, health care providers and health care seekers. This will inform strategies to improve access to immunization services, mitigation of dropouts and informing health system processes in terms of planning and implementation of targeted interventions to reach those who are likely to drop out. The prediction output is classified in categories with labels of *1* for those likely to be on time, *2* for those likely to be late and *3* for those who are likely to miss the DPT3 dose.

1.2 Problem Statement

The electronic records system implemented widely in Uganda is the DHIS2 however use of manual paper-based systems like the HMIS EPI is prevalent. This manual system uses tally sheets, child health cards, child register and attendance summary (Äijö et al., 2020). The underlying challenges of this mode of operation include, child follow up challenges, vaccine logistics issues, caretaker negligence (Dehnavieh et al., 2019).

To evaluate the immunization service utilization, DPT3 is used globally as the standard indicator, however between 2010 and 2016, studies have shown that low- and middle-income countries hit stagnant low completion rates in relation to utilization of DPT3 vaccine dosage (Kamanda, 2010; Ward et al., 2020). In a study conducted in 2016 to evaluate immunization intervention efficiency, it was discovered that the doses received in earlier weeks after birth recorded significantly higher attendance than the ones given later due to parental apathy, long waiting times and the exhaustion from the frequency of doses (Nsubuga et al., 2019; Okot, 2015).

In a bid to resolve some of these challenges, this study intends to use the existing information to study the patterns and trends of completion and hence predict the chances of a child missing a dose, being late for the scheduled and/or receiving the dose on time basing on DPT3 as the indicator. The immunization monitoring programme has a defaulter tracking list, this shows which child did not receive the scheduled vaccines. The prediction model could come in handy to identify the children before they default so they are prioritized.

1.3 Objectives

1.3.1 General Objective

This study aims at developing a random forest classification-based model that will support immunization follow up by predicting the likelihood of immunization completion.

1.3.2 Specific Objectives

i. To analyse the factors influencing immunization completion
ii. To review the techniques used by existing Electronic Immunization Records Systems to support immunization completion
iii. To review existing models used to predict immunization completion.
iv. To develop a classification model to predict immunization completion.

v. To test and validate the model.

1.3.3 Research questions

i. What are the factors influencing immunization completion?

ii. 2

iii. What methods are used to predict immunization completion?

iv. How can a prediction model be developed to predict immunization?

vi. How can the model be tested and validated?

1.4 Justification

This study seeks to address the issue of immunization completion gaps by leveraging the available data in the EIRS. There is a vast amount of data in health care and this study utilizes this opportunity to develop a decision support model that predicts the likelihood of a child completing their immunization schedule on time. This intervention is aimed at improving health systems strategies like follow up of children before they fall off the trail. The need for such an intervention is to be proactive in attaining Universal Health coverage by equipping stake holders with necessary information to ensure that children receive their vaccination doses on time and especially focusing on the ones at high risk of being late or missing the doses.

For health professionals, healthcare organizations and policymakers, the models developed are to predict probable immunization dropout patterns, therefore, aid in contingency planning, predict the possibility of dropouts basing on DPT3 immunization completion trends and devise means to prevent them and estimate the population for purposes of resource planning.

This study aims to benefit the general population, guardians, parents and caretakers through follow-ups made from facilities to increase participation in immunization activities, increase awareness of facts touching immunization and resource planning. In addition, timely immunization protects against deadly infections and spread of killer diseases and also prevent infant mortality.

To the research and academia field, this research is an incremental study to enhance the already existing study and implementation of health information records systems, it also creates a platform for future researchers to refer and review especially in regard to health informatics and

harnessing digital health interventions and health data to influence decisions making and contribute to sustainable development.

1.5 Scope and limitations

The scope of the proposed study will be limited to the use of pre-existing data collected by My Child solution at Mukono health facility and using the same to develop a model that predicts immunization completion in infants basing on DPT3 as an indicator. The secondary data has been pre-collected over the years and is ready for processing. Also, this is due to the limitations in other data collection methods like focused groups and interviews given the global travel and contact restrictions due to COVID19. DPT3 is used as it is an indicator in this research as a global standard of immunization utilization evaluation.

The anticipated limitations include data entry errors from the MyChild since most of the data is collected from handwritten HMIS forms and could be wrongly recorded.

2 Chapter 2: Literature review

2.1 Introduction

This chapter covers an empirical review of literature on relevant studies on immunization completion, that is the influencing factors, electronic immunization systems, past studies on immunization completion prediction and how classification models can be developed to predict the same. The study thoroughly examines the existing literature and the knowledge gaps and illustrates how the proposed model shall address the gaps.

2.2 Empirical review

2.2.1 Factors affecting immunization completion

An estimated number of 23million children are still unable to fully benefit from the available immunization interventions available across the globe but mostly in middle and low-income countries (Restrepo-Méndez et al., 2016). The reviewed literature covers the factors that affect immunization completion, assessing whether children received fully recommended doses of vaccine, partial or not vaccinated at all. The type of literature reviewed focused on studies between 2010 to 2020 in Low and middle-income countries and from this, the factors classified into two subsections as expounded below.

2.2.1.1 Participant facing factors

Children immunization status is mostly dependent on their caretakers, parents and guardians in particular, the major influence being maternal health-seeking behaviour and exposure (Babirye et al., 2012a). The mothers are primarily in charge of their children's health from pregnancy until the children can make sound independent decisions (Oryema et al., 2017). A study in Benin linked the likelihood of full immunization to be higher among children whose mothers had antenatal care visits unlike those who had not as well as those who did not deliver at the facility (Budu et al., 2021). Therefore, a mother must be aware of the recommended vaccines and schedules, this is influenced by the level of education as these are taught in school, their literacy levels that are the ability to read and comprehend from the child health cards or the media campaigns as availed in their areas and support from their spouses (Babirye et al., 2011a). There are tonnes of myths and beliefs in communities that are anti-vax; therefore, the knowledge of caretakers to discern myths from facts is important (Rahman & Obaida-Nasrin, 2010; Sullivan et al., 2010). Also, parents' knowledge about immunization depends on their health-seeking

behaviour, how much effort they put in to inquire about how and when to access a particular vaccine from the health caregivers, community health influencers or other mothers (Bbaale, 2013b; Restrepo-Méndez et al., 2016).

Also, the accessibility to immunization information and facilities backed up by income levels, health insurance access, media exposure and immunization records influence immunization completion (Malande et al., 2019c). While most vaccines are free, access to the same services is sometimes impaired by the distance from the service centre and the target community (Rahman & Obaida-Nasrin, 2010). In cases where the facility is five or more kilometres away, one may consider the mode of transport, for islanders who need to take a boat are likely to miss the opportunity if they cannot afford the fare while road users may be required a long trek. This hinders the likelihood of the underprivileged children receiving the dose in required doses (Malande et al., 2019a; Oleribe et al., 2017).

It is important to note that several studies have noted that for scheduled routine immunization programs, there is a significant disparity in attendance of those schedules in the earlier days of the child than those given later like measles and DPT3 (Martin K. Mutua et al., 2011; Martin Kavao Mutua et al., 2016). A study in Ethiopia found out that decline in turn up children grow older is usually brought about by the massive increase in responsibilities and workload of the caregiver hence likely to forget the scheduled dates let alone prioritise them if the child is healthy, (Mekonnen et al., 2020) therefore, this shows the need for follow up and constant reminders to the parent.

Therefore, the consideration of these variables while collecting and producing health data statistics, analytics, knowledge and evidence is important for contextualizing root causes of existing patterns. This study has captured a few of these but not all therefore recommends that incremental studies and collection of this data be done.

2.2.1.2 Health systems facing factors

The health systems driven by international health bodies and health ministries have made immunization available especially by availing free vaccination services, to ensure desirable percentages of immunization completion, the following factors ought to be considered.

Primarily, the major factor is information dissemination on immunization to empower the target participants with all necessary information on recommended doses and from where to get them.

This covers both the health administrators and caregivers of children. The information also has to address and 'demystify the existing myths around immunization and vaccines. Immunization cards that were promptly filled and referred to were seen to have improved immunization seeking behaviour as they outline the different vaccines given and the schedules (Bbaale, 2013b). Therefore, besides information, it should be noted that follow up is necessary to remind caregivers on upcoming immunization dates.

Health Management Information Systems and data quality are factors that influence immunization completion. The HMIS and data collected are used to inform and influence decisions facing service delivery, therefore it is notable that data that was clear, relevant and timely helped to make proactive interventions from stakeholders as opposed to reactive interventions which eventually cost resources and time (Chopra et al., 2020). An assessment study on missed opportunities of vaccination in Kenya, one of the findings to why children who had been at the facility still missed out on vaccination opportunities was due to limited integration of vaccination services with other health services (Li et al., 2020). HMIS integration with EIRs could help facility workers by automatically identifying a child's vaccination needs no matter what the reason for the visit was hence reducing the number of missed opportunities.

The health systems strategies on equity are a factor affecting immunization completion. While there are already set measures to improve immunization, the strategies that focused on increasing chances for equity in harder to reach participants increased the rates of completion. The most affected groups include rural areas, poverty-stricken areas, politically unstable and war-torn areas and calamity stricken areas, such include slums, refugee camps and off the grid difficult terrain areas (Martin K. Mutua et al., 2021; Ndiritu et al., 2006; Nsubuga et al., 2019). The chances of children in these areas to access services despite their desire and efforts are lessened compared to their counterparts therefore the strategies ought to consider increasing their odds (Chopra et al., 2020; Rainey et al., 2011). Interventions that focus on childhood immunization dropouts should aim at bringing more children to immunization facilities on-time immunization (Usman et al., 2010).

The procurement and supply chain of vaccines is an important factor to ensures that vaccines are always available in required amounts which in most cases has hindered immunization completion. In the literature studied, there are cases where children missed doses due to vaccines

being out of stock at health centres, they had showed up but could not receive the service (Malande et al., 2019a; Restrepo-Méndez et al., 2016; Vouking et al., 2019).

2.2.2 Techniques used in existing Electronic Immunization Records Systems use to support immunization completion

In this section, the search strategy was specifically on literature published between 2010 and 2020 on Sub-Saharan Africa with a focus on Uganda for the context of the scope of this study. The literature used was mostly from Global Health Science and practice, BMJ Global Health, Pan African Medical Journal, Frontiers in Public health among others. Keywords used electronic immunization registries, electronic immunization records systems, digital health and health decision support.

2.2.2.1 DHIS2

Being deployed in over 72 low and middle-income countries on a national scale, DHIS2, developed at the University of Ohio, is an open-source web-based HMIS. In Uganda, it is deployed countrywide and relies on data collected manually on the paper-based system of HMIS EPI (Sowe & Gariboldi, 2020). DHIS2 platform is used for health data management, data visualization, individual records, mobile tracking and integration with other software. The DHIS2 is commissioned for use under the Ministry of Health with support from Non-Profitable Organizations and it is mostly used at district level for reporting, monitoring and evaluation of routine health data on district level (Kiberu et al., 2014).

The features in DHIS2 that use data to support decision-making is the data management and analytics feature, which allows users to manage aggregate and routine data. These data visualization features are GIS, pivot tables, charts and dashboards (Dehnavieh et al., 2019).

While it is not entirely set up for only immunization records, the DHIS2 data is integrated with other EMRs to collect, aggregate and share data that is hence used for decision making beyond reporting. It also uses individual patient data like telephone contacts to disseminate SMS alerts (Kikoba et al., 2019; Maïga et al., 2019).

With these features, DHIS2 in immunization completion support improves visibility and monitoring of routine immunization indicators and decision making for stake holders.

2.2.2.2 My Child System/ Solution (MCS)

This system is a scanner-supported system from Shifo Foundation that uses smart paper technology to capture and process immunization data aimed at replacing the manual paper-based HMIS (Carnahan et al., 2020; *On the Way to Reducing the Workload for Ugandan Health Workers | by Shifo Foundation | Shifo News | Medium*, n.d.). Deployed in 12 districts of Uganda, the MyChild system used captured data to show timely data on registered children, fully vaccinated children. The system also has features for SMS follow up and monthly reports generated and integrated with DHIS reporting system (Bea & Heydari, 2014; *For Health Workers in Uganda, SMS Reminders to Parents Are "the Hope We Needed" | by Shifo Foundation | Shifo News | Medium*, n.d.). Other incorporated data use is reminders via SMS alerts to healthcare providers and seekers, auto-generated follow-up lists, stock balances, utilization and wastage rates, health facility key performance indicators and child immunization schedules (Äijö et al., 2020).

In an assessment study, this system showed reduced turn-around time with data collection and processing compared to the HMIS EPI process, standardized data format, fewer errors and follow up on mothers for missed immunization opportunities. While it has integration and data quality advantages, the noted pitfall was, in the case of child health registration card stock-outs, the health care providers had to resort to the manual system (Äijö et al., 2020; Sowe & Gariboldi, 2020).

2.2.2.3 TIMR/ZEIR

Under the Better Immunization Data (BID) initiative to better immunization data collection, quality and use, the Tanzanian Immunization Registry (TIMR) was deployed in for regions of Tanzania and the Zambia Immunization Registry (ZEMR) In the Southern provinces of Zambia focusing on unique patient identification, stock management and reporting on immunization routines. The data is then used to ensure that each child receives the right dosage at the right time, plan and forecast needed stock, determine due and overdue children as per the immunization schedules and reporting. The reports provided include the number of children vaccinated and categorize defaulter information, mapping with the facility health workers (Seymour et al., 2019).

According to Laurie et al 2019, the progress of these immunization programs is in three phases namely, strengthening data collection, improving data quality and increasing the use of data for decision making. In data use for decision making, health worker empowerment in defaulter follow up and stock balances. The data used to evaluate facility performance and address performance gaps. The EIR supports peer networking among health workers to share challenges and experiences. In general, there was a shift from data collection and reporting to the use of visualization tools to plan activities (Werner et al., 2019).

2.2.3 Existing prediction models that support immunization completion

In this section, the literature reviewed was based on studies made on the existing prediction models used in support of immunization coverage using existing data from EHRs in low and middle-income countries. Models used in the examined literature are those that covered studies that were carried out to predict immunization coverage related parameters including immunization dropout rates, vaccine coverage among others.

2.2.3.1 Forecasting models used to predict Immunization trends

Forecasting models use a series of past values to predict future events, (Adhikari & Agrawal, n.d.) used forecasting to predict trends in immunization patterns that would eventually affect immunization completion.

In a 2019 study, ARIMA was used to develop a model which forecasts monthly measles immunization coverage using data from 2014 to 2018. The study result proved that ARIMA model developed with the smallest normalized Bayesian Information Criterion can be used for forecasting and equipping decision makers with the basis to establish strategies, priorities and resource usage (Alegado & Tumibay, 2019). ARIMA used in a different study in India, research was performed to predict the demand for BCG a tuberculosis vaccine to secure vaccine stock based on the progress of the newborn babies in India (Kim et al., 2016).

ARIMA model is a time series model, in the Box-Jenkins methodology, which is used by both studies, it assumes that any time series pattern fits in any of the three statistical models; (p, d, q) Autoregressive (AR) p= denoting autoregressive terms number, Moving Average (MA) q= order of the moving average process and finally, d= degree of first differencing involved. ARIMA model is formed by the combination of appropriate AR and MA terms (Chase, 2013). In an Indonesia study, vaccine stock prediction, ARIMA is used to predict required stock levels of

vaccines at health centres and visualize the results to help the government in resource planning and allocation (Satya Sahisnu et al., 2020).

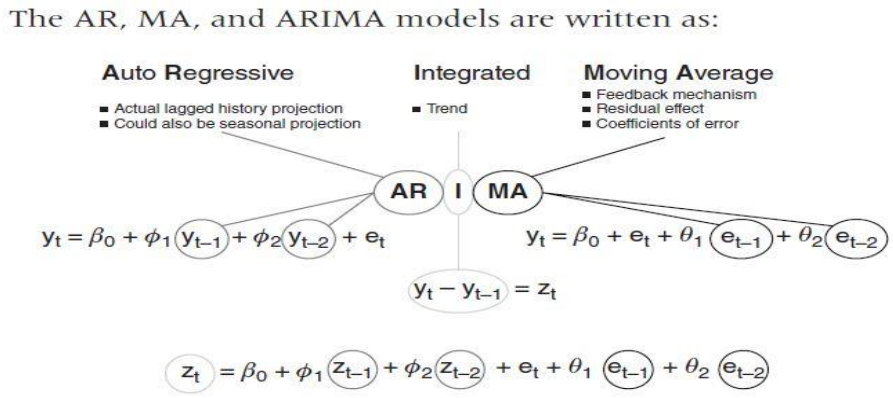


Figure 2.1 ARIMA Model formula extracted from (Chase, 2013).

ARIMA model is a time series model, in the Box-Jenkins methodology, which is used by both studies, it assumes that any time series pattern fits in any of the three statistical models: (p, d, q)

In conclusion, while these models were able to give the minimum required results, there is a gap since seasonality was not always supported since most data used required to be stationary.

