The Nelson Mandela AFrican Institution of Science and Technology

NM-AIST Repository **https://dspace.mm-aist.ac.tz**

Computational and Communication Science Engineering Masters Theses and Dissertations [CoCSE]

2021-10

A predictive model for early detection of diabetes mellitus using machine learning

Semakula, Henry

NM-AIST

https://dspace.nm-aist.ac.tz/handle/20.500.12479/1592

Provided with love from The Nelson Mandela African Institution of Science and Technology

A PREDICTIVE MODEL FOR EARLY DETECTION OF DIABETES MELLITUS USING MACHINE LEARNING

Henry Semakula

A Project Report Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

October, 2021

ABSTRACT

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose or blood sugar that over time can bring severe damage to vital organs including the heart, blood vessels, eyes, kidneys and nerves. Diabetes is therefore one of the major priorities in medical science research. Type 2 diabetes is common in adults, either because of inadequate insulin production, or when the body's cells fail to respond properly to the produced insulin. For all the diabetes cases, it's found out that 90% are Type 2 diabetes. Of the 422 million people with diabetes worldwide, 336 million people are found in developing countries, and 1.6 million people die of diabetes each year according to statistics by the World Health Organization.

Around 19.8 million adults in Africa have Type 2 diabetes but approximately 75% are unaware of their condition (undiagnosed). Most people are undiagnosed because many people lack knowledge of symptoms for diabetes, and others are not diagnosed due to lack of testing kits more especially in rural areas. African governments have scaled up purchasing and distribution of diagnostic kits but the majority of the population has not been reached. Researchers have been developing predictive models for Type 2 diabetes, but African populations are not widely included in their datasets. The developed models may therefore not accurately identify at-risk populations in the African context. The main emphasis of this research was to come up with a machine learning prediction model to find out Ugandans likely to be suffering from Type 2 diabetes (output classes: high risk or low risk), based on input symptoms.

Random Forest, Support Vector Machine, Naïve Bayes, and AdaBoost classifiers were trained on anonymised, real patient data with twelve features including age, gender (male or female), systolic blood pressure, residence (town or village), diastolic blood pressure, family Member with diabetes, alcohol intake, smoker, hypertensive, obesity, physically inactive and body mass index. This research's experimental results after the comparison of the Accuracy Score and Confusion Matrix for all the above algorithms, the Random Forest classifier emerged the premier with the accuracy score of 85.4%, thus the experimental results shown that performance of Random Forest classifier as being significant superior compared to all other the machine learning algorithms.

DECLARATION

I, Henry Semakula, do hereby declare to the Senate of The Nelson Mandela African Institution of Science and Technology that this dissertation titled "*A Predictive Model for Early Detection of Diabetes Mellitus Using Machine Learning*" is my own original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution

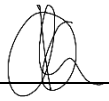
Henry Semakula

15th/11/2021

Name and Signature of the Candidate Date

The above declaration is confirmed by:

Dr. Michael Kisangiri



15th/11/2021

Name and Signature of Supervisor 1 Date

Dr. Edith Luhanga

15th/11/2021

Name and Signature of Supervisor 2 Date

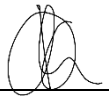
COPYRIGHT

This dissertation is copyright material protected under the Berne Convention, the Copyright Act of 1999 and other international and national enactments, in that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgment, without a written permission of the Deputy Vice Chancellor for Academic, Research and Innovation, on behalf of both the author and the Nelson Mandela African Institution of Science and Technology.

CERTIFICATION

The undersigned certify that have read and hereby recommend for acceptance by the Nelson Mandela African Institution of Science and Technology a dissertation titled, "*A Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning*" in partial fulfillment of the requirements for the degree of Master of Science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology.

Dr. Michael Kisangiri



15th/11/2021

Name and Signature of Supervisor 1 Date

Dr. Edith Luhanga

15th/11/2021

Name and Signature of Supervisor 2 Date

ACKNOWLEDGEMENT

Above all, I offer great gratitude to the Almighty God, for having blessed me and granted me a very big chance of joining a master's program at Nelson Mandela African Institution of Science and Technology in Arusha Tanzania (NM-AIST). I give gratitude to the Almighty God for the gift of life, intelligence and for the boldness, strength, good physical, mental and social well-being throughout my study period at NM-AIST as they were the main rationale towards this great attainment.

I convey my sincere exceptional gratitude to Centre of Excellence for Information Communication Technology in East Africa (CENIT@EA project) my sponsors for the scholarship they gave me and all the other monetary support rendered to me while I was at NM-AIST. I further express special thanks to my supervisors Dr. Michael Kisangiri and Dr. Edith Luhanga of the School of Computational and Communication Science and Engineering (CoCSE) at NM-AIST for their support and guidance they provided to me every time whenever I had a difficulty or any challenging situation regarding my research. Their offices were open every time and they were always welcoming and very concerned whenever I had any challenge during my research period and always gave help and support of all kinds.

Furthermore, I acknowledge my colleagues that I met at my internship place The Medical Concierge Group (TMCG) for their wonderful support and collaboration. I particularly like to single out my industrial supervisor at TMCG Dr. John Mark Bwanika. I thank him for the continuous support and for all the opportunities he gave me during my internship period at TMCG.

I am forever and I will always be thankful to everyone at NM-AIST that added value to my study including my colleagues, course mates, lecturers and friends at large for their significant role they played, they contributed a lot as they were always social, caring, concerned and cooperative during my study at NM-AIST.

Lastly, I thank my parents for their sympathetic ear and wise counsel. They were always there for me. Finally, completing this dissertation without the support of my friends would be as hard as turning water to wine. I thank my friends Mr. Clinton Chikwata and Mr. Joan Atuhaire for providing me with discussion time as well as happy distractions to rest my brain besides my research. Thanks a lot indeed and blessings to everyone.

DEDICATION

This dissertation is dedicated to my one and only lovely son Keith Hiram Semakula.

TABE OF CONTENTS

ABSTRACT i

DECLARATION ii

COPYRIGHT iii

CERTIFICATION iv

ACKNOWLEDGEMENT v

DEDICATION vi

TABE OF CONTENTS vii

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF APPENDICES xii

LIST OF ABBREVIATIONS AND SYMBOLS xiii

CHAPTER ONE 1

INTRODUCTION 1

1.1 Background of the Problem 1

1.2 Statement of the Problem 2

1.3 Rationale of the Study 3

1.4 Project Objectives 3

1.4.1 Main Objective 3

1.4.2 Specific Objectives 3

1.5 Research Questions 4

1.6 Significance of the Study 4

1.7 Delineation of the Study 4

CHAPTER TWO 6

LITERATURE REVIEW 6

2.1. Introduction 6

2.2. Machine Learning in Diabetes Predictions 6

2.2.1 Datasets 6
2.3. Benefits of Machine Learning in Diabetes Predictions 7
2.4. Machine Learning Prediction Tools for Diabetes 7
2.4.1 Omni Diabetes Calculator 7
2.4.2 Framingham Diabetes Risk Score Model 7
2.5. Related Works 8
CHAPTER THREE 10
MATERIALS AND METHODS 10
3.1 Dataset Collection 10
3.1.1 Requirements Gathering 11
3.1.2 Data Analysis for the Requirements Gathered 11
3.1.3 Requirement Analysis 11
3.2 Data Preprocessing for Machine Learning 13
3.2.1 Data Integration 13
3.2.2 Data Cleaning 13
3.3 Features Selection 14
3.4 Development of the Web Application 14
3.4.1 Software Development 14
3.4.2 Tools and Technologies for Software Development 17
3.5 Model Training and Testing 19
3.5.1 Random Forest Algorithm 19
3.5.2 Steps for Random Forest Creation/Building 19
3.6 Assumptions and Dependencies 20
CHAPTER FOUR 22
RESULTS AND DISCUSSION 22
4.1 Results 22
4.1.1 The developed System 22
4.1.2 Usability of the Web Based Predictive Model for Diabetes 24

4.1.3 System Validation 25
4.1.4 User Acceptance 26
4.1.5 Model Evaluation 27
4.2 Discussion 28
CHAPTER FIVE 30
CONCLUSION AND RECOMMENDATIONS 30
5.1 Conclusion 30
5.2 Recommendations 30
REFERENCES 32
APPENDICES 35
POSTER PRESENTATION 39

LIST OF TABLES

Table 1: Functional Requirements 12

Table 2: Non-Functional Requirements 13

Table 3: Comparison of Prototyping Methodologies 17

Table 4: System Testing Results 25

Table 5: User Acceptance Testing Response Alternatives 26

Table 6: User Acceptance Testing Results 26

Table 7: Confusion Matrix and Accuracy for the Classifiers 27

Table 8: Classification Report for the Classifiers 28

LIST OF FIGURES

Figure 1: Development Life Cycle of the Proposed System..... 15

Figure 2: Form to be filled in to make a prediction..... 22

Figure 3: An alert of all fields to be filled in before clicking predict button..... 23