2.2.3.2 Machine learning techniques encapsulation

In a study to identify children at high risk of defaulting from immunization, a digital repository with about five thousand longitudinal immunization records from India was used. Variables of this data were gender of the child, language spoken at the child's house, place of residence of the child (town or city), enrolment vaccine, timeliness of vaccination, enrolling staff (vaccinator or others), date of birth (accurate or estimated), and age group of the child. Four machine-learning techniques encapsulated in a predictive engine namely recursive partitioning, support vector machines, random forests and c forests (Chandir et al., 2018).

From this study, the results produced accuracy rates of 78.9% for recursive partitioning, 78.8% for SVMs 75.6% for random forests and 78.6% C-forest. While recursive partitioning had the highest accuracy rates, other performance metrics like sensitivity, specificity, positive predictive value, and negative predictive value were considered and the results obtained are as shown in Figure 2.2.

Performance metrics of all the study predictive models.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Negative predicted value (%)
Recursive partitioning	78.9	74.0	84.2	83.4	75.1
Support vector machines	78.8	88.9	68.0	74.9	85.1
Random forests	75.6	94.9	54.9	69.3	91.0
C-Forest	78.6	90.5	65.8	74.0	86.6

Table 2.1 Study performance metrics extracted from (Chandir et al., 2018).

The random forest model, given its high sensitivity, was able to identify the maximum number of children who would default subsequent vaccinations. On the other hand, the recursive partitioning model's high specificity at 84.2% and lowest sensitivity at 74.0%, indicated the ability to identify the highest value of children who will comply with their vaccination schedule (Chandir et al., 2018).

The identified limitations to this literature in regard to this study is the generalizability of the data, therefore, to use the model for other populations there is a need for readjustments to put other factors like languages spoken, influence of residence location and wealth index calculations into consideration. The model is also only beneficial to communities with high access to health facilities and underutilized services which is not the case of low-income countries like Uganda where availability and access to services is one of the setbacks to child immunization coverage rates (Malande et al., 2019a).

2.2.3.3 Predictive analytics and Associative Classification

In a study conducted in Pakistan, to introduce predictive analytics in the expanded programme on immunization, a model to improve coverage was proposed to identify children likely to miss any of the vaccines in the immunization schedule (Qazi et al., 2020). This study was to solve shortcomings of a previous study by Subhash et al where children were classified into two groups of defaulters and non-defaulters but as long as one dose was missed the child was labelled as likely to default (Chandir et al., 2018). This study used a dataset from a survey and for defaulter prediction and associative rules identification, 19 demographic and socioeconomic attributes were used to relate to a child's likely vaccination status (Huang & Danovaro-Holliday, 2021).

With an accuracy rate of 98%, this study used a multilayer perceptron classifier along with machine learning algorithms namely, decision trees, Support vector machine and naïve Bayes to correctly identify children at the risk of dropping out at different stages (Qazi et al., 2020).

This study informs our research on how utilizing predictive analytics can reinforce immunization programs by focusing on vulnerable participants. The researchers note that this research is not generalizable given the dataset size and number of attributes therefore contradicting results may be found when another dataset is used (Qazi et al., 2020).

2.2.4 Classification models that can be developed to predict immunization completion

In this section, we explore classification algorithms and how they can be used to achieve the aim of the study.

2.2.4.1 Naïve Bayes model

Naïve Bayes is a probabilistic classifier that employs the Bayes theorem's use of conditional probability that is observations from previous events to predict future events (Shastri et al., 2018). Bayesian theorem guides calculation of the posterior probability by the equation;

$$P(c|X) = \frac{P(c|X)P(c)}{P(X)}$$

$$P(c|x1, x2, \dots, xn) = \frac{P(x1|c)P(x2|c)\dots P(xn|c)P(c)}{P(x1)P(x2)\dots P(xn)}$$

where, P(c|X) represents the posterior probability of the target class of the given contextual feature, while P(X|c) is the likelihood which is the probability of contextual feature of a given class (Sarker et al., 2020). The posterior probability is calculated for all classes, and the class with the highest probability will be the instance's label (Karim & Rahman, 2013). The underlying architecture of Naïve Bayes depends on conditional probability. It creates trees known as Bayesian Network based on their probability of happening (Dey, 2016).

The Naïve Bayes model is generally a better classifier than a predictor and is generally more efficient and works wells on categorical data and both binary and multiclass classification (Mansotra, 2019; Shastri et al., 2018), however in this study, we are mostly interested in prediction than classification. Another major pitfall of the Naïve Bayes model is that it assumes independence among child nodes hence resulting in inaccuracy (Kotsiantis et al., 2006).

2.2.4.2 K- Nearest Neighbors

Basing on instances, the KNN algorithm uses a technique which classifies instances basing on likeness to each other within the same dataset. If an instance is labelled as a member of a category, the value of the category label of an uncategorized instance will be determined by the instance of the nearest neighbors(Kotsiantis et al., 2006; Sarker et al., 2020) . For a given integer and an observation the classifier identifies the nearest points in the data close to the observation and then estimates the conditional probability as the fraction of points (James, G., Witten, D., Hastie, T., Tibshirani, 2013).

Let $W = \{x_1, x_2, \cdots, x_n\}$ be a set of n labeled samples. The algorithm is as follows:

```
BEGIN
  Input y, of unknown classification.
  Set K, 1 ⩽ K ⩽ n.
  Initialize i = 1.
  DO UNTIL (K-nearest neighbors found)
    Compute distance from y to x_i.
    IF (i ⩽ K) THEN
      Include x_i in the set of K-nearest neighbors
    ELSE IF (x_i is closer to y than any previous nearest
      neighbor) THEN
      Delete farthest in the set of K-nearest neighbors
      Include x_i in the set of K-nearest neighbors.
    END IF
    Increment i.

  END DO UNTIL
  Determine the majority class represented in the set of K-
  nearest neighbors.
  IF (a tie exists) THEN
    Compute sum of distances of neighbors in each class
    which tied.
    IF (no tie occurs) THEN
      Classify y in the class of minimum sum
    ELSE
      Classify y in the class of last minimum found.
    END IF
  ELSE
    Classify y in the majority class.
  END IF
END
```

Figure 2.2 A sample of KNN pseudo code (Dey, 2016)

In a study such as this, KNN models are employed to classify unknown variables basing on existing variables (Mannion, 2020) for instance existing immunization attendance records alongside other demography records. The main disadvantage of KNN classifiers is the high computational power needed and memory requirement to store the entire sample since all available points are to be scanned for the most similar neighbors to be determined (Al-Dosary et al., 2019).

2.2.4.3 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm used for both classification and regression, the main objective of this algorithm is to find a hyperplane, in a dimensional space and classify the data points (James, G., Witten, D., Hastie, T., Tibshirani, 2013). Support vectors are data points that influence the positioning of hyperplanes and changing them changes the hyperplanes position. They are usually used for margin calculation, which ensures the maximum distance between the margin and the classes for minimal classification errors (Dey, 2016).

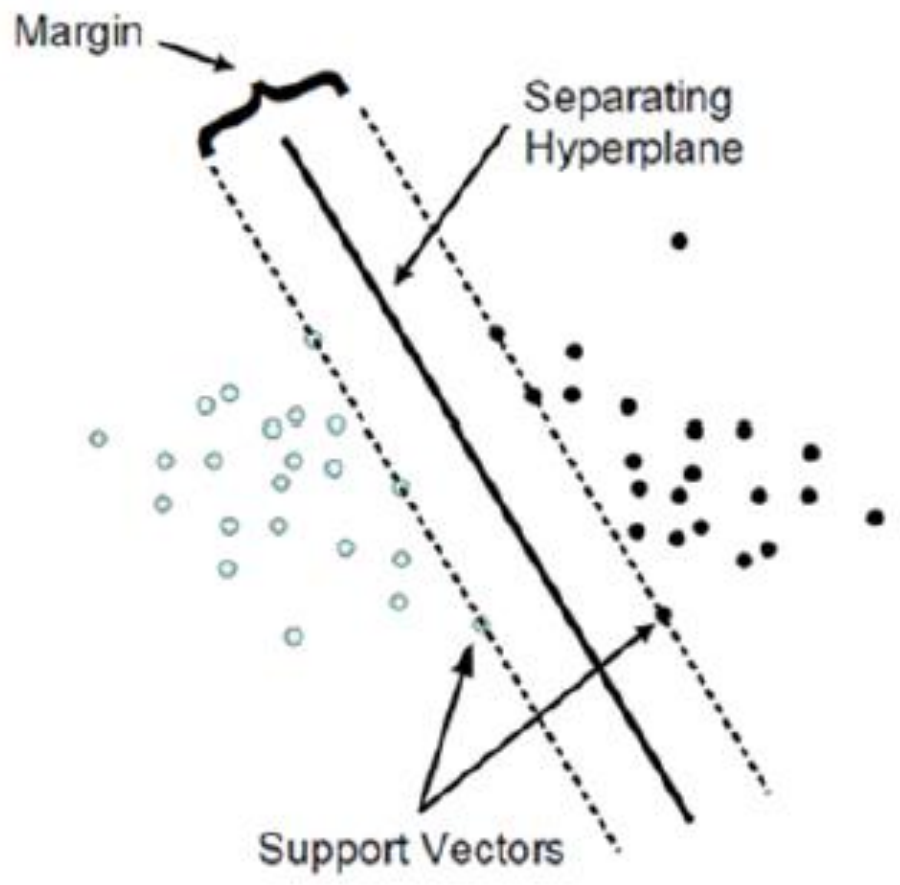


Fig 2.3 Support Vector Machine (Dey, 2016)

SVMs work best with tasks with large features in respect to the number of training instances and with separable data, however, this is not usually the case in the real world and can be solved by mapping the data on a higher dimensional space viz feature space, from where a separating hyperplane can be defined (Kotsiantis et al., 2006). This study seeks to predict the probability of completion in three classes and SVM is not ideal as it separates features into two classes (Rampisela & Rustam, 2018).

2.2.4.4 Decision trees

Decision trees are flow chart like models which use branches to group attributes together, they are mainly used for classification but also can be modelled for prediction (Dey, 2016; Mannion, 2020). Each decision tree has nodes and branches which represent a choice available to be classified in and the chance outcomes from the nodes respectively (Mannion, 2020; Song & Lu, 2015).

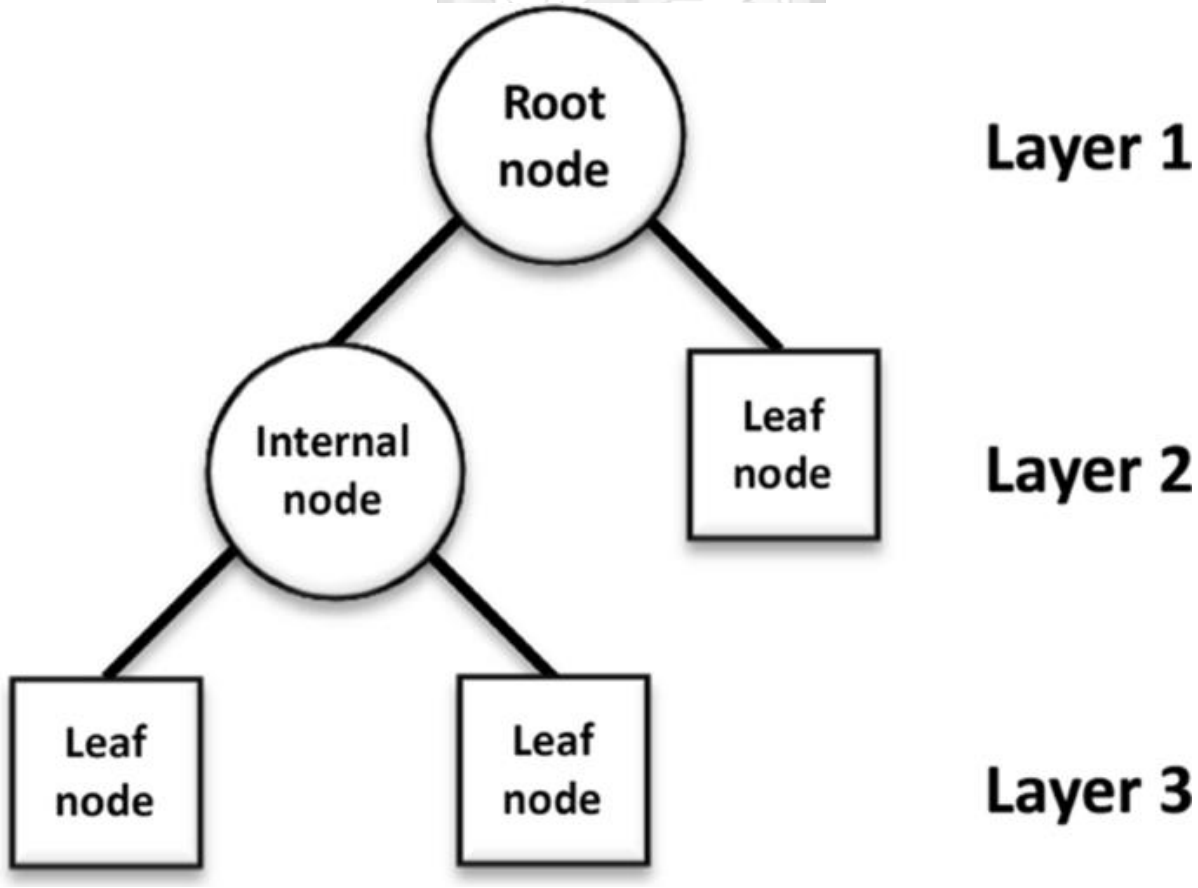


Figure 2.3 A decision tree (Sarker et al., 2020)

When building the model, the hierarchy of most important variables should be identified first then use a top-down divide and conquer approach, then splitting goes on until the predetermined conditions are met (Song & Lu, 2015; Zhang et al., 2020). Decision trees, unfortunately, have lower predictive accuracy and can be non-robust however they can be aggregated to improve their performance like using boosting, bagging and random forests.(Nordhausen, 2014).

2.2.4.5 Random Forests

The random forests (RF) are an ensemble learning classification technique that generates and combines many decision trees and aggregates them to reduce variance that would otherwise happen in single decision trees (Couronné et al., 2018; Mannion, 2020). While growing trees, the random forest selects the best feature for splitting at each node from a random set of features as opposed to selecting the most important feature hence adding randomness and considering other features (Denisko & Hoffman, 2018; Sarker et al., 2020). This reduces overfitting, improves model accuracy and exposes feature importance in prediction (Denisko & Hoffman, 2018).

For a new instance to be formed, each decision tree in the random forest classifies data from the input attributes and chooses the mode prediction for the final output (Altman, 2011). The parameters needed for an RF prediction model to be generated are a predefined number of predictive variables used by several classification trees to group the attributes (Hemmati-Sarapardeh et al., 2020).

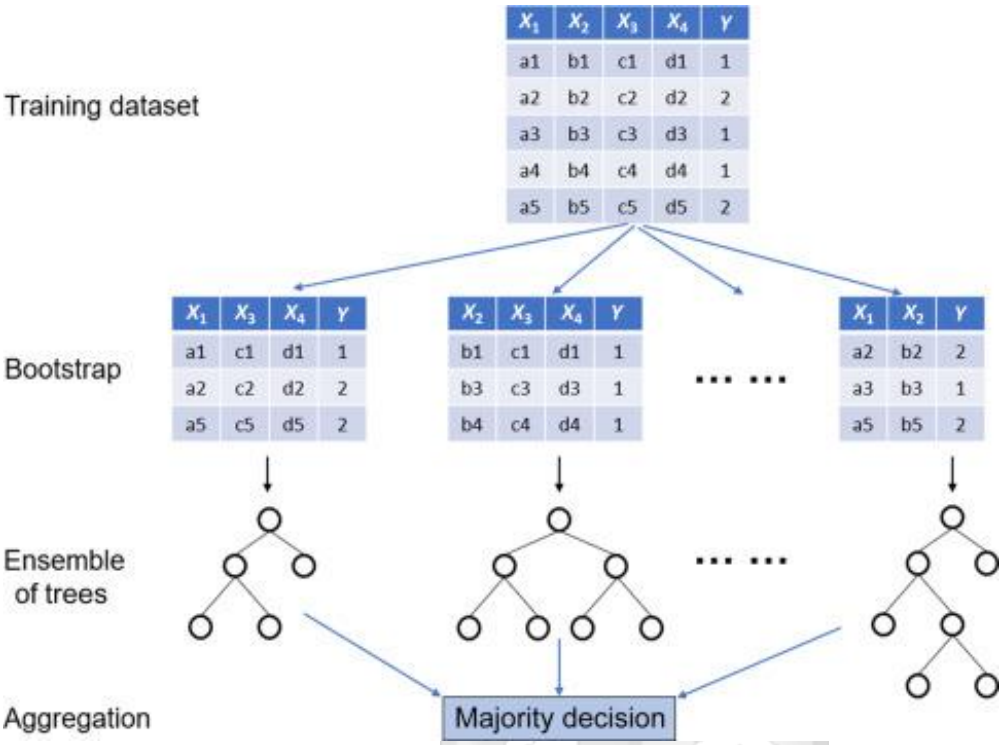


Figure 2.4A random forest classifier (Machine Learning for Subsurface Characterization | Siddharth Misra, Hao Li, Jiabo He | download, 2020)

2.3 Summary of the literature review

From the literature reviewed in the study, existing systems mostly use immunization data to solve issues related to systems facing challenges like facility performance and vaccine logistics (Äijö et al., 2020; Dehnavieh et al., 2019; Kiberu et al., 2014; Maïga et al., 2019; Werner et al., 2019). While some tackle people-facing challenges like the MyChild system sending SMS reminders for immunization and prediction of dropout rates, there are some gaps like knowing which groups need more attention.(Babirye et al., 2012b; Bbaale, 2013a) In a study in Tanzania and Zambia by Carnahan E, Ferriss E et al (2020), the result showed the importance of adding recommenders built into the system for routine monitoring therefore, this study suggests a model feature of predicting immunization completion and will use an existing repository of the MYChild System implemented in Mukono District (Carnahan et al., 2020).

The gaps of existing prediction models include seasonality of models, lack of generalizability and specificity challenges.(Chandir et al., 2018; Chase, 2013; Qazi et al., 2020) In this study's case of Uganda, improving completion rates of immunization requires to put into consideration the people facing challenges like limited access to immunization services, knowledge gap,

environmental factors and health-seeking behaviour (Babirye et al., 2012b; Kamanda, 2010; Menzies et al., 2020; Okot, 2015).

The study proposes classification models because it intends to classify the prediction results in classes; likely to miss, be timely or be late. The study shall use a random forest model to predict the likelihood of immunization completion because of its feature selection ability, accuracy rates and variance reduction.(Couronné et al., 2018; Denisko & Hoffman, 2018; Sarker et al., 2020) In comparison to its counterparts like decision trees, naïve Bayes, KNN, logistic regression and SVMs, this model is more appropriate for this study given its performance and the aim of the research.(Caruana & Niculescu-Mizil, 2006) In conclusion, this study shall use the random forest classifier to develop a model that predicts a child's immunization completion by classification. This has been selected over other classifiers because the data at hand has several attributes which alone are weak but together create a strong learner for classification (Paul & Bhatia, 2020).

Therefore, this study uses a random forest classifier over a regressor given that the data has discrete labels to it and can be defined by classes, in this instance, a participant is classified to have got the vaccine on time, late or missed.

2.4 Conceptual Framework

This section shows a diagrammatic representation of how the research objective is to be achieved based on the reviewed literature. Data will be collected from the existing database of the MyChild system implemented at Mukono health facility, which will thereby be pre-processed for missing value treatment, decoding, standardization, duplicates treatment, cleaning and encoding. After processing, the desired attributes as per literature reviewed in factors that influence immunization completion to use in this case as the predictors like immunization dates, child sex, dates of birth, child exposure to HIV and mother's previous health-seeking behaviour are selected in the final dataset which is further split into training and test datasets. The model is then developed using the training data and is tested until the right training hyperparameters are determined, using the test dataset, it is validated further for accuracy and the final result is output in visualization graphs and reports.

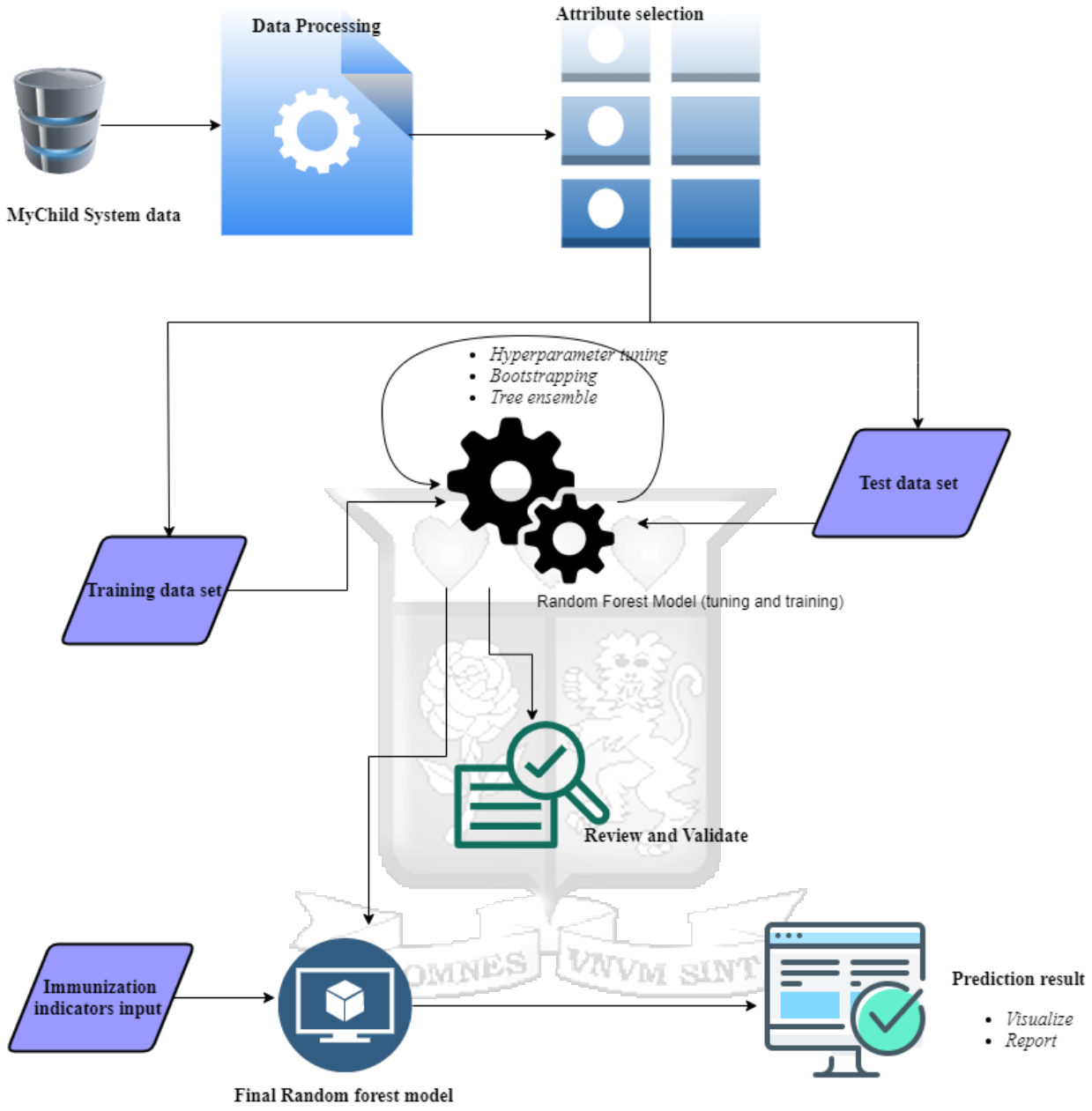


Figure 2.5 Conceptual framework of the proposed classification model to predict immunization completion

3 Chapter 3: Methodology

3.1 Introduction

This chapter is where the methodology and techniques to be used by the study to collect and analyse data and develop a novel classification model will be elaborated. This section highlights the system methodology, research design, data acquisition and processing, models development and its testing and validation. It exhibits the quality of the research and highlights a summary of possible ethical considerations

3.2 System methodology

Dynamic System Development Method (DSDM), an agile system development method will be used in this study given the advantage of frequent, iterative and incremental development changes (M, 2020). Dynamic System Development Method (DSDM) is based on the Rapid Application development approach basing on the study needs, model quality, timely delivery and incremental iterative development (Alsaqqa et al., 2020).

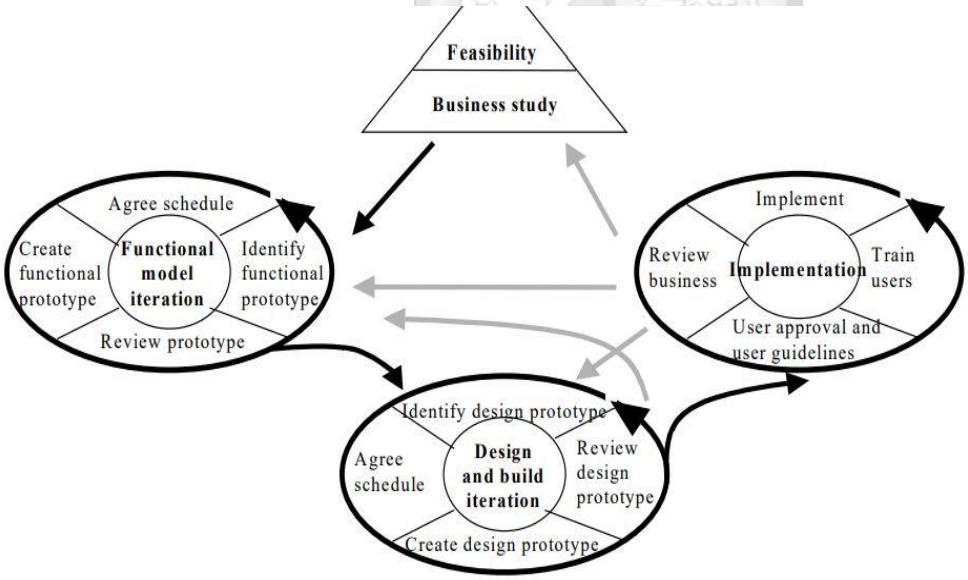


Figure 3.1 DSDM process diagram (Abrahamsson et al., no date).

3.2.1 *Feasibility study phase*

In this phase, the researcher analysed the relevant factors for system development and analyse the risk to determine technical requirements for the study. Informed by the literature review, the researcher identified Mukono health facility, with the data they had collected on immunization to

have ample attributes to use in the study. The factors used include the vaccine schedule attendance (attendance patterns), exposure to counselling (information received), exposure to tetanus vaccine (health seeking behaviour), child exposure to HIV and the HIV testing.

3.2.2 *Business study phase*

The researcher in this phase made a list of functionalities needed in the model, a high-level description of the development process, model architecture and an outline of the prototyping plan. Using the conceptual framework in Fig. 2.5, the researcher derived the functional requirements like the inputs and model outputs.

3.2.3 *Functional model iteration*

In this stage of the study, the researcher scheduled project delivery dates in phases, still considering the conceptual framework, the researcher developed a basic random forest model. Started with using defaults and review the results (OOB error) and identify then fine tuned the hyperparameters like the mtry and number of trees to find the best fitting model. This ws an iterative and incremental phase which is appropriate given the nature of this study.

3.2.4 *Design and build phase*

This stage covers designing and building the final model to be used. Testing and reviews were done on the results of the test data from each iteration to determine a model of best fit. The best fitting model was determined by the scores from the Out of Bag error and the cross-validation validation tests done at the output.

3.2.5 *Implementation phase*

The model developed from the previous phase was implemented in subsequent tests using different datasets. Users will then be trained on how it works and their feedback from interacting with the model shall be used to further develop a better model, and the process iterates. However, it should be noted that this phase is out of scope for this particular study.

3.3 Research design

Research designs are the methods used in data collection, measurements and analysis by the researcher to conduct the study. In this study, experimental research design was employed, both qualitative and quantitative methods of research were incorporated in the study. While the data to be used by the researcher is quantitative data, the researcher employs qualitative analysis when tuning parameters and determining the influence a descriptive attribute like a caregiver's

counselling status might have on the final prediction of a child's participation in immunization attendance. The modelling will be iterative to focus on enabling the model to randomly consider all possible features that influence the category the child is likely to be predicted into. Running the research this way is to achieve objectivity, eliminating possible biases and ensuring that the model is accurate and robust.

3.4 Target population

The study population are infants who have been registered between 2015 and 2020 July in the My Child database deployed at Mukono health facility, Mukono district, located in central Uganda. In this research, the interest of the researcher was historical data as extracted from an existing immunization records system, My Child solution, that pre-collected and registered individual entries of child immunization data as scheduled from birth to 14 weeks. This population was selected because it is only relevant to the study and provides the necessary data required to achieve the aim of the study. The data acquired from this populations span from 2015 to 2020 July, this duration was selected based on the availability of consistent data and the significance of its volume to be able to construct a reliable model that can accurately predict the result and that it can be tested and validated. The choice of attributes selected was informed by the factors that influence immunization completion as per the studied literature including the caregiver's health seeking behaviour indicated by counselling, tetanus protection, previous vaccine attendance, vaccine schedules and HIV exposure.

3.5 Data collection

The source of the data is a pre-existing repository of the My child Solution pilot as implemented at Mukono district health facility, the data collected is of records between 2015 to 2020. The researcher requested the data through the facility administrator and the data was availed at no cost. The data has been deidentified because it is individual-level data and for the sake of protecting the confidentiality of both caregivers and the infants. From the data collected, below is a description of the selected columns

Table 3.1 A table showing the attributes of the data collected for the study

Column name	Description
public_id	The child ID in the system. This ID is reported on the Child health forms when a child receives vaccine

sex	Child sex
birth_date	Child's birth date
pab	Was the child protected at birth: protected from neonatal tetanus before delivery? (Moth immunized)
counseling	Has counselling been given?
hiv_tested	Has the child been tested against HIV? N= No, Y=Yes and NA=The mother is HIV negative
ipv	date of administration polio vaccine dose (injection administration)
dpt_3	date of administration of third dpt dose
opv_3	date of administration of third opv dose
pcv_3	date of administration third pcv dose
bcg	date of administration of bcg dose
rv_1	date of administration rotavirus first dose
rv_2	date of administration rotavirus second dose
dpt_1	date of administration dpt first dose
dpt_2	date of administration dpt second dose
opv_0	date of administration polio vaccine first dose (orally administered)
opv_1	date of administration polio vaccine second dose (orally administered)
opv_2	date of administration polio vaccine third dose (orally administered)
pcv_1	date of administration first pcv dose
pcv_2	date of administration second pcv dose

3.6 Data Analysis and exploration

The researcher studied the data and identified the variables needed for the study model development as in Table 3.1; the data was then cleaned for missing values, outliers, duplicates and then standardized, normalized and quality checked in STATA version 14. The researcher used the available columns to create categorised data from vaccine administration dates and date of birth to determine whether a child was on time, late or missed the dose guided by the Uganda vaccine schedules by UNEPI. The data was exported to a CSV file for further analysis and model development in R statistical software package.

3.7 Model development Methodology

The model developed utilizes a random forest algorithm for classification to predict the category which a child is likely to fall into for the third DPT3 dose basing on the available predictors, this model was selected for its random feature selection which puts into account all features in the dataset for accuracy.

The parameters that were tuned for this model are the number of trees to be used in the model and the number of random variables to be used in the tree to get a stable classifier.

3.8 Testing and validation

The developed random forest model was tested using the out of bag test which is built within the running on the training data, and on running the test data, we used the confusion matrix to check the accuracy, sensitivity and specificity of the model. This method was chosen because it shows performance metrics details including accuracy of the model, the exactness of the model, the true positives and negatives rates among others which a sufficient for validation of a model.

3.9 Research Quality

Research quality consists of two criteria, validity and reliability.

Validity checks whether the study measures what it says it does and whether the results can be generalized to other settings. In this study, this was checked by running the model on a sample of test data and the results showed an accuracy rate of 76% and the study variables used can be generalized in another setting.

Reliability on the other hand is concerned with the repeatability of results, this was ensured in hyperparameter tuning and setting the seed such that the particular sequence of random numbers is reproducible

3.10 Ethical considerations

This research study is to be reviewed by the Strathmore University Ethics Review Committee and then approved to proceed. This study had the likelihood of exposing participants' personal information that could be considered confidential. To minimize this risk, we protected children and parents' confidential data by eliminating personal identifiers like names and addresses. The obtained data was requested from the facility administrator and was handled with high confidential regard as agreed. Data from secondary sources were cited for future reference.

4 Chapter 4: System Analysis, Design and Architecture

4.1 Introduction

This chapter covers the system architecture and detailed design and analysis of the proposed model while expounding the different requirements. From the conceptual framework, the model to predict immunization completion, the model will use data extracted from the My Child System database that will be pre-processed and cleaned for data quality and efficiency in further processing.

Furthermore, the chapter analyses gathered requirements and how they were collected as well as what the model's functional and non-functional requirements are. From this chapter, the system architecture is explained, and the system designs diagrammatically presented to explain how the proposed model objective shall be accomplished.

4.2 System Analysis

System analysis is the process of interpreting facts and decomposing the requirements gathered for system development and what is expected of the system during development and after it has been developed. This chapter analyses the functional and non-functional requirements, how they were gathered and analysed and how they are used to develop a model that fulfils the study objective.

4.2.1 Requirement Gathering

From Mukono health facility where the MyChild system is piloted, this study sought the requirements of the system. Bearing in mind that the experimental study is incremental and should be able to be integrated with the existing system. Therefore, the developed model was required to be interpretable and understandable buy the stakeholders for it to fulfil its intended purpose.

The requirements that were gathered including the used data set was pre-processed. This included extraction of the required dataset from the database by the stakeholder. The extracted data was then cleaned, this process included checking for missing values and outliers those were deleted, the remaining data was deidentified for confidentiality and security purposes. The researcher then derived variables from available data for instance, computing vaccine due dates from the date of birth, data was then transformed into classes by factorization.