Figure 4: Result when you have low risk of having Type 2 diabetes 24

Figure 5: Results when you have high risk of having Type 2 diabetes..... 24

LIST OF APPENDICES

Appendix 1: Questionnaire 35

Appendix 2: The Budget for Predictive Model for Early Detection of Diabetes Mellitus using Machine Lear

Appendix 3: Work Break Structure for a Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning 37

LIST OF ABBREVIATIONS AND SYMBOLS

TMCG	The Medical Concierge Group
EMR	Electronic Medical Records
DM	Diabetes Mellitus
EMoS	Embedded and Mobile Systems
BMI	Body Mass Index
RBS	Random Blood Sugar
GDM	Gestational Diabetes Mellitus
NCDs	Noncommunicable Diseases
AI	Artificial Intelligence
ML	Machine Learning
CSV	Comma Separated Values
WHO	World Health Organization
PID	Pima Indian Diabetes Dataset

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Diabetes is one of the substantial general health issues, affecting over 425 million people worldwide, mostly adults over the age of 18. The word Diabetes consists of two Greek words "Dia" meaning through, and "betes" meaning pass, which refers to a cycle of heavy thirst and abnormal frequent urination. The word "Mellitus" meanwhile is a Latin word which means "sweetened with honey", referring to the existence of sugar in urine (Guariguataa, 2014). Diabetes Mellitus can either be Type 1 diabetes or Type 2 diabetes. Type 1 diabetes (also called insulin dependent diabetes or juvenile-onset diabetes) develops when there is destruction in insulin-producing cells of the pancreas, which may be brought about by an auto-immune reaction. It is most common in children and its signs and symptoms start slowly but later increase. They include bed-wetting in children who previously never wet the bed at night, unintended weight loss, increased thirst, extreme hunger and frequent urination.

Type 2 diabetes also proclaimed by a name of non-insulin dependent diabetes is commonly in adults starting at the age of 30, with onset often occurring between 50 and 60 years. It is brought about when the body manufactures or produces insulin but the cells do not make use of the produced insulin very well or do not make use of the produced insulin at all. In recent years, the occurrence of Type 2 diabetes in both young and adult populations have been increasing, with projections showing 1.5 times rise in cases by 2030 from the 285 million cases in 2010. Both types of diabetes are associated with a large number of complications including: Blindness, heart disease, strokes and impotence in males. Some pregnant women experience gestational diabetes, which is caused by glucose intolerance (Guariguataa, 2014).

The majority of diabetes patients reside in developing countries. In Africa, an estimate of 19.8 million people has diabetes in accordance with the International Diabetes Federation, with approximately 75% of them being undiagnosed. Increased late detection and prognosis of diabetes has increased the prevalence of the disease currently standing at 15% in adults. According to International Diabetes Federation (Close, 2018) out of the 19.8 million adult population, 259 100 were newly diagnosed diabetes incident cases. Diabetes prevalence is increasing on the continent as a result of increased overweight, obesity and additional cardiovascular risk factors such as high blood pressure (Faraja, 2015). Other risk factors include family history of diabetes, age, physical inactivity, unhealthy diets and impeached glucose

tolerance (David, 2011). It is on record that in Sub-Saharan Africa, the number of diabetes cases will rise by 69% in adults aged 20 to 29 years between 2010 and 2030 (David, 2013).

Uganda is a country in East Africa and has a population of over 47 million people (Worldometer, Uganda Population, 2021). Prevalence of diabetes is estimated to be around 1.4%, which is about 500 000 people. Around 21% of diabetics have Type 2 diabetes. The rural Eastern Uganda has the highest occurrence of diabetes and prediabetes, accounting for 7.4% and 8.6% respectively of national cases, with the mean age being 35 years (David, 2013). Diabetes is usually diagnosed through blood tests carried out on a person's blood glucose levels either using fasting glucose test or the random glucose test or the A1c test. In rural areas and low-income countries, diabetes diagnosis is still low due to little knowledge of symptoms for diabetes, and others are not diagnosed due to lack of testing kits (Pastakia, 2018). The disease has shown a big effect on the economy both direct and indirect in Uganda. The direct burden is in form of the health system cost incurred by the society while the indirect cost burden is in form of losses made in form time spent taking care of diabetic patients by their family members among other losses like money, pain to the family and loved ones in case of death (Joses, 2009). It is therefore, important to ensure early detection and management, as research shows early detection could allow reversal (Roy, 2019).

Machine-learning models allow automated identification mechanism for patients who are at high risk of contracting Type 2 diabetes when the models are trained with the risk factors of Type 2 diabetes (An-Dinh, 2019). Machine-learning models save time that would have been spent carrying out different blood tests like the fasting glucose test as described in chapter two.

1.2 Statement of the Problem

Late detection of diabetes can lead to severe complications such as blindness, impotence in male, kidney failure, cholesterol and heart diseases. There is also both a direct and indirect economic burden brought about by diabetes, particularly in low- and middle-income countries. Although various governments, including Uganda, have scaled up the purchasing and distribution of diabetes diagnostic kits, a majority of the population remains unknowledgeable about symptoms of the disease, which results in late seeking of care, and those in rural areas do not have enough access to diagnosis services.

Scientists around the world have successfully developed machine learning models for early prediction of diabetes, using various risk factors such as being born in a family with history of diabetes, eating unhealthy diets, being overweight and obesity, people do not engage in

physical exercises, cigarette smoking, heart diseases, and population of older people etc. as features. However, the African populations are severely underrepresented in these datasets, and therefore, the models may not be as accurate in predicting diabetes for them.

In Uganda, predictive models for diabetes are not so common and not well embraced. Therefore, the aim of this research was to develop a prediction model for the Ugandan population that could be deployed via a web application and be used across healthcare facilities in Uganda to help predict the risks of individuals acquiring diabetes at early stages.

1.3 Rationale of the Study

In order to curb down the increasing number of people with diabetes and getting new patients with diabetes, ease, early detection and management of diabetes is of a great importance towards managing people with diabetes and can be achieved with proper and early screening of people regardless whether they have diabetes or not. When a person is screened, his/her results turn out to be positive that is he/she has diabetes then he/she can start early diabetes medication. This will reduce on the serious complications brought by late detection of diabetes for example precipitate heart disease and stroke, blindness, limb amputations, and kidney failure. With the help of machine learning, early detection of diabetes mellitus can be made easily by evaluating the risk factors that cause diabetes mellitus.

1.4 Project Objectives

1.4.1 Main Objective

To develop a web-based predictive model for possible cases of Type 2 diabetes based on input symptoms and support early referral for testing.

1.4.2 Specific Objectives

(i) To collect data for the risk factors that cause diabetes mellitus to be used for building a predictive model for early detection of Diabetes Mellitus.

(ii) To determine the most important features for predicting possible case of Type 2 diabetes

(iii) To build a prediction model using classification methods.

(iv) To evaluate and validate the prediction model.

(v) To develop a web application based on the best algorithm.

1.5 Research Questions

(i) What kind of data is needed for predicting the patients' future likelihood of having diabetes mellitus?

(ii) What are the key important features used for predicting risk factors of Type 2 diabetes?

(iii) What classification algorithms perform best in predicting possible cases of Type 2 diabetes in Ugandan patients?

(iv) How can it be sure that the developed model works as was expected?

(v) What do Ugandan clinicians require in a web-based application for the deployment of a machine learning prediction model for Diabetes 2?

1.6 Significance of the Study

Due to the rising number of late detection and prognosis of diabetes, and with the advancement of technology in the medical sector like the use of Machine Learning and Artificial Intelligence (AI), early detection of diabetes can be made easily through developing and building a predictive model to detect diabetes at an early stage. Thus, making it easy for the patient or a person screened with diabetes since diabetes is still at an early stage for him/her to get treatment easily and prevent him/her from momentous difficulties such as heart disease and stroke, blindness, impotence, and kidney non-performance. With the help of machine learning, early detection of diabetes mellitus can be made easily by evaluating the risk factors that cause diabetes mellitus.