4.2.2 Functional Analysis/Requirements

The following are the functional requirements that are expected to be incorporated in the proposed model to meet the objective of the study:

i. The model should extract the attributes as loaded in the files and randomly select a subset of predictor variables from the dataset and build many decision trees. The model should show the predictions of the decision trees in the random forest
ii. The system should display a report of the final prediction showing the results of the prediction, the accuracy and the variable importance.

4.2.3 Non-functional Analysis/Requirements

4.2.3.1 Scalability

The proposed model should be able to handle a large file without its functionality being compromised.

4.2.3.2 Usability

The proposed model is intended for use by a trained team at health facility level with the ability to understand the data. Therefore, the system should be aligned with the objectives of the health facility stake holders ensuring accuracy since the results will be used to make decisions.

4.2.3.3 Reliability

The proposed model should have accurate output after training and parameter tuning so as to be able to self-retrain and give accurate output. The output by the model must be consistent when the same instance is repeated. The systems should always be able to extract imported data and in an event of failure, it should be able to be restored to functional state by the administrator.

4.2.3.4 Security

The proposed model involved deidentification of data. The participants have to be protected from any form of data leaks. Being health data, it is confidential and even when being shared for decision making, the privacy of the participants must be protected as all times.

4.3 System Architecture

This section shows the layout and components of the implementation of the model guided by the conceptual framework in Fig 2.6. as showed in Fig 4.1. The input of the system are pre-

processed files of data from the My child system and loaded into a graphical interface where the data will be submitted and then processed by the model. An output to be downloaded by the user is then given and it shows results that is the prediction of which category of immunization completion status a child will be, the accuracy of the model on that data sample and the important variables used in that data. These categories will be given as output which can be exported by the user.

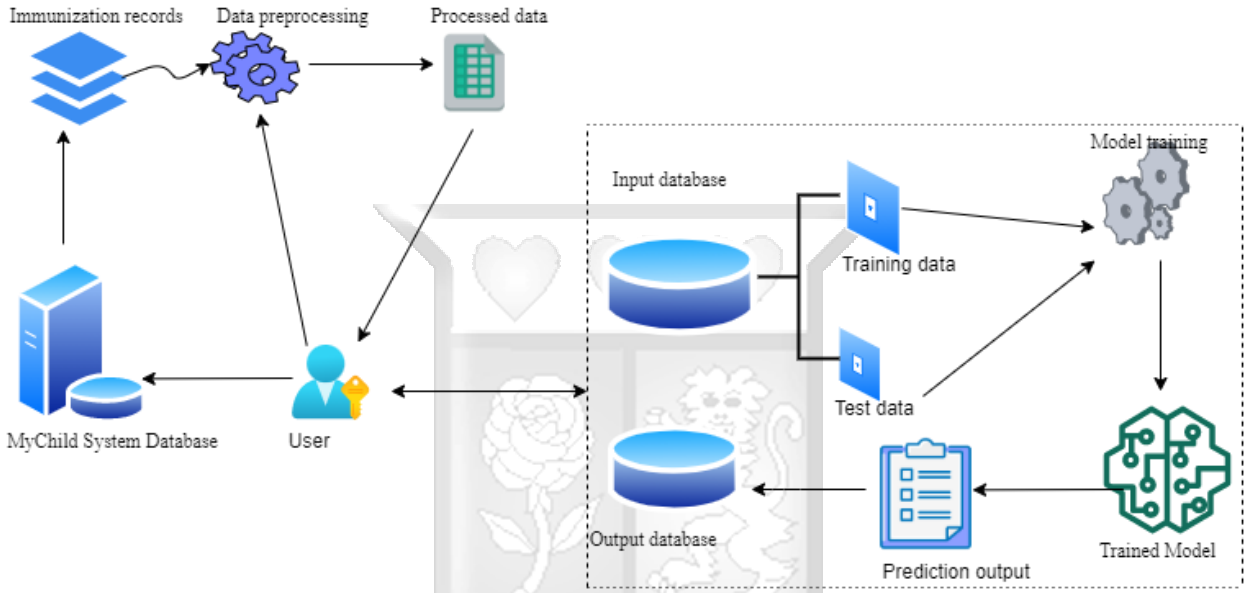


Figure 4.1 System Architecture

4.4 System Design

This section represents concepts of the solution, the interaction, modules, data representations, allocations of functions to users, software and hardware. This representation intends to highlight the details of the decisions about functionality organization.

4.4.1 Use Case diagrams

This depicts a user's interaction with the system, Tables 4.1, and 4.2show use case scenarios and Figure 4.2 illustrates the interaction of a user (actor) and the system.

Table 4.1 Use Case Scenario 1

Use Case	**Model development, training and testing**
Primary Actors	**Researcher/Developer**

Pre-Condition	Dataset standardized, uploaded and split into training and testing data Packages imported
Post-condition	Well-functioning model
Major Steps Performed	
Actor	**System**
The developer selects training data	
The developer tunes the hyperparameters	
	The algorithm performs bootstrap aggregation
	The algorithm randomly selects features to form datasets and random decision trees
	The algorithm reiterates, takes majority vote and compiles OOB error
Developer loads test data	
	System runs model as pretrained
	System generates prediction and saves data in the output database
Developer retrieves, views, analyses the output	
Developer validates model	
	System saves the new state of the model

Table 4.2 Use Case Scenario 2

Use Case	**Import immunization data**
Primary Actors	**User**
Pre-Condition	Data has been pre-processed

	User was pre-registered and logged into system
Post-condition	System stores data into the database
Main success scenario	
Actor	**System Responsibility**
User imports dataset	
	System saves dataset
User selects dataset for prediction	
	System loads the dataset for prediction
	System returns prediction results with categories, recommendation, and accuracy and saves in the output database
User retrieves/views/exports results	

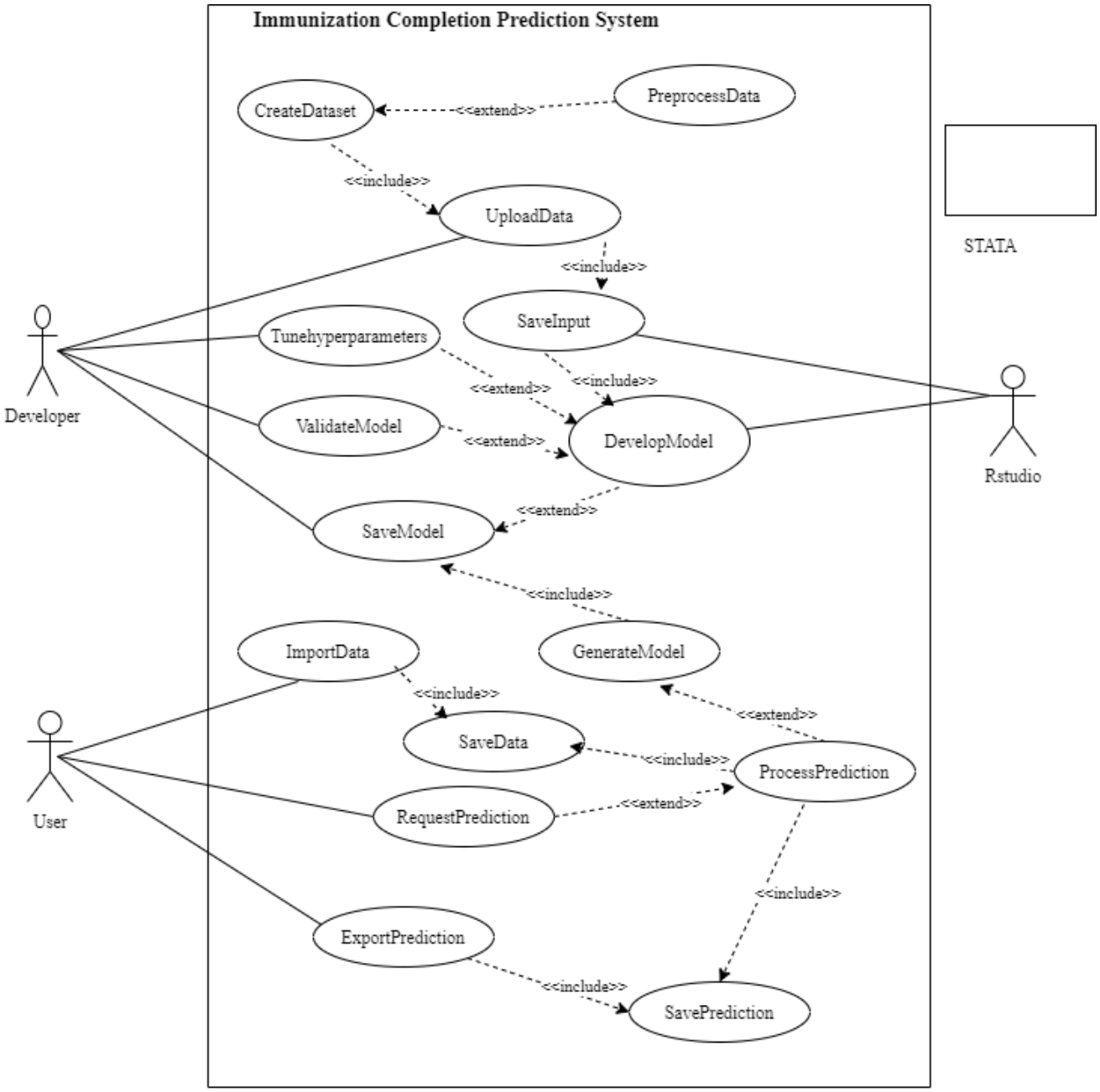


Figure 4.2 Use case diagram

4.5 System sequence diagram

A system sequence diagram in Figure 4.3, elaborates the behaviour of the system to complement the use case in Figure 4.2. The diagram shows the user importing data into the system a SaveData instance is created and the imported data is saved in a database. The user proceeds to make a request to create a prediction instance, the instance generates an existing model, in this model, the dataset of choice is loaded and processed. When the model generates a prediction, it

then saves the output in a database from which the user can export and view the data. The output contains the prediction, recommendation and accuracy.

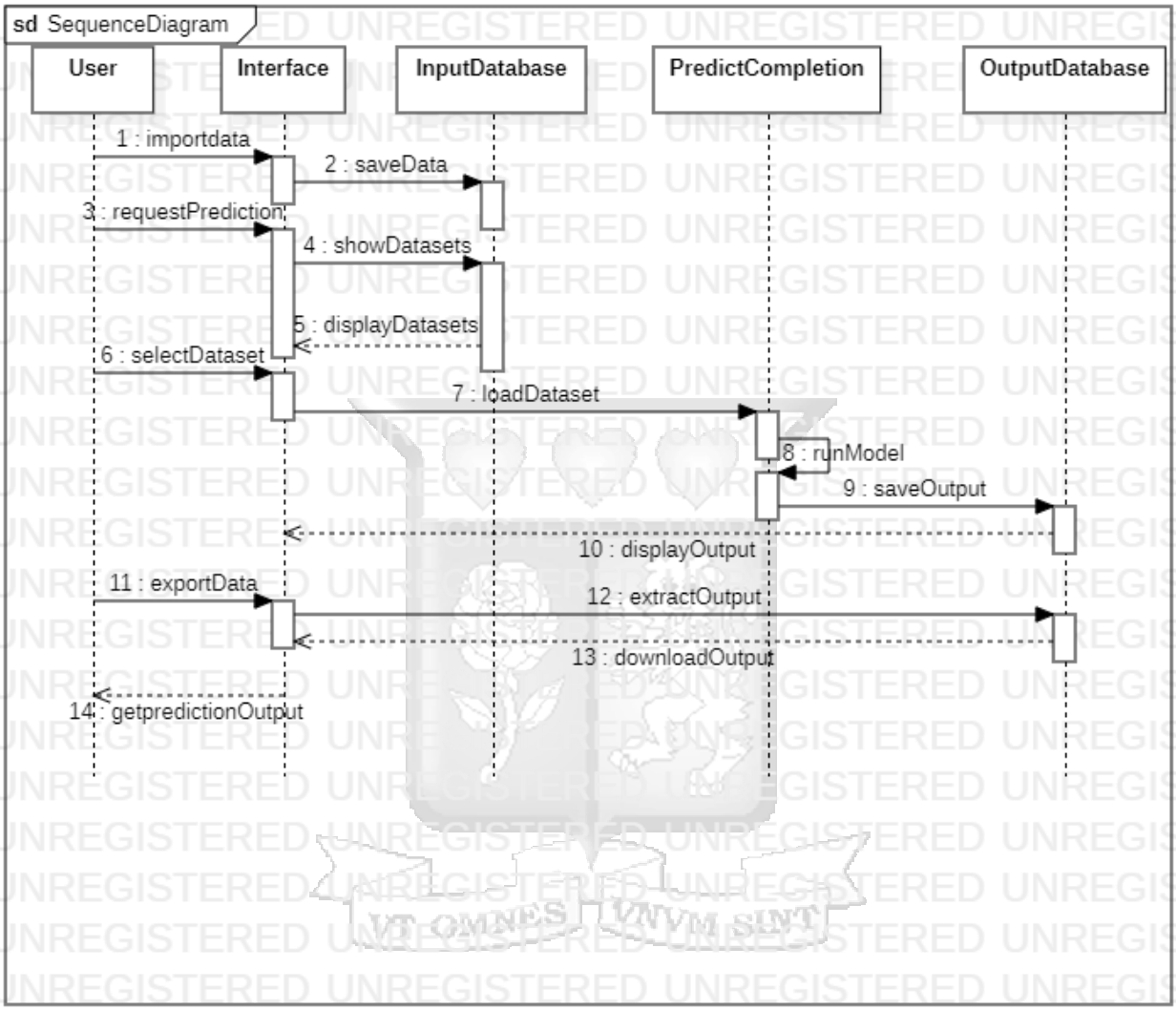


Figure 4.3 Sequence Diagram

5 Chapter 5: Implementation and validation

5.1 Introduction

This chapter focuses on how the model was developed, trained and tested. Implementation covers the tools and processes used to develop a random forest-based classification model to predict whether a child will be late, on time or miss the dpt3 vaccine dose. This section details how data was cleaned and pre-processed, the libraries used in model development, parameter tuning and how these affected the model development. The system validation section covers details on how the developed trained model performs on testing data and its accuracy.

5.2 System implementation

5.2.1 Data Pre-processing, normalization, and loading

The data was pre-processed in STATA version 14, and the first step was checking for and removal of duplicate observations. The data also had incomplete entries we rendered as irrelevant and those were eliminated leaving the dataset with data of children from 2015 January to 2020 July, and from this data, the transformation of variables with vaccine reception dates to categories of whether the vaccine was administered on time, late or missed coded 1,2,3 respectively. From this, outliers were detected and removed too. Other variables were derived from the date of birth and that is due months of the different schedules and the birth month, these would be used as predictors in the model. For the sake of privacy and confidentiality, automatic public IDs were generated to replace the actual public IDs. Finally, we generated and exported a CSV file from STATA with 22 variables and 71,157 entries and loaded it into R.

```
> dim(registry)
[1] 71157    22
> str(registry)
spec_tbl_df[,22] [71,157 x 22] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ public_id   : chr [1:71157] "1015-5185" "1002-2821" "1002-5014" "1066-8529" ...
 $ birthmonth : num [1:71157] 12 8 1 5 3 10 3 4 6 11 ...
 $ due1        : num [1:71157] 12 8 1 5 3 10 3 4 6 11 ...
 $ due2        : num [1:71157] 1 10 3 7 5 11 5 6 8 1 ...
 $ due3        : num [1:71157] 2 10 4 8 6 12 6 7 9 2 ...
 $ due4        : num [1:71157] 3 11 4 9 6 1 7 8 10 3 ...
 $ bcg1        : num [1:71157] 3 3 3 3 3 3 3 2 3 3 ...
 $ polio       : num [1:71157] 3 3 3 3 3 3 3 3 3 3 ...
 $ polio1      : num [1:71157] 3 3 3 3 3 3 3 3 3 3 ...
 $ dpt1        : num [1:71157] 3 3 3 3 3 3 3 3 3 3 ...
 $ pcv1        : num [1:71157] 3 3 3 3 3 3 3 3 3 3 ...
 $ rotav1      : num [1:71157] 3 3 3 3 3 3 3 3 3 3 ...
 $ polio2      : num [1:71157] 3 3 3 3 3 3 3 3 2 3 ...
 $ dpt2        : num [1:71157] 3 3 3 3 3 3 3 3 2 3 ...
 $ pcv2        : num [1:71157] 3 3 3 3 3 3 3 3 3 3 ...
 $ rotav2      : num [1:71157] 3 3 3 3 3 3 3 3 2 3 ...
 $ dpt3        : num [1:71157] 3 3 3 2 3 3 3 3 2 2 ...
 $ childsex    : chr [1:71157] "M" "M" "F" "M" ...
 $ pab         : chr [1:71157] "OTHER" "OTHER" "OTHER" "OTHER" ...
 $ conselled   : logi [1:71157] FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ hiv_exposed: chr [1:71157] "OTHER" "OTHER" "OTHER" "OTHER" ...
 $ hiv_tested : chr [1:71157] "No" "No" "No" "No" ...
```

Figure 5.1A screenshot of code and output showing data structure and dimension

After loading data, null variables were observed so further cleaning was done to drop those null entries leaving the dataset with 53,442 entries. After, the predictor and target variables were converted into factors by applying the *lapply()* function, it is important to note that the random forest works best when the response variable is a factor. Converting these variables into factors shows that they are not logical, nor are they integers but categories.

A new dataset that combines the IDs with the rest of the variables was created. From this data, dataset for the model was ready for use

```
> clean = registry %>% drop_na()
> dim(clean)
[1] 53442    22
```

Figure 5.2 A screenshot of the code dropping null entries and the dimension of the new dataset

```
convert <- lapply(clean[2:22], factor)
View(convert)

#bind with ID
immunization <- cbind(clean[1], convert)
dim(immunization)
str(immunization)
```

*Figure 5.3*A screenshot showing the conversion to factors and binding the public ID to create a new data frame to use in model development

5.2.2 Development environment

The development of the model was done in RStudio, R version 4.0.5 on a 64bit Windows 10 machine with Core i5 and 4 GB RAM. This environment was preferred because it is open source given the budget and its statistical and graphical techniques but also because it allows the user to add functionality as well as extend packages and libraries

The libraries used in developing models are shown in the table below.

Package	Library	Function
tidyverse	readr	Read rectangular data
dplyr	dplyr	Used for data manipulation
caret	caret	Classification, model tuning and testing

randomForest	randomForest	Create and analyse random forests
shiny	shiny	To build web interactive web applications

Table 5.1 Packages and libraries used in developing the model

5.2.3 Random Forest Model Components

This section entails the details of the model, after data was loaded, the model training and model structure and the implementation of the model.

5.2.3.1 Data partitioning

A seed of 123 was set to enable the sampling to be able to reproduce a particular sequence of random numbers. Using the function *initial_split(),* created a split of the data into two samples of 75% and 25% of the original dataset to be used for training and testing respectively.

Insert code and output

5.2.3.2 Random forest model training

Using training data, a random forest algorithm set to default parameters, on checking variable importance, date_of_birth was the least important therefore we choose to drop it from the model since it did not affect the accuracy of the model.

5.2.3.2.1 Hyperparameter tuning

The first hyperparameter we tuned was the mtry this is the number of variables randomly sampled as candidates at each split, this was done using the *tuneRF()* function to choose the best fit mtry with the least errors. We chose 4, with the OOB error of percentage 0%.

Second tuned hyperparameter was the Ntree, the number of decision trees to grow for the whole model. According to Random Forest package description, Ntrees are encouraged to be set to a relatively big number such that each predictor variable gets predicted a few times (Denisko & Hoffman, 2018; *How to Implement Random Forests in R | R-Bloggers*, n.d.; *Random Forest - an Overview | ScienceDirect Topics*, n.d.-a).

```
#get optimal mtry
mtry <- tuneRF(train[-1], train$dpt3, ntreeTry=1000,
               stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE)
best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]
print(mtry)
print(best.m)


#random forest

classifier = randomForest(dpt3~., data=train[-1],
                          importance = TRUE, ntree = 1000,mtry = 4,
                          replace=TRUE, random_state = 0)
print(classifier)
```

Figure 5.4 A screenshot of the code to get optimal mtry parameter and random forest model training

```
> mtry <- tuneRF(train[-1], train$dpt3, ntreeTry=1000,
+                stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE)
mtry = 4  OOB error = 0%
Searching left ...
mtry = 3        OOB error = 0.03%
-Inf 0.01
Searching right ...
mtry = 6        OOB error = 0%
```

Figure 5.5 A screenshot of the output mtry

5.2.3.2.2 Final model training

After hyperparameter tuning, the model was trained using the randomForest() algorithm function as in the Figure 4.4 above, from this we had OOB error of and important variables as shown in the figures below.

```
> print(classifier)

Call:
 randomForest(formula = dpt3 ~ ., data = train[-1], importance = TRUE,      ntree = 1000, mtry = 4, replace = TRUE, rando
m_state = 0)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 23.88%
Confusion matrix:
    1     2     3 class.error
1 254   198   677   0.7750221
2 143 10998  4083   0.2775880
3 124  4348 19257   0.1884614
```

Figure 5.6 A screenshot of the output of the trained model classifier

```
> print(classifier)

Call:
 randomForest(formula = dpt3 ~ ., data = train[-1], importance = TRUE,      ntree = 1000, mtry = 4, replace = TRUE, rando
m_state = 0)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 23.88%
Confusion matrix:
    1     2     3 class.error
1 254   198   677   0.7750221
2 143 10998  4083   0.2775880
3 124  4348 19257   0.1884614
```

Figure 5.7The output of the trained model

```
#Variable Importance
randomForest::importance(classifier)
varImpPlot(classifier)
```

Figure 5.8 Screenshot of the code for variable importance

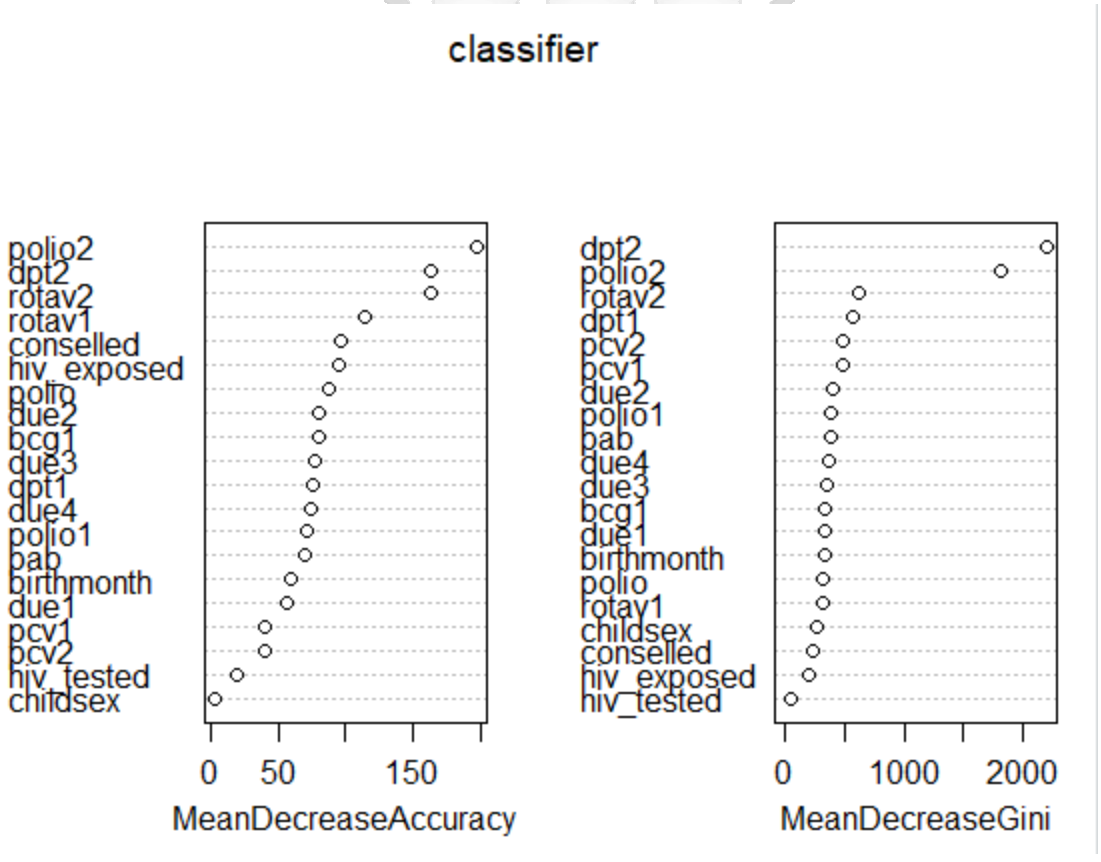


Figure 5.9 A screenshot of the variable importance plot

```
> randomForest::importance(classifier)
                    1         2          3 MeanDecreaseAccuracy MeanDecreaseGini
birthmonth  10.593839 39.057561  25.096029              59.42138         324.11312
due1        11.382659 37.264431  23.972086              55.58625         324.79881
due2         8.196883 50.341436  35.177501              80.63085         393.83940
due3         7.294038 49.201477  32.473976              76.41102         350.45016
due4         9.924880 48.442368  28.817226              74.41332         360.70164
bcg1        11.247432 39.646437  50.349215              79.36574         331.98995
polio       17.402864 54.675847  44.795540              86.65947         319.83399
polio1      16.742750 50.115090  50.969238              70.88603         389.72354
dpt1        22.642315 44.551460  51.417496              75.05638         560.92704
pcv1        23.036723 28.350239  31.119696              40.45081         485.65809
rotav1      30.701192 15.986326  98.012766             113.84817         307.53142
polio2      32.292337 52.454488 172.539979             196.64511        1812.94397
dpt2        55.426811 25.201105 170.801622             163.35041        2207.78879
pcv2        23.048017 28.668115  31.382283              40.35242         490.58454
rotav2      22.103296 35.416161 140.852638             162.53375         621.73903
childsex     9.828484 -5.842924   7.809551               2.63266         269.35124
pab         12.360567 30.945138  52.747194              70.18167         377.92015
conselled   19.475267 76.075901  42.866046              96.40910         230.48680
hiv_exposed 21.233836 54.434542  71.801815              94.50637         188.87638
hiv_tested  -3.407469  2.210088  21.981155              18.88408          43.33847
```

Figure 5.10 A screenshot of variable importance details

5.2.3.3 Random forest model structure

This section explains what happens at each point as per the figure below

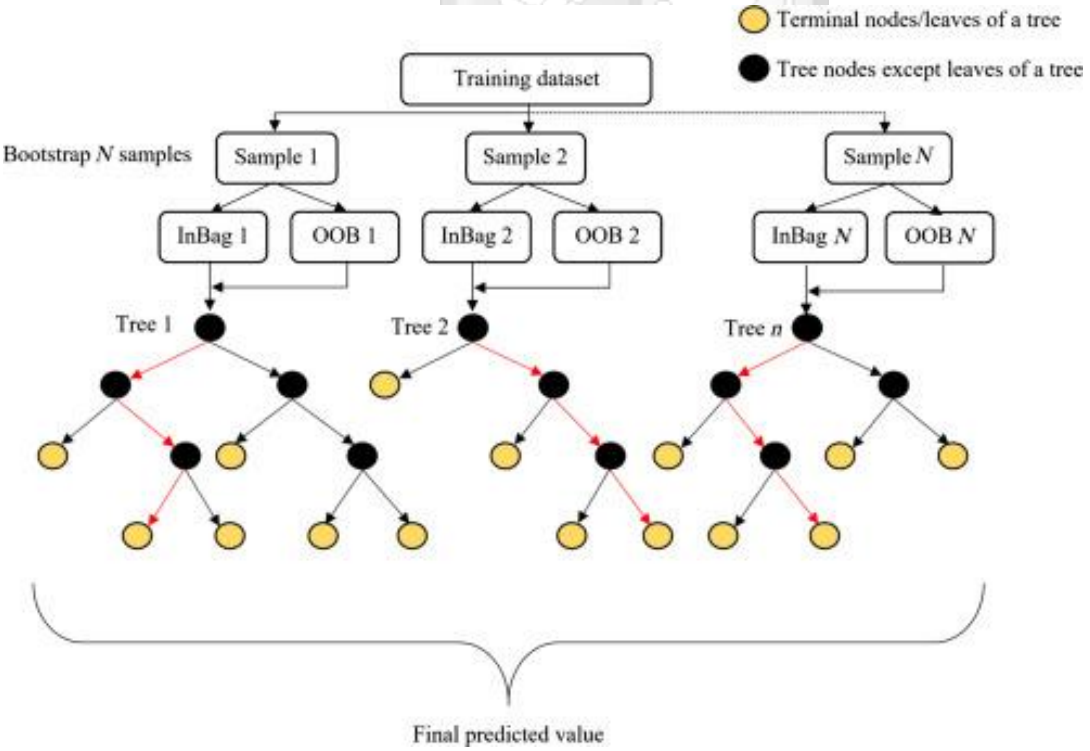


*Figure 5.11 A structure of a random forest (*Random Forest - an Overview | ScienceDirect Topics, *n.d.-b)*

5.2.3.3.1 Training dataset

This is the input of the model, it is usually pre-processed and made ready to be used for the model development. In the RF formula, from this dataset we make clear which variables are predictors and which one is the target variable to be predicted.

5.2.3.3.2 Bootstrap sampling

This a resampling method used in the random forest algorithm which works by sampling variables randomly from the dataset with replacement and the bootstrap is used to compute the statistic as assigned. It should be noted that the algorithm used in this study, while bootstrapping, for each sample, it leaves out some data that will not be used in the decision tree known as the Out of Bag. On getting a prediction outcome, it is used to calculate the OOB score which is the number of correctly predicted rows from the OOB sample.

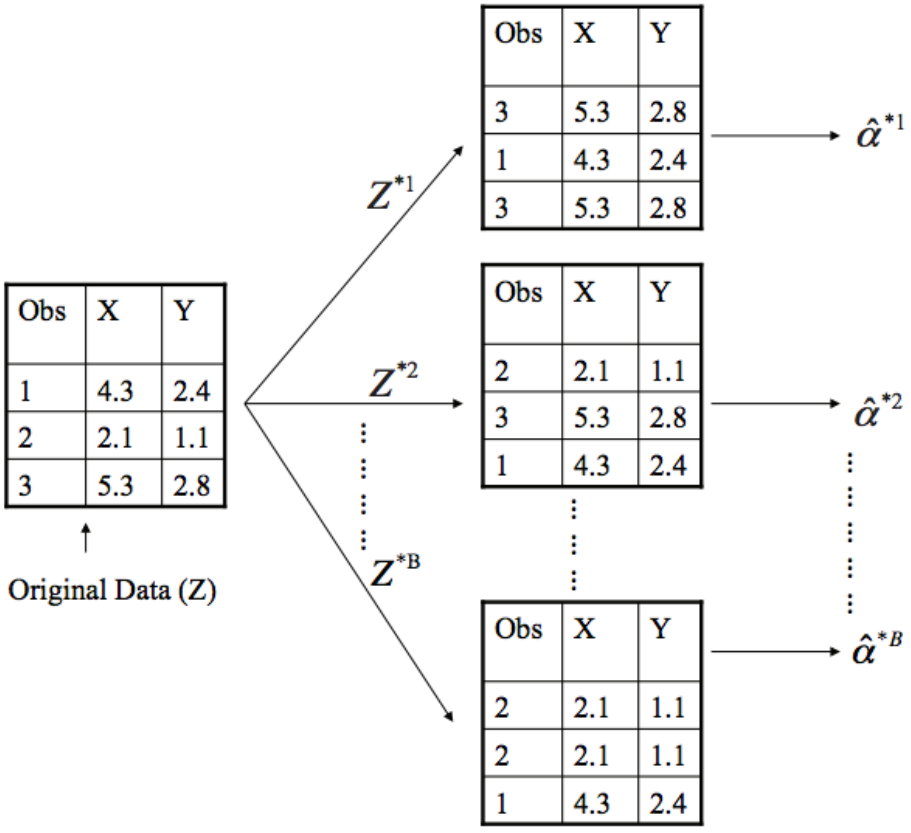


Figure 5.12 An illustration of bootstrapping process (Resampling Methods · UC Business Analytics R Programming Guide, n.d.)

5.2.3.3.3 Decision tree ensembling

In the random forest algorithm, a group of decision trees as defined by the researcher are created using the bootstrapped data and each tree gives its prediction as output as illustrated in figure 5.13.

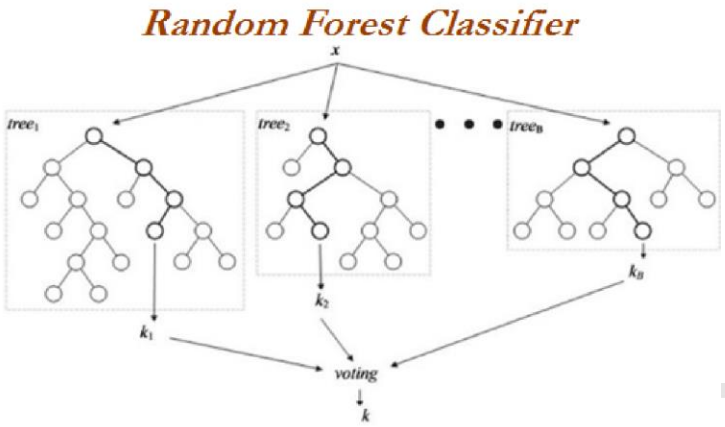


Figure 5.13 A structure of a random Forest(Random Forest Algorithm- An Overview | Understanding Random Forest, n.d.)

It should be noted that each tree grows like an inverted tree with the root at the top, the root node is where all variables to be used in that tree are. Then attributes following a particular condition are split on branches, following the conditions on the internal nodes. A child node is a sub node of a root node and the leaf nodes represent the best predictor.