1.7 Delineation of the Study

This research puts only much focus on Type 2 diabetes excluding the other diabetes mellitus types that is Type 1 diabetes and gestational diabetes. Amongst the very many risk factors of Type 2 diabetes, only twelve were considered and assessed that is, furthermore, this research study never aimed at developing a mobile app but the developed web-based application can be accessed using a mobile phone and is mobile responsive as predefined by the developer, and also it does not allow users (Health Workers) to decide which data to analyse because only twelve risk factors were selected as listed below:

(i) Age of respondent

(ii) Gender (Male or Female)

(iii) Systolic blood pressure (mmHg)

(iv) Residence (Town or Village)

(v) Diastolic blood pressure (mmHg)

(vi) Family Member with Diabetes (0=No, 1=Yes)

(vii) Alcohol intake (0=No, 1=Yes)

(viii) Smoker (0=No, 1=Yes)

(ix) Hypertension (0=Normal, 1=Hypertension)

(x) Obesity - Known to have obesity (0=No, 1=Yes)

(xi) Achieves WHO recommended minimum physical activity level (0=No, 1=Yes)

(xii) Body Mass Index (kg/ m^2)

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction

For over decades, most recent research studies have focused on early detection of diabetes mellitus as for late detection of diabetes mellitus has for over time been a big challenge worldwide when it comes to managing and treating diabetes mellitus. Late detection of diabetes mellitus has over many complications to the human body including: blindness, damage of nerves, stroke, heart problem, impotence in men and many others. Late detection of diabetes mellitus being a challenge, ways on how to predict the high risk of developing diabetes or possible cases of Type 2 diabetes is mandatory to prevent development of diabetes.

Many researchers have come out with various solutions to curb down occurrence of the problem, amongst the solutions, is the use of artificial intelligence techniques for diabetes mellitus prediction such as use of deep learning and machine-learning algorithms in diabetes prediction. Computational intelligence or the artificial intelligence techniques for example deep learning and machine learning algorithms in predicting diabetes disease are known in the medical field and researchers are working on improving the accuracy of the different algorithms. Many research studies have been done trying out different classifiers and building new models with a motivation of improving the accuracy of diabetes prediction.

2.2. Machine Learning in Diabetes Predictions

2.2.1 Datasets

Diabetes mellitus has been around for many decades but there is a small number of publicly available datasets that have been collected and being used for diabetes prediction. However, the accuracy of the machine learning and deep learning algorithms depends on the size of the datasets (Ajiboye, 2015). The most common diabetes datasets used for predicting the disease include the Pima Indian Diabetes (PID) Dataset found in this link https://www.kaggle.com/uciml/pima-indians-diabetes-database which has 8 attributes with 768 records or instances, where 65.1 % (500 instances) are non-diabetic and 34.9 % (268 instances) are diabetic. The dataset has been used by many researchers like Han (2018) who used the PID dataset to develop Type 2 diabetes mellitus prediction model based on data mining and many other researchers.

2.3. Benefits of Machine Learning in Diabetes Predictions

According to Faruque (2019), machine learning algorithms provide efficient results without taking long period of time thus, making the screening of Type 2 diabetes easy and allowing many screening tests to be done in a short time. The fasting glucose test requires a person to fast for more than 8 hours before the test can be done. With machine learning however, results can be obtained within an average of 5 minutes time with the main cost being the model building time.

2.4. Machine Learning Prediction Tools for Diabetes

2.4.1 Omni Diabetes Calculator

The Omni Diabetes Calculator lets the user see the probability of them contracting diabetes. The system requires the user to first do basic laboratory tests (fasting glucose and High-Density Lipoprotein (HDL) cholesterol tests), check their blood pressure, and then input the results into the Omni Diabetes Calculator (Białek, 2020). Fill in the required appropriate boxes: age, weight, sex, height, the laboratory results of fasting glucose and HDL, family history, average systolic blood pressure and your ethic origin. The Omni Diabetes Calculator gives out results of whether a person has a high risk or a low risk of contracting diabetes in the near future say in the next 7.5 years. The Omni Calculator still depends on the fasting glucose and HDL test results without the laboratory test results it cannot calculate to tell whether a people are at a high risk or low risk of having diabetes hence, it's limitation since without the fasting glucose and HDL cholesterol test results it cannot calculate to tell whether someone is at a high risk or low risk of having diabetes.

2.4.2 Framingham Diabetes Risk Score Model

Framingham Diabetes Risk Score Model (FDRSM) is a model for assessing the probability of having Type 2 diabetes however, it does not give reliable fatal cardiovascular disease risk approximate of Type 2 diabetes according to Ruth (2007). The model was developed using 337 diabetic patients in the Framingham cohort which is a small number when developing diabetes prediction models. Machine learning models require large sums of data for training the algorithms in order to have a big accuracy in the prediction. Like the Omni diabetes Calculator, the Framingham Diabetes Risk Score Model also requires the laboratory tests for fasting glucose and HDL cholesterol tests for it to assess the probability of occurrence of Type 2 diabetes.

2.5. Related Works

Prediction of gestational diabetes mellitus by maternal factors and biomarkers from 11 to 13 weeks and above by Surabhi (2011) in University College Hospital, London, UK sought to develop a predictive model for gestational diabetes in maternal women of 11 to 13 weeks+. The research mainly focused on diabetes in women which is gestational diabetes mellitus, still not all women but only pregnant women of 11 to 13 weeks+, maternal age, racial origin, body mass index, and previous history of gestational diabetes were the assessed factors to prognosticate the future occurrence of gestational diabetes mellitus in women. The limitation of the research is that its concentration is only on diabetes in women (gestational diabetes mellitus) yet diabetes occurrence is more in men than in women and type 2 diabetes has the biggest percentage of all the diabetes cases (Amy, 2019).

The research study by Henry *et al.* (2009) had an aim to develop the two risk-scoring systems for prognosticating the occurrence of diabetes mellitus in U.S adults 45 to 64 years. The research focused on the age bracket of adults of 45 to 64 years of age. The research limitations were not considering the knowledge of parental diabetes or the family history of diabetes on the people the research was conducted on, also the research had a race limitation and the age bracket group since diabetes mellitus can even be found with the person of age below 45 or above 64 years of age.

The research study done in Africa by Eric (2021) at Kwame Nkrumah University of Science and Technology (KNUST) in Ghana looked at developing a predictive model and feature importance for early detection of Type 2 diabetes mellitus. This work focused on a precise prediction of Type 2 diabetes by comparing multiple machine learning techniques for modelling but it the research dataset size was very small. The dataset consisting of 219 Type 2 diabetic patients and 219 non-diabetic patients was used which is very small when it comes to training and testing machine learning algorithms. However, the research was done in an African context since the associated risk factors of diabetes from an African population were collected and used. The research was very good because the associated risk factors of diabetes from an African population were collected and used but only affected by the size of the dataset used (438 records).

The research study by Patrick (2021) sought to develop an Augmented Diabetes Prediction System in Mbarara region in Uganda. From his research survey carried out in January 2021, he sought to use the Naïve Bayes and Decision Tree algorithms to develop a system with an aim of evaluating the accuracy of the algorithms. However, the dataset was very small as it is to most

of the predictive models since few datasets are published out there for usage when developing machine learning diabetes predictive models. Furthermore, accuracy of the algorithms he used alone without other evaluation metrics would not be depended on because even there was no cross validation performed on the machine-learning algorithms. His research does not point out whether sensitivity or specificity of the used algorithms was a major concern, thus the research does not show that the Augment Diabetes Prediction System major aim was either detection of patients with or without diabetes.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Dataset Collection

Secondary data of patients who were screened for Type 2 diabetes were collected from Kampala (capital city of Uganda) that is from Mulago Kiruddu hospital and St. Apollo Health Center, from the Eastern Uganda cities of Jinja and Iganga (Iganga Hospital), and from The Medical Concierge Group (TMCG) clinic. The data contained 18 features as listed below:

(i) Age of respondent

(ii) Gender (Male or Female)

(iii) Systolic blood pressure (mmHg)

(iv) Residence (Town or Village)

(v) Diastolic blood pressure (mmHg)

(vi) Family Member with Diabetes (0=No, 1=Yes)

(vii) Alcohol intake (0=No, 1=Yes)

(viii) Smoker (0=No, 1=Yes)

(ix) Hypertension (0=Normal, 1=Hypertension)

(x) Obesity - Known to have obesity (0=No, 1=Yes)

(xi) Achieves WHO recommended minimum physical activity level (0=No, 1=Yes)

(xii) Body Mass Index (kg/ m^2)

(xiii) Occupation (1=Subsistence 2=Trading 3=Formal 4=Mechanic)

(xiv) Socio-economic status quintile (0=Lowest 1=Second 2=Middle 3=Fourth 4=Highest)

(xv) Education level (1=None 2=Primary 3=Secondary 4=Tertiary)

(xvi) Stress score in 3 categories (1=Low, 2=Moderate, 3=High)

(xvii) Known to have diabetes (0=No, 1=Yes)

(xviii) Knowledge about diabetes in 4 categories (1=Very Low, 2=Low, 3=Moderate, 4=Good)

A total of 2500 records were obtained, 20% diabetic cases and 80% non-diabetic cases. Altogether, data from 2 hospitals and 3 clinics was collected for this web based predictive model for early detection of diabetes on over 2500 patients where by 1228 (49.12%) were males and 1272 (50.88%) were females.