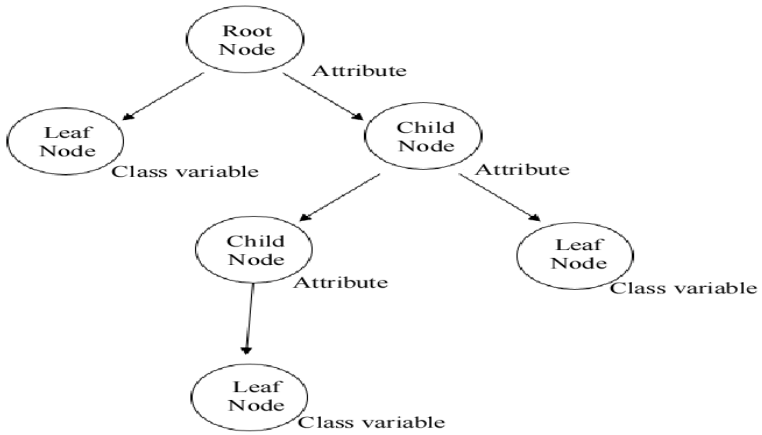


Figure 5.14 Decision tree structure (Decision Tree Structure (Martínez et Al., 2009) | Download Scientific Diagram, n.d.)

5.2.3.3.4 Aggregation

In this random forest, a prediction inference is made by aggregating the predictions made by each tree in the model. In this case of classification of categorical variables, the category the appears most is selected as the final prediction.

5.3 System Implementation

After the random forest was tuned to an acceptable OOB error percentage of 23.88, variable importance was also checked to make sure all variables used were relevant. The output was visualized as shown in Figure 5.7 above.

This model developed was deemed fit by the researcher given its accuracy and performance rate was then used on the testing data from the output, this model would be tested and validated.

5.4 Model Testing and results

The model was tested on the testing data that was 25% of the entire dataset with the code and outcome as shown below.

```
#check with test
pred = predict(classifier, newdata = test, type = "class")
pred

plot(pred)
```

Figure 5.15 A screenshot showing the code running the model on testing data

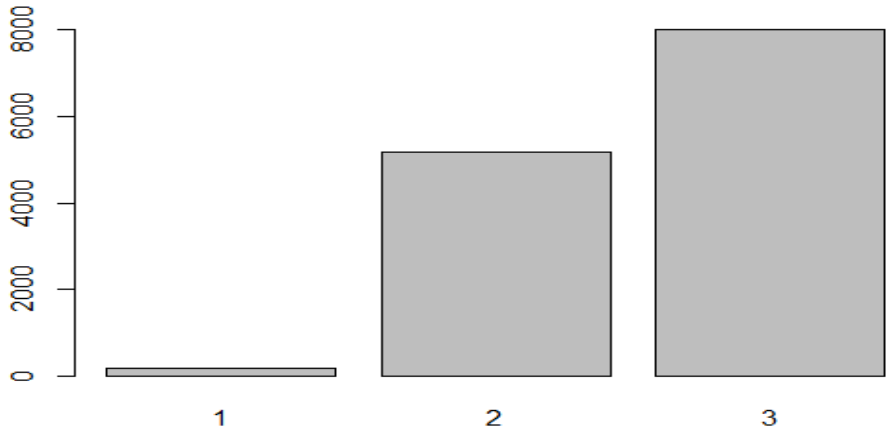


Figure 5.16 A screenshot of the output of prediction on test data.

Confusion matrix was then used to calculate the accuracy and performance of the data

```
> confusionMatrix(pred, test$dpt3)
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   87   47   36
         2   68 3652 1464
         3  229 1335 6442

Overall Statistics

               Accuracy : 0.7621
                 95% CI : (0.7547, 0.7693)
    No Information Rate : 0.5945
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5214

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity          0.226562   0.7255   0.8111
Specificity          0.993604   0.8160   0.7113
Pos Pred Value       0.511765   0.7045   0.8046
Neg Pred Value       0.977483   0.8310   0.7198
Prevalence           0.028743   0.3768   0.5945
Detection Rate       0.006512   0.2734   0.4822
Detection Prevalence 0.012725   0.3880   0.5993
Balanced Accuracy    0.610083   0.7707   0.7612
```

Figure 5.17 A screenshot of the code and output of the confusion matrix.

From this output, the prediction results are as follows

From the training set, for each category classified as 1, 87 entries were predicted correctly while 47 were predicted into class 2 and 36 into class 3. This class was least accurately predicted class at a balanced accuracy percentage of 51%, sensitivity rate of 33% and specificity rate of 99%.

Category 2, classified 3652 entries correctly, classifying 68 in class 1 and 1464 in class 3, the balanced accuracy percentage of category two was 77%, sensitivity rate of 72.5 and specificity of 81%.

Category three, the model predicted 6442 entries correctly in class 3 while misclassifying 229 entries incorrectly in class 1 and 1335 into class 2. The overall balanced accuracy for category 3 was 76%, with a sensitivity rate of 81% and specificity of 71%.

Overall, the accuracy of the model was 76% within a 95% confidence interval of 75.47 and 76.93, no information rate of 59.45% and a kappa of 52.14%.

6 Chapter 6: Results and Discussions

6.1 Introduction

This chapter contains the researcher's findings during the study and particularly in relation to each of the objectives as set out at the beginning of the study. This study aimed to develop a classification model that predicts the likelihood of immunization completion. The sections will discuss the findings in relation to each specific objective.

6.2 Factors influencing immunization completion

In this study, basing on the variable importance output, the researcher found that the factors that were most important were the previous vaccine schedule attendance (polio2, dpt2, rotav2 and rotav1), then the counselling status and whether child was exposed to HIV. The least important variables noted were child sex, earlier vaccines received and the child's birth month.

From the literature reviewed many scholars were in argued that caretaker attitude, literacy levels, house hold income and health seeking behaviour influenced their response to immunization, this study confirms that caretaker behaviour indicated by the variables of previous schedule attendance, counselling and exposure to HIV greatly affected the target variable.

From the literature it was noted that vaccine logistics, location of the facilities and data quality affected immunization control, while this data was not available in the data set used in this study, when observing the dataset it was noticed that some children received a dose in one of the vaccines scheduled on the same day but missed the other or were late for it. From this the researcher assumed that it was due to vaccine shortages or ignorance of the caregiver on which vaccines to get from the facility.

In conclusion, the findings greatly relate with the studied literature especially caregiver involvement and knowledge in immunization programs. The other factors examined in the literature review but variables not present in the study dataset include levels of income, beliefs of the caregivers, location of the facility, distance from facility, availability of the child health card and the order of birth of the child. If present, they could have greatly contributed to the performance of the model and helped to make more informed conclusions.

6.3 Techniques used by existing Electronic Immunization Records Systems to support immunization completion

This study used the MyChild system's secondary data to train and develop this model. From the reviewed literature, it was noted that of all the systems used in most parts of East Africa that is

DHIS2, TIMR and the MyChild system, the DHIS2 is widely used and the rest were pilot implementations. All the studies on these systems indicated that their sole purpose was data collection and reporting as the main function. Therefore, these systems use data analysis techniques to analyse and report data, visualization, reports graphs and pivot tables.

It is noted that these systems are not exclusively electronic given the challenges with electricity supply, stationery and other maintenance issues, hence manual systems are still used in many an occasion. These are evident in the data inconsistencies that are brought about by human error like forgetting.

In conclusion, from this study, the model developed shows that data available in these systems could be further processed and used for decision making. The literature reviewed bears recommendations of improved data quality for better use of the data and systems integration which in regard to this study would have aided the researcher to develop a more robust and accurate model.

6.4 Models used to predict immunization completion

The models that have previously been used to predict immunization completion as studied in the literature in relation to this study were; one aimed at predicting children at a high risk of defaulting and classify them into two categories (Chandir et al., 2018) while another predicted at what point a child was likely to drop off (Qazi et al., 2020). This study while it also uses predictive analysis, it differs from both because it predicts the likelihood of completion into three classes.

In addition, both these studies use encapsulation of different machine learning techniques in their model while for this study, only Random Forest algorithm is used. This explains the disparity is accuracy levels, given that each technique used contributes to the performance metrics of the model.

While both models as reviewed had a limitation of generalizability, it is important to note that compared to this study, these studies had more socio-economic variables which contributed highly to the performance of the model. Such variables include location of the participants, languages spoken and wealth index among others. The model developed in this research study used only the data as available from the MyChild system and it did not contain socio-economic variables about the study group.

6.5 Developing a classification model to predict immunization completion

This study used a random forest classifying algorithm, using a dataset with 22 variables and 53,442 entries split into two datasets, training and testing data. The algorithm was trained on training data, hyperparameters tuned to use 4 mtry and 1,000 Ntrees which means at each split, four random variables were randomly sampled to create a bootstrapped dataset and 1000 decision trees were used in the model. The model when tested on test data it gave an output of 76.2% accuracy.

For this model to perform better in terms of accuracy, the number of trees would be increased at the cost of speed. Also increasing the number of relevant variables will increase its accuracy if the variables have a high importance in predicting the target variable.

The research model was mainly limited by the available data, using secondary data came with challenges of missing data, duplicates and withholding of data deemed confidential by the stakeholders which would have boosted the model performance especially details of the caregivers.

7 Chapter 7: Conclusion and Recommendation

7.1 Conclusions

In conclusion, the researcher set out to develop a model that predicts the likelihood of a child to complete immunization, and they found out that for this to be done, factors influencing immunization completion had to be put into consideration. These factors include the attendance history, health seeking behaviour, knowledge of the caregiver on immunization among others. Also a target variable that would be based on for the outcome was identified, in this study the third dose of DPT vaccine was used because it is a major indicator of a fully immunized child under the age of one (1), alongside measles, PCV and OPV.

In comparison to other models that have been developed for a similar purpose, this model introduces classification into three categories that is, timely, late or missed. This is to emphasize the importance of a vaccine being received within the expected timelines for it to be considered fully effective. In addition, this study shows that better performance of models is rooted in the amount of data used to increase the ability of the model to learn and ogive more accurate output.

It should be noted that given the quality and availability of data, the research did not consider some factors that would have been relevant and impactful. These are socio economic factors that include, education levels, languages spoken, income levels of the participant's household. This data was not available in the dataset neither is it collected at the facility in this case study. Therefore, it is diligent to note that while the results of this study are functional, including these variables in the dataset would have produces a more accurate practical result.

One contributions of this study is to add value to immunization data systems, in addition to reporting the system is able to show future predictions that can be used by the stakeholders to make impactful decisions. In this case, to use these predictions to lay follow up and outreach strategies to make sure that each child has an opportunity and is at less risk of missing their scheduled vaccine and enhance their chances of receiving their vaccine on time. Additionally, from this study, other scholars may build a better model or use the study to make further research on the topic.

7.2 Recommendations

Based on this study, the following are the recommendations.

In regard to the factors that influence completion, it is important that more variables are available in the data. These could be considered for inclusion, caretaker location, income index, levels of education, religion, availability of childbirth card on visit among others.

Also, it is recommended that health system is designed to integrate and operate alongside other systems to improve functionality of existing systems without having to create an independent system afresh.

7.3 Future Work

Incremental to this study, the use of socio-economic attributes like the language spoken, income levels, education levels of the caretakers would add greater insight to the study and produce more accurate results.

Use of pattern analysis as a probabilistic classifier may be used in a study such as this. The use of such algorithms gives an output with a confidence value associated with the result, hence increasing insight for stakeholders to make decisions.

References

Adamu, A. A., Uthman, O. A., Sambala, E. Z., Ndwandwe, D., Wiyeh, A. B., Olukade, T., Bishwajit, G., Yaya, S., Okwo-Bele, J. M., & Wiysonge, C. S. (2019). Rural-urban disparities in missed opportunities for vaccination in sub-Saharan Africa: a multi-country decomposition analyses. *Human Vaccines and Immunotherapeutics*, *15*(5), 1191–1198. https://doi.org/10.1080/21645515.2019.1575163

Adhikari, R., & Agrawal, R. K. (n.d.). *An Introductory Study on Time Series Modeling and Forecasting*.

Äijö, A., Schäffner, I., Waiswa, P., Kananura, R. M., Tessma, M. K., & Hanson, C. (2020). Assessment of a novel scanner-supported system for processing of child health and immunization data in Uganda. *BMC Health Services Research*, *20*(1), 1–9. https://doi.org/10.1186/s12913-020-05242-1

Al-Dosary, N. M. N., Al-Hamed, S. A., & Mohamed Aboukarima, A. (2019). K-Nearest neighbors method for prediction of fuel consumption in tractor-chisel plow systems. *Engenharia Agricola*, *39*(6), 729–736. https://doi.org/10.1590/1809-4430-Eng.Agric.v39n6p729-736/2019

Alegado, R. T., & Tumibay, G. M. (2019). Forecasting Measles Immunization Coverage Using ARIMA Model. *Journal of Computer and Communications*, *07*(10), 157–168. https://doi.org/10.4236/jcc.2019.710015

Alsaqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile Software Development: Methodologies and Trends Sentiment Analysis View project Blockchain technologies View project. *Article in International Journal of Interactive Mobile Technologies*. https://doi.org/10.3991/ijim.v14i11.13269

Altman, J. (2011). Advances in Intelligence and Soft Computing. In *Educacion*.

Babirye, J. N., Engebretsen, I. M. S., Makumbi, F., Fadnes, L. T., Wamani, H., Tylleskar, T., & Nuwaha, F. (2012a). Timeliness of Childhood Vaccinations in Kampala Uganda: A Community-Based Cross-Sectional Study. *PLoS ONE*, *7*(4), e35432. https://doi.org/10.1371/journal.pone.0035432

Babirye, J. N., Engebretsen, I. M. S., Makumbi, F., Fadnes, L. T., Wamani, H., Tylleskar, T., &

Nuwaha, F. (2012b). Timeliness of Childhood Vaccinations in Kampala Uganda: A Community-Based Cross-Sectional Study. *PLoS ONE*, *7*(4), e35432. https://doi.org/10.1371/journal.pone.0035432

Babirye, J. N., Rutebemberwa, E., Kiguli, J., Wamani, H., Nuwaha, F., & Engebretsen, I. M. (2011a). More support for mothers: A qualitative study on factors affecting immunisation behaviour in Kampala, Uganda. *BMC Public Health*, *11*(1), 1–11. https://doi.org/10.1186/1471-2458-11-723

Babirye, J. N., Rutebemberwa, E., Kiguli, J., Wamani, H., Nuwaha, F., & Engebretsen, I. M. (2011b). More support for mothers: A qualitative study on factors affecting immunisation behaviour in Kampala, Uganda. *BMC Public Health*, *11*(1), 723. https://doi.org/10.1186/1471-2458-11-723

Bbaale, E. (2013a). Factors influencing childhood immunization in Uganda. *Journal of Health, Population and Nutrition*, *31*(1), 118–127. https://doi.org/10.3329/jhpn.v31i1.14756

Bbaale, E. (2013b). Factors influencing childhood immunization in Uganda. *Journal of Health, Population and Nutrition*, *31*(1). https://doi.org/10.3329/jhpn.v31i1.14756

Bea, A., & Heydari, A. (2014). *Simple message delivery system The development process of a simple message delivery add-on to be used in a work environment*. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-188173

Budu, E., Seidu, A.-A., Agbaglo, E., Armah-Ansah, E. K., Dickson, K. S., Hormenu, T., Hagan, J. E., Adu, C., & Ahinkorah, B. O. (2021). Maternal healthcare utilization and full immunization coverage among 12–23 months children in Benin: a cross sectional study using population-based data. *Archives of Public Health*, *79*(1), 34. https://doi.org/10.1186/s13690-021-00554-y

Carnahan, E., Ferriss, E., Beylerian, E., Mwansa, F. D., Bulula, N., Lyimo, D., Kalbarczyk, A., Labrique, A. B., Werner, L., & Shearer, J. C. (2020). Determinants of Facility-Level Use of Electronic Immunization Registries in Tanzania and Zambia: An Observational Analysis. *Global Health: Science and Practice*. https://doi.org/10.9745/GHSP-D-20-00134

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ACM International Conference Proceeding Series*, *148*, 161–168.

https://doi.org/10.1145/1143844.1143865

Chandir, S., Siddiqi, D. A., Hussain, O. A., Niazi, T., Shah, M. T., Dharma, V. K., Habib, A., & Khan, A. J. (2018). Using Predictive Analytics to Identify Children at High Risk of Defaulting From a Routine Immunization Program: Feasibility Study. *JMIR Public Health and Surveillance*, *4*(3), e63. https://doi.org/10.2196/publichealth.9681

Chase, C. W. (2013). Demand-Driven Forecasting. In *Demand-Driven Forecasting*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118691861

Chopra, M., Bhutta, Z., Blanc, D. C., Checchi, F., Gupta, A., Lemango, E. T., Levine, O. S., Lyimo, D., Nandy, R., O'brien, K. L., Okwo-Bele, J. M., Rees, H., Soepardi, J., Tolhurst, R., & Victora, C. G. (2020). Addressing the persistent inequities in immunization coverage. In *Bulletin of the World Health Organization* (Vol. 98, Issue 2). https://doi.org/10.2471/BLT.19.241620