3.1.1 Requirements Gathering

Data collection was carried out for a timeframe of three months that is, November 2020 to January 2021. Interviews were carried out with 8 key informants (5 doctors and 3 clinical laboratory technicians). Simple random sampling, which is a probability sampling method, was used in choosing the 8 key informants and to generalize the risk factors of Type 2 diabetes or possible cases of Type 2 diabetes. The interview questions were in a structured form and the ultimate purpose was to identify the main risk factors of Type 2 diabetes that they relied on.

3.1.2 Data Analysis for the Requirements Gathered

Data analysis was done using the descriptive measures that gave a descriptive coefficient which summarized the data collected. A fully filled data file was sent into electronic form in Google drive excel sheet. The excel sheet for the data that was in a CSV format was extracted and uploaded into a Jupyter notebook which is part of anaconda and used for data cleaning and analyzing data with python programming language.

3.1.3 Requirement Analysis

User requirements were gathered, examined and then grouped into categories that is functional requirements and non-functional requirements.

(i) Functional Requirements

Functional requirements define the basic system behaviour, in principle functional requirements demonstrate the ability of the system (Gabriela, 2017). Tells what input the system should take in, what kind of information should the system give as output, what type of computations and processing should the system do, as well as the timing and synchronization of the data and processes mentioned above. The Table 1 shows the functional requirements of predictive model for early detection of Type 2 diabetes.

Table 1: Functional Requirements

Subsystem	Requirement	Description
System Form	All the fields of the form must be filled in.	The system should provide a person a slot to the age. The system should give a chance to a person to select the gender, Male or Female. The system should give a chance to a person to select the location, Village or Town. The system should provide a person a slot to enter the body mass index (BMI) value. The system should give a chance to a person to choose whether the person drinks alcohol or not. The system should give a chance to someone to choose whether the person smokes or not. The system should provide a person a slot to enter the systolic blood pressure readings. The system should give a chance to a person choose whether the person has a family member with diabetes or not. The system should give a chance to a person to choose whether they are hypertensive or not. The system should provide a person a slot to enter diastolic blood pressure readings. The system should give a chance to a person to choose whether they are obese or not. The system should give a chance to a person to choose whether they are physically inactive or not.
Prediction	Predict Button	The system predict button should be clicked after filling in all the fields of the system form.
Results form	Click the Predict Button	After filling in the system form and clicking system predict button, the predicted results are displayed based on the inputs, either having low risk or high risk of having Type 2

Subsystem	Requirement	Description
		diabetes.

(ii) Non-Functional Requirements

Nonfunctional requirements of a system majorly define the constraints that the developers of a system ought to adhere to during the design and implementation (development) of that particular system (Shahid, 2017). In this study, they describe of how the predictive model should make its prediction of possible cases of Type 2 diabetes, they are elucidated in Table 2.

Table 2: Non-Functional Requirements

Requirement	Description
Performance	The system shall work on user inputs in the shortest possible time, availing up-to-date predictive responses.
Usability	The system shall be easy to use and understand by users since it follows conventional standards in computing.
Robustness	The system has the ability to recover from failure in case of problems with connection of either hardware or software
Availability	The system shall be available to the users whenever needed as long as they are connected to the internet.
Language	The system shall be availed and documented in only English language

3.2 Data Preprocessing for Machine Learning

The data preprocessing began with importing all the crucial libraries which included the Pandas library, NumPy library, Pickle library, Seaborn library, and many other libraries. Data preprocessing was divided into two stages that are:

3.2.1 Data Integration

Data integration, which involved consolidating data from the various sources, and was done by using the uniform access integration technique.

3.2.2 Data Cleaning

Data cleaning, which involved identifying and handling missing values, fixing structural errors, filtering unwanted outliers and removing duplicates in the dataset. Missing values were handled

by using the Pandas method of forward fill. The ultimate goal was to improve on the data quality and increase the overall productivity.

3.3 Features Selection

From the total of 18 features, the following were chosen as being most relevant for predicting Type 2 diabetes. The selection was done based on the domain expertise of the doctors, and amongst the very main risk factors of Type 2 diabetes used in related works:

(i) Age of respondent

(ii) Gender (Male or Female)

(iii) Systolic blood pressure (mmHg)

(iv) Residence (Town or Village)

(v) Diastolic blood pressure (mmHg)

(vi) Family Member with Diabetes (0=No, 1=Yes)

(vii) Alcohol intake (0=No, 1=Yes)

(viii) Smoker (0=No, 1=Yes)

(ix) Hypertension (0=Normal, 1=Hypertension)

(x) Obesity - Known to have obesity (0=No, 1=Yes)

(xi) Achieves WHO recommended minimum physical activity level (0=No, 1=Yes)

(xii) Body Mass Index (kg/ m^2)

3.4 Development of the Web Application

3.4.1 Software Development

The Evolutionary prototyping methodology was chosen for development of the web application due to an application being novel and therefore, end-users did not fully know their requirements, and required multiple chances to provide feedback and revised the system based on their usage experiences and secondly the system was complex and required testing and validation at multiple times. In addition, evolutionary prototyping also starts implementation

when the requirements are properly understood, unlike other prototyping methodologies like rapid prototyping (Hamoodi, 2014), and has an advantage in that the building and development of the system is speeded up, and user engagement with the system during the development of the system is more likely to meet user requirements.

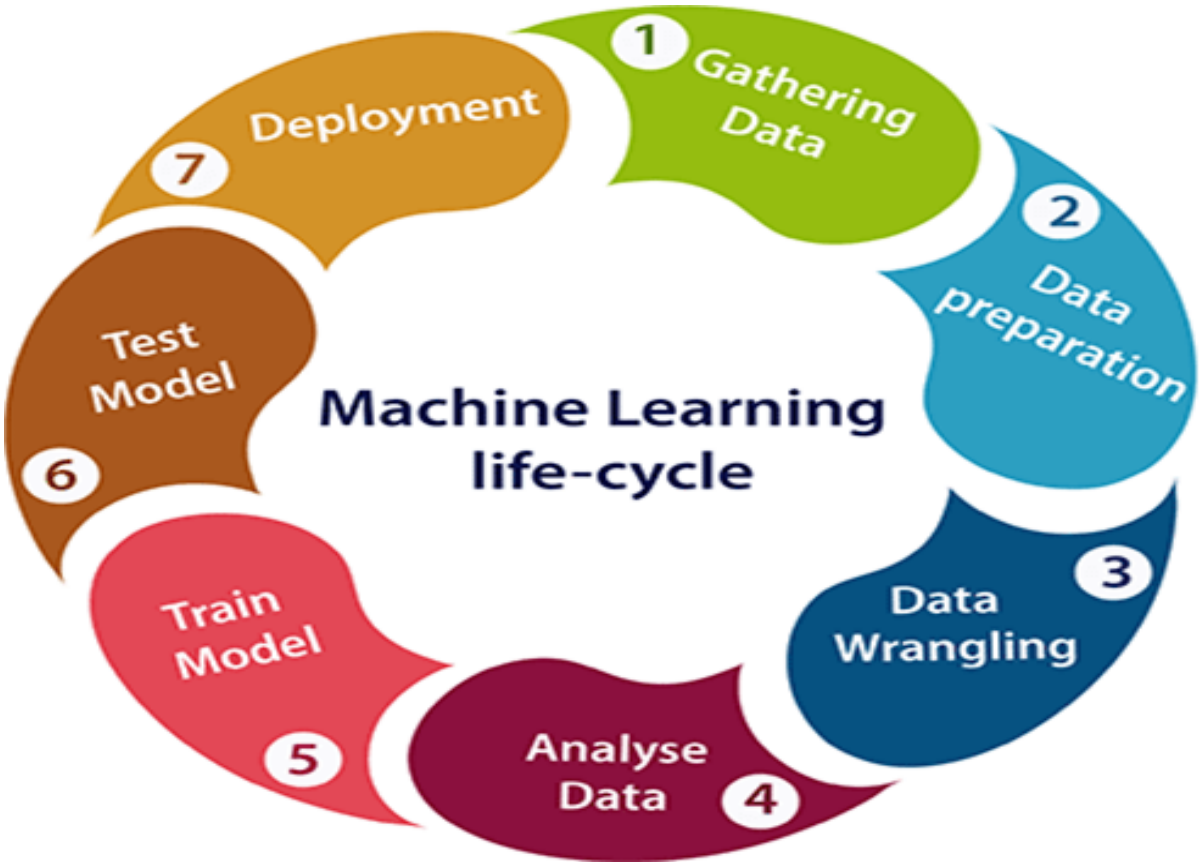


Figure 1: Development Life Cycle of the Proposed System

The screening tool for predicting early detection of Type 2 diabetes went through seven steps as the development life cycle of the proposed system.

(i) Gathering Data

Secondary data of patients who were screened for Type 2 diabetes were collected from different hospitals and health centres and the gathered data contained 18 features.

(ii) Data Preparation

Data preparation also known as data pre-processing, it is the process of reshaping or metamorphosing raw data in order that data scientists and analysts can run it through machine learning algorithms to unveil insights or make prognostications. Jupyter notebook is the web application which was used during data preprocessing, data wrangling, data analysis, model training and model testing.

(iii) Data Wrangling

Data wrangling is also known as data munging, it is the process of polishing and integrating begrimed and complex data sets for easy access and analysis. Using Jupyter notebook as the web application, data set was cleaned, shuffled to avoid biasness in when picking features from the data set, and the missing values in the data set were handled using the Pandas method of forward fill.

(iv) Data Analysis

Predictive analytics was carried out, it encompassed the use of data, statistical algorithms and machine learning techniques to find out the probability of future outcomes of Type 2 diabetes based on historical data.

(v) Model Training

Training the algorithm or Model training, 70% of the data that was 1750 records from the data set were used for training the algorithm. And the Ten-Fold cross validation was employed during the training of the algorithm.

(vi) Model Testing

Testing the algorithm or Model testing, 30% of the data was used that is 750 records from the data set were used for testing the algorithm.

(vii) Deployment

Heroku cloud hosting platform was used in the deployment of the developed predictive model for predicting early detection of Type 2 diabetes. Table 3 compares Waterfall and Spiral Methodologies versus Evolutionary Prototyping.

Table 3: Comparison of Prototyping Methodologies

Methodology	Flexibility	Client Interaction	Phase Containment of Error
1. Waterfall	No	One Time	Low
2. Evolutionary Prototyping	Fully	Frequent	High
3. Spiral	Few	Some time	Low

3.4.2 Tools and Technologies for Software Development

For client-side scripting languages, Hypertext Markup Language (HTML), Cascading Style Sheet (CSS) and JavaScript were used for the development of the system, and for the back-end Python was the main language that was used when developing the system. Programming codes for the client-side language are executed within the browser at the client side because the code execution process is at the front end of the user' computer and makes rendering the page easy.

(i) Hypertext Markup Language

The HTML is Hypertext Markup Language, and HTML scripting language is a static client-side scripting Language basically used for formatting, setting forth of data to the web browser. The HTML as a scripting language is the widely used scripting language for developing websites and uses Tags, HTML has both opening and closing tags. The commonly used HTML tags include; Paragraph tags (My Paragraph), Header tags (h1 to h6) take an example of (<h4>Heading</h4>), anchor tag for linking different web pages (Home Page), image tag for inserting in image in a web page (<img src="">) and other many tags. The HTML was used in developing the front end or user interface of the web-based Predictive model because HMTL is simple to learn and allow incorporating in other programming languages like CSS and JavaScript among others.

(ii) Cascading Style Sheet

Cascading Style Sheet (CSS) was concerned with directing how the HTML appeared on the web page of the system. The main purpose of CSS was in formatting the color, font-size, font

of the web page of the system. In other words, CSS was used as makeup of the pages of the developed system, CSS gave the background look and styled the appearance of the system in terms of background images, colors, headers, footers among other things related to the appearance of the developed system.

(iii) Python

Python was used as a general-purpose programming language and was the main programming language used during the development of the system. Python was in both font and back end of the system unlike the HTML, CSS and JavaScript which were only used for the font end development. A python web framework called flask was used in developing the system, python flask helped in providing useful tools and features for creating web-based applications.

(iv) Jupyter Notebook

The Jupyter Notebook is a free, an open-source, interactive web application that permits researchers to generate, to combine software code and distribute documents that merge live code, equations, computational output, conceptualization, and other multimedia resources. The name, Jupyter, is derived from the main supported programming languages that the platform supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows programmers to keep in touch codes programs in Python. Jupyter notebook was used during data preprocessing, data wrangling, data analysis, model training and model testing.

(v) PyCharm

The PyCharm is a dedicated python Integrated Development Environment (IDE) used in computer programming. PyCharm was used due to its ability to have a connection with a version control or source control like GitHub, and was used for sharing the project codes with GitHub.

(vi) GitHub

The GitHub is a web-based code hosting platform for source control and working as a team through partnership, GitHub offers the distributed source control and source code management functionality of Git, plus its own feature. GitHub also provides access to source control and security to the programming codes. The GitHub was used due to its ability to have a connection with a hosting platform like Heroku.

(vii) Heroku Cloud Hosting Platform

Heroku is as a service supporting several programming languages as a Platform-as-a-Service (PaaS) that enables developers to set up, assemble, run, and operate applications entirely in the cloud. The Heroku was used in the deployment of the screening tool for predicting early detection of Type 2 diabetes. Heroku also has SQLAlchemy as a database for storing the application data, SQLAlchemy is the Python SQL toolkit and Object Relational Mapper that gives application developers the full ability and flexibility of SQL.

3.5 Model Training and Testing

A binary classification model to predict the risk/possible cases of Type 2 diabetes (high risk and low risk output classes) was developed using the random forest classifier and binary variable in the form of high risk and low risk of having diabetes mellitus, two main things were done that was,

(i) Estimating the parameters for the machine learning methods, in machine learning lingo, estimating the parameters is called "**training the algorithm**", 70% of the data that was 1750 records were used for training the algorithm. And the Ten-Fold cross validation was employed during the training of the algorithm.

(ii) Evaluating how well the machine learning methods worked, in machine learning lingo, evaluating the method is called "**testing the algorithm**", 30% of the data was used that is 750 records were used for testing the algorithm.

3.5.1 Random Forest Algorithm

In the first place, random forest algorithm can be used for both classification and regression tasks. The random forest classifier as known as random decision forest is ensemble learning method for classification, regression and other tasks that operates by setting up a numerous of decision trees at training time. In coming up with the model for predicting possible cases of Type 2 diabetes, a binary classification model was used, the random forests were used to build numerous decision trees and merged them in conjunction to get a more precise and stable prediction.

3.5.2 Steps for Random Forest Creation/Building

Random forests are built or created from decision trees, decision trees are simple to set up, simple to use, and uncomplicate to elucidate but in operation they are not that amazing. To restate from the elements of statistical learning also known as the holy writ of machine

learning, decision trees have one characteristic that prevent them from being the ideal tool for predictive learning that is inexactitude or imprecision. Decision trees work great with data used to generate them but they are not pliable when it comes to classifying new samples. The good news is random forest combines the intelligibility of decision trees with flexibility resulting into a vast improvement in precision:

(i) Randomly Select some Features from the Original Dataset: Create a Bootstrapped Dataset. To create a bootstrapped dataset, randomly pick samples from the original dataset. The important detail is that it's authorized to use the same sample more than once. Bootstrapping the data in addition to using the aggregate to make a decision is called bagging. When creating a bootstrapped dataset some samples are left out from the original dataset, in other they are not picked. Those are called Out-Of-Bag dataset.

(ii) Among the Selected Features, Calculate the Node using the Best Dissociate Point: Create Decision Tree using the Bootstrapped Dataset, but only pick the Random Subset of Variables (or Columns) at each Step. Using a bootstrapped sample and considering only a subset of the variables at each step results into a wide variety of trees thus making random forests more effective than individual decision trees.

(iii) Dissociate the Node into Daughter Nodes using the Best Split

(iv) Perform the Steps from (i) to (iii) Continuously until a Certain Number of Nodes has been Reached

(v) Build Forest by Continuously Performing Steps from (i) to (iv) Respectively for a Number Times to Create the Desired Number of Trees

With the above five steps done (i) to (v), then the random forest classifier is created, thus the prediction has to be done.

3.6 Assumptions and Dependencies

(i) It is presumed that the system users (doctors and health or medical workers) have access to a smartphone, personal computers and Internet to use the developed web-based predictive model for early detection of Type 2 diabetes.

(ii) The primary users of the system which are the doctors and health or medical workers shall have basic knowledge about connecting computing devices.

(iii) The primary users of the system which are the doctors and health or medical workers shall have basic knowledge about Type 2 diabetes.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results

4.1.1 The developed System

The developed system was hosted at Heroku.com, it just requires one to be connected to the internet and open the preferred browser of choice, then type in this URL https://diabetes-mellitus-predictor.herokuapp.com/ . The system will display a page that has to be filled in all the fields of that page as shown in Figure 2.