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 1–14. https://doi.org/10.1186/s12859-018-2264-5

Decision Tree Structure (Martínez et al., 2009) | Download Scientific Diagram. (n.d.). Retrieved April 27, 2021, from https://www.researchgate.net/figure/Decision-Tree-Structure-Martinez-et-al-2009_fig1_330262276

Dehnavieh, R., Haghdoost, A. A., Khosravi, A., Hoseinabadi, F., Rahimi, H., Poursheikhali, A., Khajehpour, N., Khajeh, Z., Mirshekari, N., Hasani, M., Radmerikhi, S., Haghighi, H., Mehrolhassani, M. H., Kazemi, E., & Aghamohamadi, S. (2019). The District Health Information System (DHIS2): A literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. In *Health Information Management Journal* (Vol. 48, Issue 2). https://doi.org/10.1177/1833358318777713

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 115, Issue 8, pp. 1690–1692). National Academy of Sciences. https://doi.org/10.1073/pnas.1800256115

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer*

Science and Information Technologies, *7*(3), 1174–1179. www.ijcsit.com

For health workers in Uganda, SMS reminders to parents are "the hope we needed" | by Shifo Foundation | Shifo News | Medium. (n.d.). Retrieved September 27, 2020, from https://medium.com/shifo-news/for-health-workers-in-uganda-sms-reminders-to-parents-are-the-hope-we-needed-9201bfa3c231

Frøen, J. F., Myhre, S. L., Frost, M. J., Chou, D., Mehl, G., Say, L., Cheng, S., Fjeldheim, I., Friberg, I. K., French, S., Jani, J. V., Kaye, J., Lewis, J., Lunde, A., Mørkrid, K., Nankabirwa, V., Nyanchoka, L., Stone, H., Venkateswaran, M., … Flenady, V. J. (2016). eRegistries: Electronic registries for maternal and child health. *BMC Pregnancy and Childbirth*, *16*(1). https://doi.org/10.1186/s12884-016-0801-7

Hemmati-Sarapardeh, A., Larestani, A., Nait Amar, M., & Hajirezaie, S. (2020). Intelligent models. In *Applications of Artificial Intelligence Techniques in the Petroleum Industry* (pp. 23–50). Elsevier. https://doi.org/10.1016/b978-0-12-818680-0.00002-3

Hosseinpoor, A. R., Bergen, N., Schlotheuber, A., Gacic-Dobo, M., Hansen, P. M., Senouci, K., Boerma, T., & Barros, A. J. D. (2016). State of inequality in diphtheria-tetanus-pertussis immunisation coverage in low-income and middle-income countries: a multicountry study of household health surveys. *The Lancet Global Health*, *4*(9), e617–e626. https://doi.org/10.1016/S2214-109X(16)30141-3

How to implement Random Forests in R | R-bloggers. (n.d.). Retrieved April 27, 2021, from https://www.r-bloggers.com/2018/01/how-to-implement-random-forests-in-r/

Howarth, A., Quesada, J., Silva, J., Judycki, S., & Mills, P. R. (2018). The impact of digital health interventions on health-related outcomes in the workplace: A systematic review. *DIGITAL HEALTH*, *4*, 205520761877086. https://doi.org/10.1177/2055207618770861

Huang, Y., & Danovaro-Holliday, M. C. (2021). Characterization of immunization secondary analyses using demographic and health surveys (DHS) and multiple indicator cluster surveys (MICS), 2006–2018. *BMC Public Health*, *21*(1), 1–14. https://doi.org/10.1186/s12889-021-10364-0

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer*.

https://www.springer.com/gp/book/9781461471370%0Ahttp://www.springer.com/us/book/9781461471370

Kamanda, B. C. (2010). *IMMUNIZATION COVERAGE AND FACTORS ASSOCIATED WITH FAILURE TO COMPLETE CHILDHOOD IMMUNIZATION IN KAWEMPE DIVISION, UGANDA BATARINGAYA COS KAMANDA*. University of the Western Cape. http://etd.uwc.ac.za/xmlui/handle/11394/2595

Karim, M., & Rahman, R. M. (2013). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications*, *06*(04), 196–206. https://doi.org/10.4236/jsea.2013.64025

Kiberu, V. M., Matovu, J. K., Makumbi, F., Kyozira, C., Mukooyo, E., & Wanyenze, R. K. (2014). Strengthening district-based health reporting through the district health management information software system: The Ugandan experience. *BMC Medical Informatics and Decision Making*, *14*(1). https://doi.org/10.1186/1472-6947-14-40

Kikoba, B. R., Kalinga, E., & Lungo, J. (2019). Integrating electronic medical records data into national health reporting system to enhance health data reporting and use at the facility level. *IFIP Advances in Information and Communication Technology*, *551*, 532–543. https://doi.org/10.1007/978-3-030-18400-1_44

Kim, K.-W., Li, G., Park, S.-T., & Ko, M.-H. (2016). A Study on Birth Prediction and BCG Vaccine Demand Prediction using ARIMA Analysis. *Indian Journal of Science and Technology*, *9*(24), 1–7. https://doi.org/10.17485/ijst/2016/v9i24/96151

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Li, A. J., Tabu, C., Shendale, S., Sergon, K., Okoth, P. O., Mugoya, I. K., Machekanyanga, Z., Onuekwusi, I. U., Sanderson, C., & Ogbuanu, I. U. (2020). Assessment of missed opportunities for vaccination in Kenyan health facilities, 2016. *PLoS ONE*, *15*(8 August). https://doi.org/10.1371/journal.pone.0237913

M, E. M. (2020). A Review on Success Factors of Agile Software Development. *International*

Journal of Applied Science & Engineering, *8*(1). https://doi.org/10.30954/2322-0465.1.2020.4

Machine Learning for Subsurface Characterization | Siddharth Misra, Hao Li, Jiabo He | download. (n.d.). Retrieved April 6, 2021, from https://b-ok.africa/book/5503972/d50f15

Machingaidze, S., Wiysonge, C. S., & Hussey, G. D. (2013). Strengthening the Expanded Programme on Immunization in Africa: Looking beyond 2015. *PLoS Medicine*, *10*(3), e1001405. https://doi.org/10.1371/journal.pmed.1001405

Maïga, A., Jiwani, S. S., Mutua, M. K., Porth, T. A., Taylor, C. M., Asiki, G., Melesse, D. Y., Day, C., Strong, K. L., Faye, C. M., Viswanathan, K., O'Neill, K. P., Amouzou, A., Pond, B. S., & Boerma, T. (2019). Generating statistics from health facility data: The state of routine health information systems in Eastern and Southern Africa. In *BMJ Global Health* (Vol. 4, Issue 5). https://doi.org/10.1136/bmjgh-2019-001849

Malande, O. O., Munube, D., Afaayo, R. N., Annet, K., Bodo, B., Bakainaga, A., Ayebare, E., Njunwamukama, S., Mworozi, E. A., & Musyoki, A. M. (2019a). Barriers to effective uptake and provision of immunization in a rural district in Uganda. *PLoS ONE*, *14*(2). https://doi.org/10.1371/journal.pone.0212270

Malande, O. O., Munube, D., Afaayo, R. N., Annet, K., Bodo, B., Bakainaga, A., Ayebare, E., Njunwamukama, S., Mworozi, E. A., & Musyoki, A. M. (2019b). Barriers to effective uptake and provision of immunization in a rural district in Uganda. *PLOS ONE*, *14*(2), e0212270. https://doi.org/10.1371/journal.pone.0212270

Malande, O. O., Munube, D., Afaayo, R. N., Annet, K., Bodo, B., Bakainaga, A., Ayebare, E., Njunwamukama, S., Mworozi, E. A., & Musyoki, A. M. (2019c). Barriers to effective uptake and provision of immunization in a rural district in Uganda. *PLOS ONE*, *14*(2), e0212270. https://doi.org/10.1371/journal.pone.0212270

Mannion, N. (2020). *Predictions of Changes in Child Immunization Rates Using an Automated Approach: USA*. http://norma.ncirl.ie/4368/1/niallmannion.pdf

Mansotra, V. (2019). A Model for Accurate Prediction of Child Immunization Data for Knowledge Discovery using Bayesian TAN and Naive Bayes Classifiers. *International Journal of Recent Technology and Engineering*, *8*(4), 3335–3343.

https://doi.org/10.35940/ijrte.d8118.118419

Mekonnen, Z. A., Mekonnen, Z. A., Gelaye, K. A., Were, M. C., & Tilahun, B. (2020). Timely completion of vaccination and its determinants among children in northwest, Ethiopia: A multilevel analysis. *BMC Public Health*, *20*(1). https://doi.org/10.1186/s12889-020-08935-8

Menzies, N. A., Suharlim, C., Resch, S. C., & Brenzel, L. (2020). The efficiency of routine infant immunization services in six countries: A comparison of methods. *Health Economics Review*, *10*(1), 1–11. https://doi.org/10.1186/s13561-019-0259-1

Mutua, Martin K., Kimani-Murage, E., & Ettarh, R. R. (2011). Childhood vaccination in informal urban settlements in Nairobi, Kenya: Who gets vaccinated? *BMC Public Health*, *11*(1), 1–11. https://doi.org/10.1186/1471-2458-11-6

Mutua, Martin K., Mohamed, S. F., Porth, J. M., & Faye, C. M. (2021). Inequities in On-Time Childhood Vaccination: Evidence From Sub-Saharan Africa. *American Journal of Preventive Medicine*, *60*(1), S11–S23. https://doi.org/10.1016/j.amepre.2020.10.002

Mutua, Martin Kavao, Kimani-Murage, E., Ngomi, N., Ravn, H., Mwaniki, P., & Echoka, E. (2016). Fully immunized child: coverage, timing and sequencing of routine immunization in an urban poor settlement in Nairobi, Kenya. *Tropical Medicine and Health*, *44*(1), 13. https://doi.org/10.1186/s41182-016-0013-x

Nakatudde, I., Rujumba, J., Namiiro, F., Sam, A., Mugalu, J., & Musoke, P. (2019). Vaccination timeliness and associated factors among preterm infants at a tertiary hospital in Uganda. *PLOS ONE*, *14*(9), e0221902. https://doi.org/10.1371/journal.pone.0221902

Nawaz, R., Khan, S. A., & Khan, G. S. (2015). Swot Analysis of District Health Information System in Khyber Pakhtunkhwa. *Journal of Medical Science*, *13*(2), 109–125.

Ndiritu, M., Cowgill, K. D., Ismail, A., Chiphatsi, S., Kamau, T., Fegan, G., Feikin, D. R., Newton, C. R. J. C., & Scott, J. A. G. (2006). Immunization coverage and risk factors for failure to immunize within the Expanded Programme on Immunization in Kenya after introduction of new Haemophilus influenzae type b and hepatitis b virus antigens. *BMC Public Health*, *6*(1), 1–8. https://doi.org/10.1186/1471-2458-6-132

Nordhausen, K. (2014). An Introduction to Statistical Learning-with Applications in R by Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani. *International Statistical*

Review, *82*(1), 156–157. https://doi.org/10.1111/insr.12051_19

Nsubuga, F., Kabwama, S. N., Ampeire, I., Luzze, H., Gerald, P., Bulage, L., & Toliva, O. B. (2019). Comparing static and outreach immunization strategies and associated factors in Uganda, Nov-Dec 2016. *Pan African Medical Journal*, *32*. https://doi.org/10.11604/pamj.2019.32.123.16093

Okot, G. (2015). *Assessing Immunisation Dropout Rates among Children Under One Year of Age in Aswa County, Gulu District, 2015*. http://dspace.ciu.ac.ug/xmlui/handle/123456789/892

Oleribe, O., Kumar, V., Awosika-Olumo, A., & Taylor-Robinson, S. D. (2017). Individual and socioeconomic factors associated with childhood immunization coverage in Nigeria. *Pan African Medical Journal*, *26*. https://doi.org/10.11604/pamj.2017.26.220.11453

On the way to reducing the workload for Ugandan health workers | by Shifo Foundation | Shifo News | Medium. (n.d.). Retrieved September 27, 2020, from https://medium.com/shifo-news/on-the-way-to-reducing-the-workload-for-ugandan-health-workers-b6d6729f36d1

Oryema, P., Babirye, J. N., Baguma, C., Wasswa, P., & Guwatudde, D. (2017). Utilization of outreach immunization services among children in Hoima District, Uganda: a cluster survey. *BMC Research Notes*, *10*(1), 111. https://doi.org/10.1186/s13104-017-2431-1

Paul, S., & Bhatia, D. (2020). Smart healthcare for disease diagnosis and prevention. In *Smart Healthcare for Disease Diagnosis and Prevention*. Elsevier. https://doi.org/10.1016/C2018-0-03178-3

Pochet, N. L. M. M., & Suykens, J. A. K. (2006). Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. *Ultrasound in Obstetrics and Gynecology*, *27*(6), 607–608. https://doi.org/10.1002/uog.2791

Qazi, S., Usman, M., & Mahmood, A. (2020). A data-driven framework for introducing predictive analytics into exQazi, S., Usman, M. and Mahmood, A. (2020) 'A data-driven framework for introducing predictive analytics into expanded program on immunization in Pakistan', Wiener Klinische Wochenschrift. *Wiener Klinische Wochenschrift*. https://doi.org/10.1007/s00508-020-01737-3

Rahman, M., & Obaida-Nasrin, S. (2010). Factors affecting acceptance of complete

immunization coverage of children under five years in rural Bangladesh. *Salud Pública de México*, *52*(2). https://doi.org/10.1590/s0036-36342010000200005

Rainey, J. J., Watkins, M., Ryman, T. K., Sandhu, P., Bo, A., & Banerjee, K. (2011). Reasons related to non-vaccination and under-vaccination of children in low and middle income countries: Findings from a systematic review of the published literature, 1999-2009. In *Vaccine* (Vol. 29, Issue 46). https://doi.org/10.1016/j.vaccine.2011.08.096

Rampisela, T. V, & Rustam, Z. (2018). Classification of Schizophrenia Data Using Support Vector Machine (SVM) Recent citations Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification Arfiani et al Classification of Schizophrenia Data Using Support Vector Machine (SVM). *IOP Conf. Series: Journal of Physics: Conf. Series*, *1108*, 12044. https://doi.org/10.1088/1742-6596/1108/1/012044

Random Forest - an overview | ScienceDirect Topics. (n.d.-a). Retrieved April 6, 2021, from https://www.sciencedirect.com/topics/engineering/random-forest

Random Forest - an overview | ScienceDirect Topics. (n.d.-b). Retrieved April 27, 2021, from https://www.sciencedirect.com/topics/engineering/random-forest

Random Forest Algorithm- An Overview | Understanding Random Forest. (n.d.). Retrieved April 27, 2021, from https://www.mygreatlearning.com/blog/random-forest-algorithm/

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research*, *8*(3), 148–151. https://doi.org/10.4103/picr.PICR_87_17

Resampling Methods · UC Business Analytics R Programming Guide. (n.d.). Retrieved April 27, 2021, from https://uc-r.github.io/resampling_methods

Restrepo-Méndez, M. C., Barros, A. J. D., Wong, K. L. M., Johnson, H. L., Pariyo, G., Wehrmeister, F. C., & Victora, C. G. (2016). Missed opportunities in full immunization coverage: Findings from low- and lower-middle-income countries. *Global Health Action*, *9*(1). https://doi.org/10.3402/gha.v9.30963

Sarker, I. H., Alqahtani, H., Alsolami, F., Khan, A. I., Abushark, Y. B., & Siddiqui, M. K. (2020). Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. *Journal of Big Data*, *7*(1), 51.

https://doi.org/10.1186/s40537-020-00328-3

Satya Sahisnu, J., Natalia, F., Vincenttius Ferdinand, F., Sudirman, S., & Seong Ko, C. (2020). VACCINE PREDICTION SYSTEM USING ARIMA METHOD. *ICIC Express Letters Part B: Applications ICIC International*, *2020*(6), 567–575. https://doi.org/10.24507/icicelb.11.06.567

Seymour, D., Werner, L., Mwansa, F. D., Bulula, N., Mwanyika, H., Dube, M., Taliesin, B., & Settle, D. (2019). Electronic Immunization Registries in Tanzania and Zambia: Shaping a Minimum Viable Product for Scaled Solutions. *Frontiers in Public Health*, *7*. https://doi.org/10.3389/fpubh.2019.00218

Shastri, S., Kour, P., Gupta, A., Sambyal, S., Singh Bhadwal, A., Sharma, A., Mansotra, V., & Sharma, A. (2018). Development of a Data Mining Based Model for Classification of Child Immunization Data. *International Journal of Computational Engineering Research*, *8*(6), 41–49. www.ijceronline.com

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Sowe, A., & Gariboldi, M. I. (2020). An assessment of the quality of vaccination data produced through smart paper technology in The Gambia. *Vaccine*, *38*(42), 6618–6626. https://doi.org/10.1016/j.vaccine.2020.07.074

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Sullivan, M. C., Tegegn, A., Tessema, F., Galea, S., & Hadley, C. (2010). Minding the immunization gap: Family characteristics associated with completion rates in rural Ethiopia. *Journal of Community Health*, *35*(1), 53–59. https://doi.org/10.1007/s10900-009-9192-2

Tamirat, K. S., & Sisay, M. M. (2019). Full immunization coverage and its associated factors among children aged 12-23 months in Ethiopia: Further analysis from the 2016 Ethiopia demographic and health survey. *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-019-7356-2

UNICEF. (2015). *UNICEF's Approach to Digital Health*.