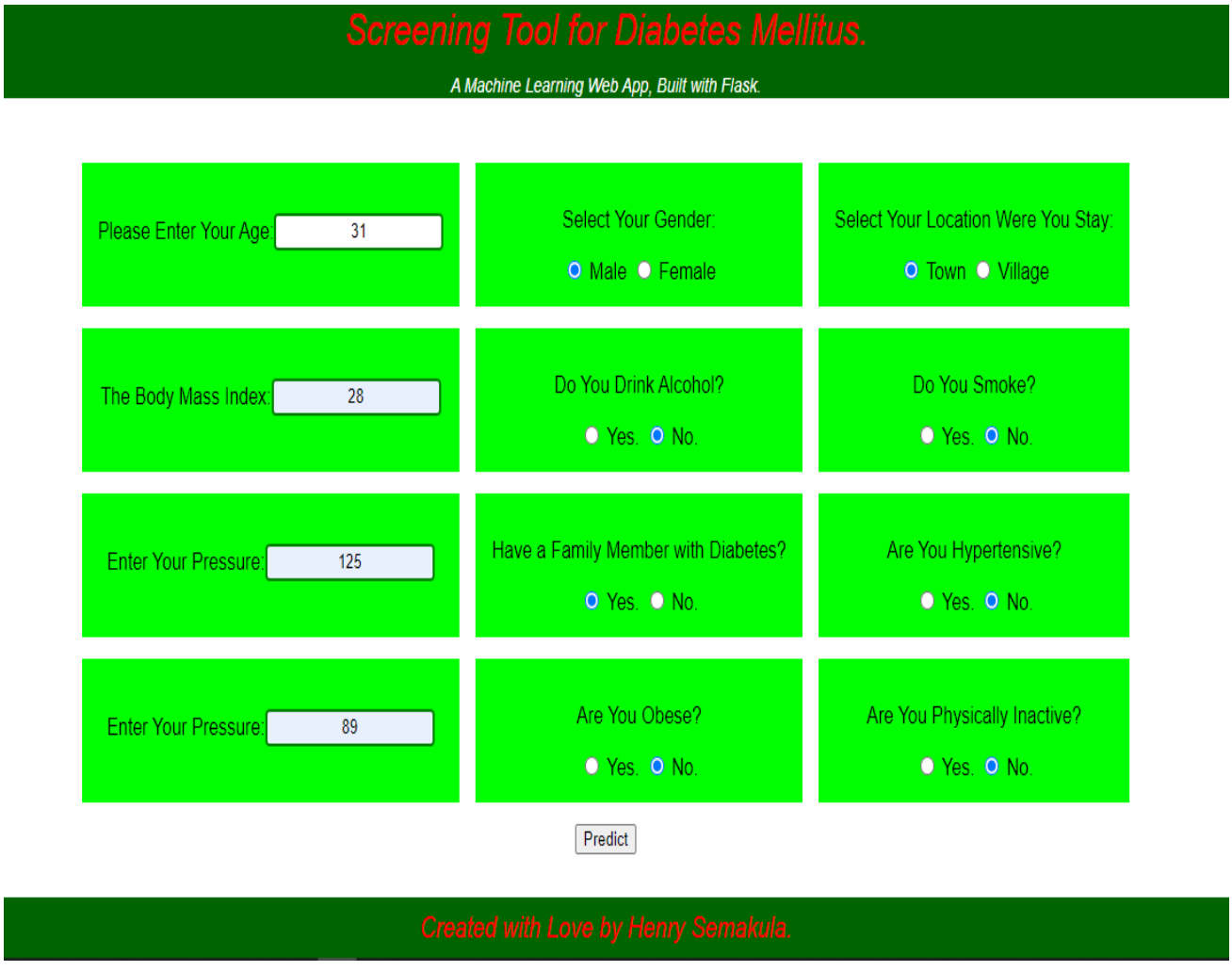


Figure 2: Form to be filled in to make a prediction

If any of the fields is not filled in or one of the fields is not filled in when the predict button is clicked, a prediction cannot be made. The system will prompt to fill in all the fields before one click the predict button or before a prediction is made as shown in the Fig. 3.

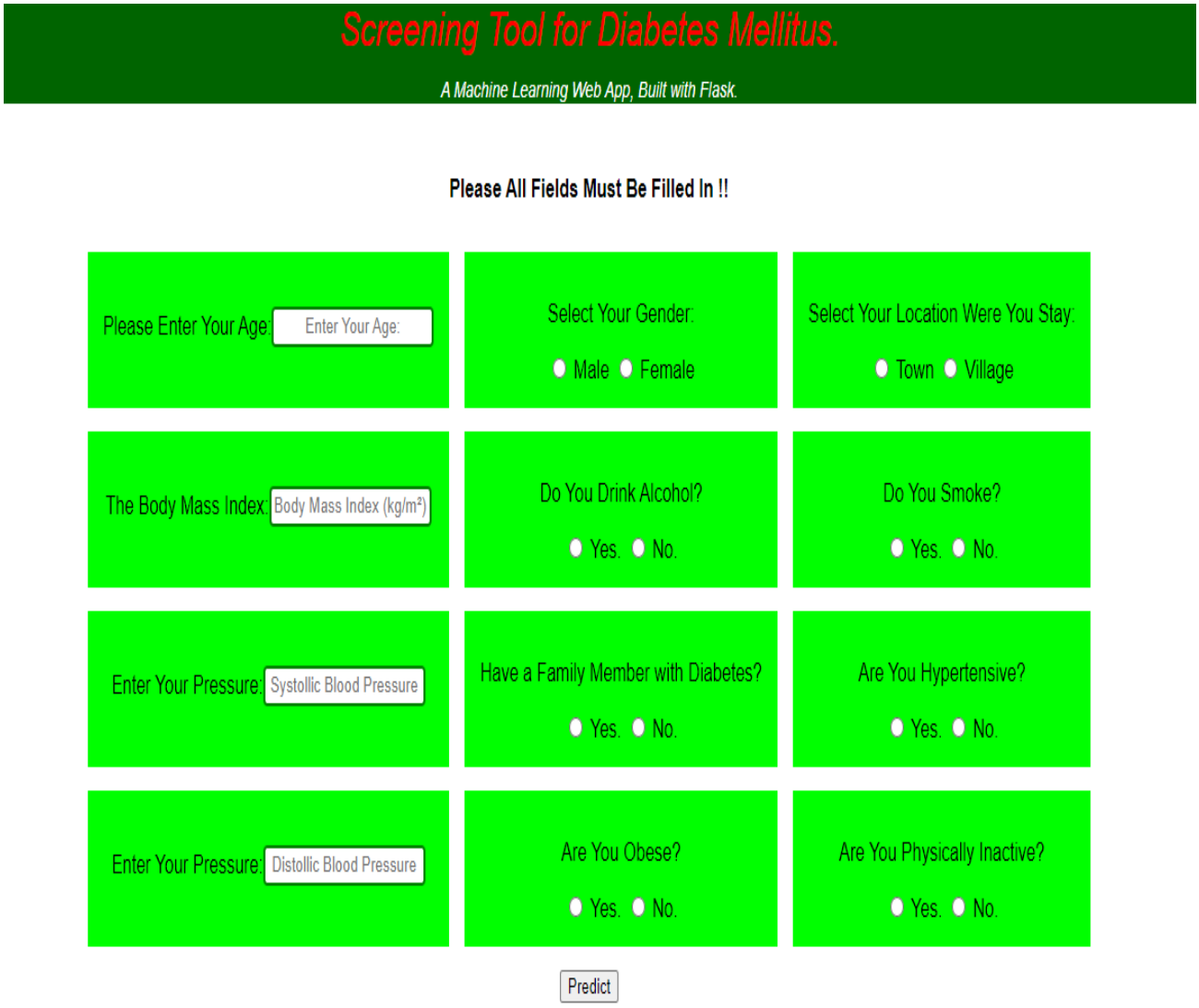


Figure 3: An alert of all fields to be filled in before clicking predict button

After filling in all the fields, then just click the predict button to get predicted results. The expect result page to be displayed with results of either having low or high risk of having Type 2 diabetes as shown in Fig. 4 and Fig. 5 respectively. When the risk of having Type 2 diabetes is low, the results appear as shown in Fig. 4.

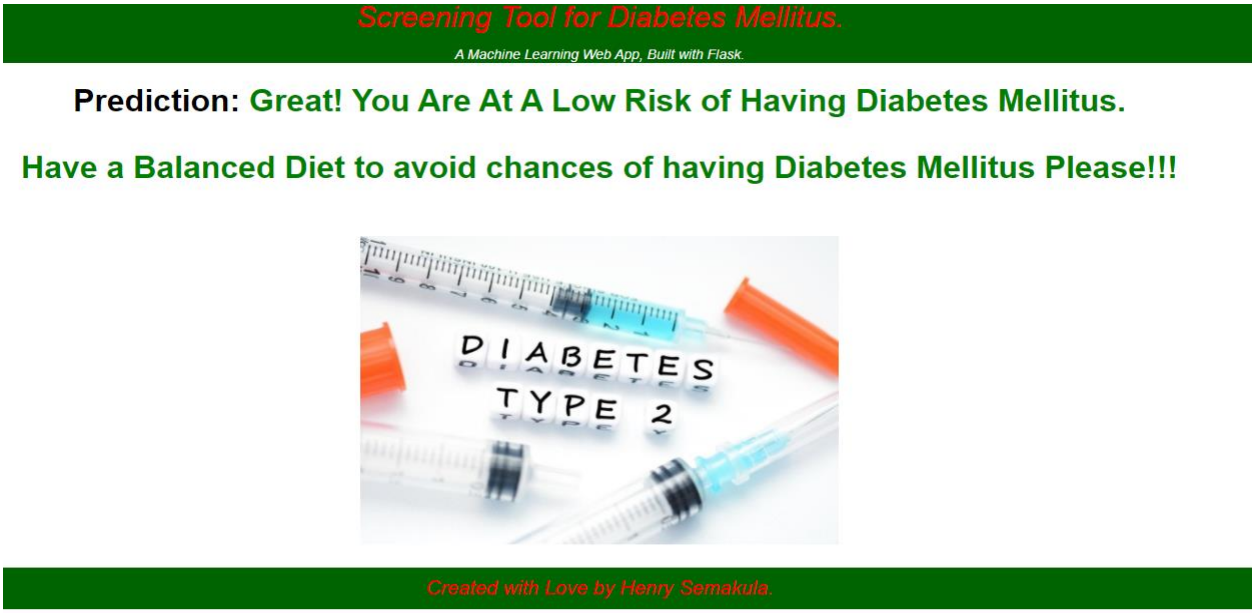


Figure 4: Results when there is a chance of low risk of having Type 2 diabetes

When the risk of having Type 2 diabetes is high, the results appear as shown in Fig. 5.

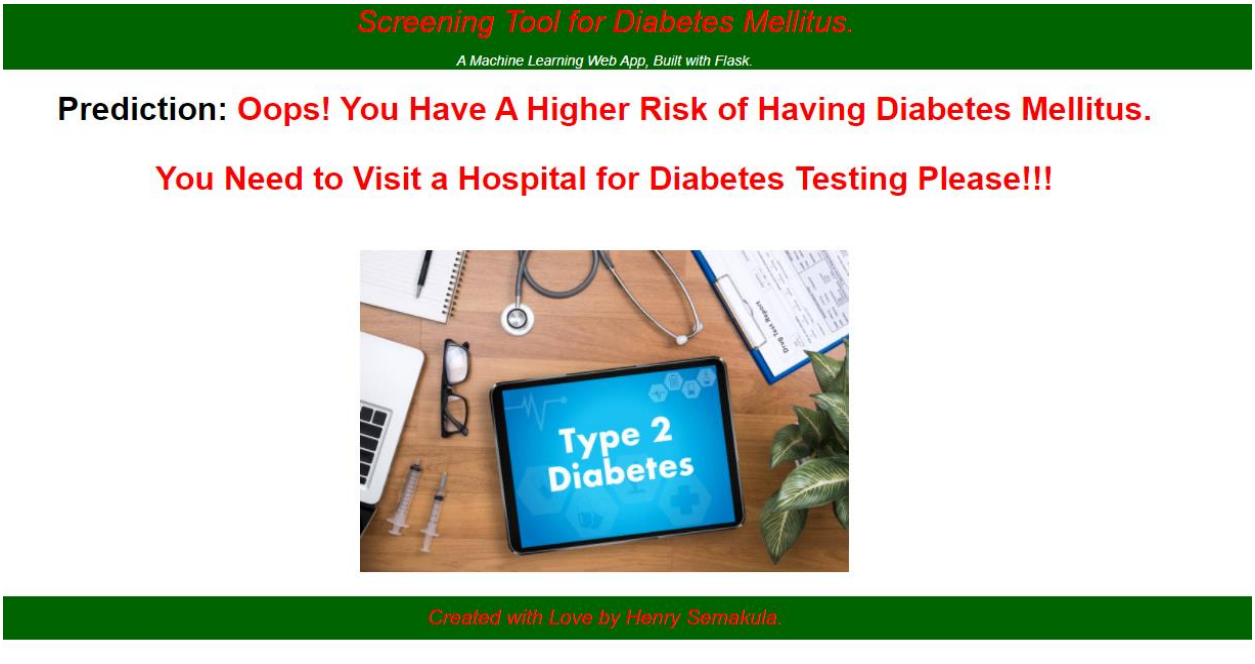


Figure 5: Results when there is a chance of high risk of having Type 2 diabetes

4.1.2 Usability of the Web Based Predictive Model for Diabetes

The Medical Workers and the other people that used the web based Predictive Model for Early Detection of diabetes mellitus using machine learning found the system so user friendly and simple to use. Since it only required the user to input or answer the questions that were in a

form format of the system, and then just within one click, that is clicking the Predict button, the system does the rest of the analysis and made a screening prediction. The prediction or the result of the diabetes screening is either the screened person has a high chance or risk of having Type 2 diabetes or has a low chance of having Type 2 diabetes as simple as that, as shown in Fig. 4 and Fig. 5 respectively.

4.1.3 System Validation

(i) Unit Testing

The unit testing was meant for substantiating the functional behaviour of the web-based predictive model. For this system, units which were tested were as follows: the function to make prediction, function to capture all input fields, function for querying the machine learning model, just to mention a few.

(ii) System Testing

After the unit testing and model integration to the web-based application, the system testing was done with the ultimate aim of verifying the developed web-based application whether it met specified business requirements of making diabetes predictions. Both local and deployed versions of the system passed the different tests of the system test.

Table 4: System Testing Results

Requirement	Description	Test Score
Home Page Loading	Web-based application shall load the home page and allow the users to input risk factors of diabetes.	Pass
Check-up Input fields of the form before making a prediction	The web-based application shall verify that all input fields of the form are filled in before it makes a prediction. If not it prompt the user to fill in all the fields then it can make a prediction.	Pass
Input Check-up	The input should be numeric, if not the web-based application can not make a prediction	Pass
Prediction Check-up	The web-based application shall make a prediction after all the input fields are filled in and are numeric, if not no prediction.	

4.1.4 User Acceptance

The user acceptance test was done by the potential users or doctors and medical workers to see how they can rate the developed software system. The ultimate intention of the user acceptance testing was to assess if the software produced is working properly or giving the right predictions as expected by the doctors and medical workers. The survey form had an alternative to give feedback on a lamina of 1 to 5 where 1 was for strongly disagreeing and 5 was for strongly agreeing. Table 5 shows the different alternative to give as answers to the questions and their meanings.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
5	4	3	2	1

Table 5: User Acceptance Testing Response Alternatives

Table 6 shows summary of the results we got from User Acceptance Testing tests of the web-based application.

Table 6: User Acceptance Testing Results

Validation Feature	Respondents					Mean Score
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	
The Web-based application is easy to use	1	0	4	3	12	4.05
The web-based application will impress the doctors & Medical workers and other frontline health staff to use it.	1	0	4	7	8	4.05
The web-based application will allow the doctors & Medical workers to access captured patient information.	1	0	3	5	11	4.25

The web-based Predictive Model for Diabetes was totally a new experience to the Medical Workers in the hospitals and the clinics where it was used for testing, the system was efficient

and less time consuming. The system was able to carry out diabetes screening for many patients taking a short time to give out results and it was efficient. The Medical Workers were taking a lot of time to do screening for diabetes since many tests were carried out with the previous methods like testing the person's glucose level in a blood test, carrying out a random glucose test or A1c test. With the web-based model the person can be screened for diabetes without being inconvenienced since screening is done without having the person to first fast for 8 hours like the fasting glucose test. The previous methods of screening for diabetes weren't only time consuming but also expensive that why many people don't get for screening for diabetes.

4.1.5 Model Evaluation

The model evaluation was done using the evaluation metrics:

(i) Confusion Matrix

Confusion Matrix summarizes the performance of the categorization algorithms on test data. It gives the percentage accuracy of predicting true positive (TP), true negative (TN), false positives (FP) and false negatives (FN) instances. The metric was used for disease prediction, it is important to have high accuracy for TP and low number of FN, to ensure those at risk are identified early for diagnostic screening.

Table 7: Confusion Matrix and Accuracy for the Classifiers

Classifier	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)	Accuracy (100 %)
Random Forest	140	599	7	4	85.4
Naïve Bayes	51	551	64	84	80.2
AdaBoost	69	563	43	75	84.2
Support Vector Machine	4	601	5	140	80.7

(ii) Accuracy Score

The accuracy Score which gives the percentage of correct predictions made by the Model.