Usman, H. R., Kristensen, S., Rahbar, M. H., Vermund, S. H., Habib, F., & Chamot, E. (2010). Determinants of third dose of diphtheria-tetanus-pertussis (DTP) completion among children who received DTP1 at rural immunization centres in Pakistan: A cohort study. *Tropical Medicine and International Health*, *15*(1), 140–147. https://doi.org/10.1111/j.1365-3156.2009.02432.x

Vouking, M. Z., Mengue, C. M. A., Yauba, S., Edengue, J. M., Dicko, M., Dicko, H. M., & Wiysonge, C. S. (2019). Interventions to increase the distribution of vaccines in sub-saharan africa: A scoping review. In *Pan African Medical Journal* (Vol. 32). https://doi.org/10.11604/pamj.2019.32.14.17225

Ward, K., Mugenyi, K., MacNeil, A., Luzze, H., Kyozira, C., Kisakye, A., Matseketse, D., Newall, A. T., Heywood, A. E., Bloland, P., & Pallas, S. W. (2020). Financial cost analysis of a strategy to improve the quality of administrative vaccination data in Uganda. *Vaccine*, *38*(5), 1105–1113. https://doi.org/10.1016/j.vaccine.2019.11.030

Wariri, O., Edem, B., Nkereuwem, E., Nkereuwem, O. O., Umeh, G., Clark, E., Idoko, O. T., Nomhwange, T., & Kampmann, B. (2019). Tracking coverage, dropout and multidimensional equity gaps in immunisation systems in West Africa, 2000-2017. *BMJ Global Health*, *4*(5), 1–10. https://doi.org/10.1136/bmjgh-2019-001713

Werner, L., Seymour, D., Puta, C., & Gilbert, S. (2019). Three waves of data use among health workers: The experience of the better immunization data initiative in Tanzania and Zambia. *Global Health Science and Practice*, *7*(3). https://doi.org/10.9745/GHSP-D-19-00024

WHO. (2018). *20 Million Children Miss Out on Lifesaving Measles, Diphtheria and Tetanus Vaccines in 2018*. 2018–2021.

World Health Organization. (2019). Immunization coverage. In *Fact sheet*.

Zhang, Z., Zhao, Z., & Yeom, D. S. (2020). Decision Tree Algorithm-Based Model and Computer Simulation for Evaluating the Effectiveness of Physical Education in Universities. *Complexity*, *2020*. https://doi.org/10.1155/2020/8868793

Zida-Compaore, W. I. C., Ekouevi, D. K., Gbeasor-Komlanvi, F. A., Sewu, E. K., Blatome, T., Gbadoe, A. D., Agbèrè, D. A., & Atakouma, Y. (2019). Immunization coverage and factors associated with incomplete vaccination in children aged 12 to 59 months in health structures in Lomé. *BMC Research Notes*, *12*(1), 84. https://doi.org/10.1186/s13104-019-4115-5

Appendices

Appendix A: Data cleaning, transformation and standardization

```
clear
set more off
/*use "C:\Users\userx\Desktop\UG_Mukono_register_Nov_2020 (1)_Bertha.csv", clear*/
insheet using"C:\Users\userx\Desktop\UG_Mukono_register_Nov_2020 (1)_Bertha.csv", clear

**Check for duplicate variables (based on ID)
duplicates report public_id
duplicates drop

**Date of birth
*Tranformation of dates from CSV to STATA format
generate int birth=date( birth_date,"DMY",2099)
format birth %td
label var birth "Child birht date"
drop birth_date

**Generate year of birth
gen year=year(birth)
gen month=month(birth)

**Limit analysis to children born between 2015 and 2020
drop if year<=2014
drop if year>=2021
label var year "Year of birth"

**Drop Children born after June 2020
drop if year==2020 & month>=7

*Due dates for immunisation(birth, 6 weeks, 10 weeks, 14 weeks)
gen day1=birth
gen day2=(birth+42)
gen day3=(birth+70)
gen day4= (birth+98)
```

```
format day1 %td
format day2 %td
format day3 %td
format day4 %td

*Due months for immunisation-based on (birth, 6 weeks, 10 weeks, 14 weeks)
gen month_day1=month(day1)
gen month_day2=month(day2)
gen month_day3=month(day3)
gen month_day4=month(day4)
label var month_day1 " Vaccines at birth"
label var month_day2 " Vaccines at 6 weeks"
label var month_day3 " Vaccines at 10 weeks"
label var month_day3 " Vaccines at 14 weeks"



**Date for IPV
generate int ipv_date=date(ipv,"DMY",2099)
format ipv_date %td
label var ipv_date "IPV administration date"
drop ipv

**Date of DP3
generate int dpt_3_date=date(dpt_3,"DMY",2099)
format dpt_3_date %td
label var dpt_3_date "DP3 administration date"
drop dpt_3

**Date of OPV 3
generate int opv_3_date=date(opv_3,"DMY",2099)
format opv_3_date %td
label var opv_3_date "DP3 administration date"
drop opv_3
```

```
**Date of PCV 3
generate int pcv_3_date=date(pcv_3,"DMY",2099)
format pcv_3_date %td
label var pcv_3_date "DP3 administration date"
drop pcv_3

**Date of Measles 1
generate int measles_1_date=date(measles_1,"DMY",2099)
format measles_1_date %td
label var measles_1_date "Measles administration date"
drop measles_1

**Date of BCG
generate int bcg_date=date(bcg,"DMY",2099)
format bcg_date %td
label var bcg_date "bcg administration date"
drop bcg


*Date of Rotavirus 1
generate int rv_1_date=date(rv_1,"DMY",2099)
format rv_1_date %td
label var rv_1_date "Rota administration date"
drop rv_1

*Date of Rotavirus 2
generate int rv_2_date=date(rv_2,"DMY",2099)
format rv_2_date %td
label var rv_2_date "DP3 administration date"
drop rv_2


*Date of DPT 1
generate int dpt_1_date=date(dpt_1,"DMY",2099)
```

```
format dpt_1_date %td
label var dpt_1_date "DPT 1 administration date"
drop dpt_1

*Date of DPT 2
generate int dpt_2_date=date(dpt_2,"DMY",2099)
format dpt_2_date %td
label var dpt_2_date "DPT 2 administration date"
drop dpt_2

*Date of OPV O
generate int opv_0_date=date(opv_0,"DMY",2099)
format opv_0_date %td
label var opv_0_date "OPV0 administration date"
drop opv_0

*Date of OPV 1
generate int opv_1_date=date(opv_1,"DMY",2099)
format opv_1_date %td
label var opv_1_date "OPV1 administration date"
drop opv_1

*Date of OPV 2
generate int opv_2_date=date(opv_2,"DMY",2099)
format opv_2_date %td
label var opv_2_date "OPV 2 administration date"
drop opv_2

*Date of PCV 1
generate int pcv_1_date=date(pcv_1,"DMY",2099)
format pcv_1_date %td
label var pcv_1_date "PCV 1 administration date"
drop pcv_1
```

```
*Date of PCV 2
generate int pcv_2_date=date(pcv_2,"DMY",2099)
format pcv_2_date %td
label var pcv_2_date "PCV 2 administration date"
drop pcv_2


 **Date of Measles 2
generate int measles_2_date=date(measles_2,"DMY",2099)
format measles_2_date %td
label var measles_2_date "Measles administration date"
drop measles_2

*****Estimation of delays
*Generate date due for the vaccine

sca sixweekv= (birth+42)
sca tenweekv=(birth+70)
sca fourteenweekv=(birth+98)

/*how to deal with negatives-drop*/
/*generate categories of ontime, delay, did not get*/
*Vaccines administered at birth (BCG,OPV0)- 0 days
gen delay_bcg1=bcg_date -birth
gen delay_polio=opv_0_date-birth


*Vaccines administered at 6 weeks (Polio1, DPT1,PCV1, Rota1)-42 days
gen delay_polio1=opv_1_date-(birth+42)
gen delay_dpt1=dpt_1_date-(birth+42)
```

```
gen delay_pcv1=pcv_1_date-(birth+42)
gen delay_rotav1=rv_1_date-(birth+42)


*Vaccines administered at 10 weeks-70days
gen delay_polio2=opv_2_date-(birth+70)
gen delay_dpt2=dpt_2_date-(birth+70)
gen delay_pcv2=pcv_2_date-(birth+70)
gen delay_rotav2=rv_2_date-(birth+70)

*Vaccines administered at 14 weeks-98 days
gen delay_polio3=opv_3_date-(birth+98)
gen delay_dpt3=dpt_3_date-(birth+98)
gen delay_pcv3=pcv_3_date-(birth+98)


**check
sum delay_polio3 delay_dpt3 delay_pcv3
sum delay_polio2 delay_dpt2 delay_pcv2 delay_rotav2
sum delay_polio1 delay_dpt1 delay_pcv1 delay_rotav1


*One vaccine for each based on the one with better date?


*Birth
gen bcg1=.
replace bcg1=1 if delay_bcg1==0
replace bcg1=2 if delay_bcg1 >0 & delay_bcg1 <.
replace bcg1=3 if delay_bcg1==.
label define bcg1 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label value bcg1
label var bcg1 "BCG immunisation status"
```

```
gen polio=.
replace polio=1 if delay_polio==0
replace polio=2 if delay_polio >0 & delay_polio <.
replace polio=3 if delay_polio==.
label define polio 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var polio "Polio immunisation status"

*

**At 6 weeks
gen polio1=.
replace polio1=1 if delay_polio1==0
replace polio1=2 if delay_polio1 >0 & delay_bcg1 <.
replace polio1=3 if delay_polio1==.
label define polio1 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var polio1 "Polio1 immunisation status"

gen dpt1=.
replace dpt1=1 if delay_dpt1==0
replace dpt1=2 if delay_dpt1 >0 & delay_dpt1 <.
replace dpt1=3 if delay_dpt1==.
label define dpt1 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var dpt1 "DPT immunisation status"

gen pcv1=.
replace pcv1=1 if delay_pcv1==0
replace pcv1=2 if delay_pcv1 >0 & delay_pcv1 <.
replace pcv1=3 if delay_pcv1==.
label define pcv1 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var dpt1 "PCV1 immunisation status"

gen rotav1=.
replace rotav1=1 if delay_rotav1==0
replace rotav1=2 if delay_rotav1 >0 & delay_rotav1 <.
```

```
replace rotav1=3 if delay_rotav1==.
label define rotav1 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var rotav1 "ROTA1 immunisation status"

*

** At 10 weeks
gen polio2=.
replace polio2=1 if delay_polio2==0
replace polio2=2 if delay_polio2 >0 & delay_polio2 <.
replace polio2=3 if delay_polio2==.
label define polio2 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var polio1 "Polio2 immunisation status"

gen dpt2=.
replace dpt2=1 if delay_dpt2==0
replace dpt2=2 if delay_dpt2 >0 & delay_dpt2 <.
replace dpt2=3 if delay_dpt2==.
label define dpt2 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var dpt2 "DPT immunisation status"

gen pcv2=.
replace pcv2=1 if delay_pcv1==0
replace pcv2=2 if delay_pcv1 >0 & delay_pcv1 <.
replace pcv2=3 if delay_pcv1==.
label define pcv2 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var pcv2 "PCV2 immunisation status"

gen rotav2=.
replace rotav2=1 if delay_rotav2==0
replace rotav2=2 if delay_rotav2 >0 & delay_rotav2 <.
replace rotav2=3 if delay_rotav2==.
label define rotav2 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var rotav2 "ROTA2 immunisation status"
```

```
*

** At 14 weeks
gen polio3=.
replace polio3=1 if delay_polio3==0
replace polio3=2 if delay_polio3 >0 & delay_polio3 <.
replace polio3=3 if delay_polio3==.

label define polio3 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var polio3 "Polio3 immunisation status"

gen dpt3=.
replace dpt3=1 if delay_dpt3==0
replace dpt3=2 if delay_dpt3 >0 & delay_dpt3 <.
replace dpt3=3 if delay_dpt3==.

label define dpt3 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var dpt3 "DPT immunisation status"

gen pcv3=.
replace pcv3=1 if delay_pcv3==0
replace pcv3=2 if delay_pcv3 >0 & delay_pcv3 <.
replace pcv3=3 if delay_pcv3==.

label define pcv3 1 "On time" 2 "Late" 3 "Missed/Not Recieved"
label var pcv3 "PCV3 immunisation status"


**********************************************************
**********************************************************
**Independent variables
**sex
encode sex, generate(childsex)
label var childsex "Child Sex"
```

```
**Protection at Birth- Vaccine acceptance of the mother?
replace pab="NA" if pab==""
encode pab, generate(motherpab)
label var motherpab "mother protected at birth"
tab motherpab

**Mothers health seeking behaviour?
replace counseling="FALSE" if counseling==""
encode counseling, generate(conselled)
tab counseling


**Child has been HIV exposed
replace hiv_exposed="NA" if hiv_exposed==""
encode hiv_exposed, generate(hiv_exposed1)
tab hiv_exposed1
lab define newlabel 1 "No" 2 "NA" 3 "Yes", modify
lab val hiv_exposed1 newlabel


**Has the child been HIV tested
replace hiv_exposed_tested="NA" if hiv_exposed_tested==""
encode hiv_exposed_tested, generate(exposed_tested)
tab exposed_tested
lab define newlabel 1 "No" 2 "NA" 3 "Yes", modify
lab val exposed_tested newlabel



drop sex pab oedema birth_weight registry_weight_for_age ///
 llin counseling height registry_height_length_for_age ///
 disability registry_weight_for_height muac hiv_exposed_tested ///
 hiv_exposed hiv_tested vitamin_a deworming feeding_code_0_1 ///
 feeding_code_6_7 feeding_code_7_24 ipv_date dpt_3_date opv_3_date ///
```

```
 pcv_3_date measles_1_date bcg_date rv_1_date rv_2_date dpt_1_date ///
 dpt_2_date opv_0_date opv_1_date opv_2_date pcv_1_date pcv_2_date ///
 measles_2_date delay_* day*

export delimited using "C:\Users\userx\Desktop\Immunisation Registry Clean.csv", replace
file C:\Users\userx\Desktop\Immunisation Registry Clean.csv saved
```

Appendix B: R implementation code

```
#load packages

library(tidyr)
library(caret)
library(randomForest)
library(party)
library(rsample)
library(readr)

# load data

registry <-mukono_immunization_model_data
view(registry)
str(registry)
dim(registry)
summary(registry)

#clean data of null variables
clean = registry %>% drop_na()
dim(clean)

#convert to factors
convert <- lapply(clean[2:22], factor)
View(convert)

#bind with ID
immunization <- cbind(clean[1], convert)
dim(immunization)
str(immunization)



#Split data into training and testing
set.seed(123)
# choosing 75% of the data to be the training data
data_split <- initial_split(immunization, prop = .75)
```

```
# extracting training data and test data as two seperate dataframes
train <- training(data_split)
test  <- testing(data_split)

head(train)
head(test)


#random forest before hyperparameter tuning, using defaults

def_cat = randomForest(dpt3~., data=train[-1],
                       importance = TRUE, ntree = 500 ,mtry = 2)
print(def_cat)
#get optimal mtry
mtry <- tuneRF(train[-1], train$dpt3, ntreeTry=1000,
               stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE)
best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]
print(mtry)
print(best.m)


#random forest

classifier = randomForest(dpt3~., data=train[-1],
                       importance = TRUE, ntree = 1000,mtry = 4,
                       replace=TRUE, random_state = 0)
print(classifier)

#Variable Importance
randomForest::importance(classifier)
varImpPlot(classifier)

#check with test
pred = predict(classifier, newdata = test, type = "class")
pred
```

```
plot(pred)

#confusion matrix
confusionMatrix(pred, test$dpt3)
```

Appendix C: Strathmore University Ethics Review Approval Certificate

Strathmore UNIVERSITY

4th May 2021

Ms Kembabazi Bertha
bertha.kembabazi@strathmore.edu

Dear Ms Kembabazi,

RE: A Classification Model Leveraging Electronic Immunization Records to Predict Child Immunization Completion. Case Study: Mukono Health Facility

This is to inform you that SU-IERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-IERC0933/20.** The approval period is **4th May 2021 to 3rd May 2022.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and also obtain other clearances needed

Yours sincerely,

for: Dr Virginia Gichuru,
Secretary; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE
(SU-IERC)
04 May 2021
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Cc: Prof Fred Were,
Chairperson; SU-IERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix D: Originality Summary report

Curiginal

Document Information

Analyzed document	A classification model leveraging Electronic Immunization Records to predict Child Immunization completion.docx (D109940413)
Submitted	6/30/2021 11:17:00 AM
Submitted by	
Submitter email	Bertha.Kembabazi@strathmore.edu
Similarity	5%
Analysis address	library.strath@analysis.urkund.com