Table 8: Classification Report for the Classifiers

		Precision	Recall	F1-Score	Support
Random	0	0.50	0.65	0.54	108
Forest	1	0.95	0.90	0.96	642
Classifier	Accuracy			0.85	750
	Macro avg	0.73	0.77	0.74	750
	Weighted avg	0.89	0.86	0.90	750
Naïve Bayes	0	0.38	0.44	0.41	115
Classifier	1	0.90	0.87	0.88	635
	Accuracy			0.80	750
	Macro avg	0.64	0.66	0.64	750
	Weighted avg	0.82	0.80	0.81	750
Support	0	0.03	0.44	0.05	9
Vector	1	0.99	0.81	0.89	741
Machine	Accuracy			0.81	750
Classifier	Macro avg	0.51	0.63	0.47	750
	Weighted avg	0.98	0.81	0.88	750
AdaBoost	0	0.48	0.62	0.54	112
Classifier	1	0.93	0.88	0.91	638
	Accuracy			0.84	750
	Macro avg	0.70	0.75	0.72	750
	Weighted avg	0.86	0.84	0.88	750

4.2 Discussion

In this research study, the findings have shown that the majority of the stakeholders in the health sector and the general public at large are interested in the web-based Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning. It showed that many people and where the majority of the respondents believed the web application would reduce or solve the problem of late detection of Type 2 diabetes as the identified problem in this research problem statement. Many of the users, those who accessed and used the web-based application agree that the application is very easy to use and navigate. The respondents showed their confidence in the web-based application for early detection of Type 2 diabetes as a second alternative to the blood glucose level testing to detect diabetes.

Compared to the previous related works or developed predictive models for early detection of Type 2 diabetes, this web based predictive model is better in a way that it was built and developed using the dataset which was in an African or Ugandan context since data was collected from patients in Uganda. Yet most of the developed predictive models for Type 2 diabetes are developed using a dataset of European context, thus making this web based

predictive model for early detection of type diabetes better than the previously developed models.

The Ministry of Health (MOH) of Uganda and other health support bodies have not been using a similar web-based application for diabetes detection, but according to the responses from the public believes in the ideas thus, gives an opportunity to run it as a prototype in the existing health systems.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

According to the chapter four, the research results and discussion were presented, furthermore, the analysis and results for this research study were described and the development of a web-based Predictive Model for Early Detection of Diabetes mellitus using Machine Learning was done.

In the previous chapter additionally, a medical worker or doctor uses the internet and any browser to get access to the developed web-based system by typing in the URL (https://diabetes-mellitus-predictor.herokuapp.com/) of the system. When the system loads, the medical worker or doctor will have to fill in the risk factors of diabetes mellitus as they appear on the form of the system and then click the predict button so as to get results or do a diabetes screening of a patent. Results or screening results will appear after clicking the predict button on the system, as long as all the fields on the form of the system are filled in since all risk factors are important in making diabetes prediction or screening of diabetes so as to have accuracy results or prediction. If one or any of the fields in the system form is not filled in, even if you click the predict button the system will not make a prediction but it will advise you to fill in all the fields before clicking the predict button or before the system makes a prediction.

With the challenges of screening of diabetes in Uganda and Africa at large, screening has effects on individual people on whom the screening of diabetes has been carried out on, also effects on the health systems and as well as effects on society at large that may include; Policies, guidelines and practices for screening for Diabetes Mellitus have been found out to have implications on individuals that are screened for diabetes, health systems and society as a whole as pointed out in Chapter Two. However, with the web-based screening using the developed a web based Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning, challenges such as time and high cost of screening are addressed since the web-based screening is less time consuming and not expensive.

5.2 Recommendations

There is less attention given to prediction of diabetes in women, and most of the research studies 65% done on diabetes prediction usually use Pima Indian Diabetes (PID) dataset for diabetes prediction which is extremely a small dataset. This research focused on specifically

Type 2 diabetes and left out other types of diabetes but however, prediction was for both men and women and never used the Pima Indian Diabetes dataset instead data was collected and formed a dataset that was used. Therefore, I recommend for future research about diabetes prediction new datasets and big datasets size should be used so as to have big accuracy in the prediction, focus should also be put to the other types of diabetes mellitus say Type 1 diabetes, the gestational diabetes in women in order to address the less attention given to prediction of diabetes in women and the frequent use of the Pima Indian Diabetes dataset.

There is also less embracement of technology usage in healthcare or machine learning prediction models in healthcare more especially in Africa, so African governments must invest in sensitization on the usage of machine learning in healthcare so as researchers and scientists are more encouraged in developing more predictive machine learning models to be used in healthcare. The medical workers, nurses, doctors and every person or individuals especially those that are above 30 years of age in Uganda and Africa at large should put much focus on the results of this research and start using the developed web based Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning so as to find out whether they stand a high chance of getting diabetes mellitus or not. If they have a high chance of getting diabetes mellitus, they should start watching or monitoring their diet, do enough physical exercises or visit the hospital for proper advice in order to avoid getting diabetes mellitus.

The African governments and their policy makers, other stakeholders such as the WHO should Promote the use of web-based detection of Diabetes Mellitus and encourage people to use the developed web-based Prediction Model to have Early Detection of Diabetes Mellitus using Machine Learning. https://diabetes-mellitus-predictor.herokuapp.com/. Since with early detection of a high risk of getting disease such as diabetes mellitus, management or treatment can easily be provided to minimize the long-term complication of Diabetes Mellitus as discussed earlier in chapter two. Organizations and companies dealing with integrating ICT and Machine Learning into Medical Care and Health care should encourage and do promotion for this web-based detection of Diabetes Mellitus as well as sponsoring it.

Much emphasis now should be put on Mobile Application Development for the Early Detection of diabetes Mellitus as a recommendation for this research so as to improve on easy and early detection of diabetes and also consider improving approaches and methods that are both qualitative and quantitative in nature so as to have a real and perfect description of the study and its findings.

REFERENCES

Aidehi, V. V. & Mujumdar, A. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165, 292-299.

Ajiboye, A. R., Abdullah-Arshah, R., Qin, H., & Isah-Kebbe, H. (2015). Evaluating the Effect of Dataset Size on Predictive Model using. *International Journal of Software Engineering & Computer Sciences, 1*, 75-84.

Amy, G., & Huebschmann, R. R. (2019, August 27). Sex differences in the burden of Type 2 diabetes and cardiovascular risk across the life course. *Springer Link*, 62, 1761–1772.

An-Dinh, S. M. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *Medical Informatics and Decision Making, 19*(1), 20-46.

Białek, Ł. (2020). *Diabetes Risk Calculator*. https://www.omnicalculator.com/health/risk-dm

Close, A. M. (2018). International Diabetes Federation 2017. https://www.google.com

Eliana, M. W., Maria, R. T., Maicon, F., Janet, T., Maria, A. D., Maria, A. C., Bruce, B. D. & Maria, I. S. (2012). Gestational diabetes and pregnancy outcomes. *Pregnancy and Childbirth,* 12(23), 22-30.

Eric, A., Emmanuel, A., Kolog, E., Afrifa, Y., Bright, A., Christian, O., Enoch, O. A., Emmanuel, A., Wei, W., & Antonia, Y. T. (2021). Predictive Model and Feature Importance for Early Detection of Type II Diabetes Mellitus. *Research Gate*, 6(17) 21-43.

Faraja, S., & Chiwanga, M. A. (2015). Diabetic foot: Prevalence, knowledge, and foot self-care practices among diabetic patients in, Tanzania. *Journal of Foot and Ankle Research*, 8(20), 1-7.

Faruque, M. F., & Asaduzzaman, S. I. H. (2019). *Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. 2019 International Conference on Electrical, Computer and Communication Engineering* . Cox'sBazar, Bangladesh. https:// www. google. com.

Guariguataa L. D. (2014). *IDF Atlas*. https:// diabetesatlas. org/ upload/ resources/ previous/ files/ 7/ IDF%20Diabetes%20Atlas%207th.pdf

Hamoodi, S. A. (2014). Website Development Life Cycle. *International Journal of Computer and Information Technology, 03*(05), 22-79.

Han-Wu, S. Y. (2018). *Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 100-107*. https://www.google.com

Henry, S., Kahn, M., Yiling, J., Cheng, M. P., Theodore, J., Thompson, M., Giuseppina, I. M. P., & Edward, W. G. P. (2009). Two Risk-Scoring Systems for Predicting Incident Diabetes Mellitus in U.S. Adults Age 45 to 64 Years. *Annals of Internal Medicine, 150*(11), 752-765.

Joses, M. K., Hama. B. S., Luis, G. S., & Saidou, P. B. (2009). Economic burden of diabetes mellitus in the WHO African region. *International Health and Human Rights*, 9(6), 19-40.

Leonor, G., David, W., Clara W., & Nigel, U. (2011). The International Diabetes Federation diabetes atlas methodology for estimating global and national prevalence of diabetes in adults. *Diabetes Research and Clinical Practice*, 94(3), 322-32.

Mesnita, G. (2017). Change of Functional Requirements For Information Systems Integration With Internet of Things. *Journal of Software & Systems Development*, 2017, 1-18.

Muhammad, S., & Khawaja, A. T. (2017). *Impact of Avoiding Non-functional Requirements in Software Development Stage*. https://www.semanticscholar.org

Partick, O. (2021). *Mbarara University works*. https:// lms. must. ac. ug/ claroline/ work/ user_ work.php?cmd=exDownload&authId=10912&assigId=3&workId=14&cidReset=true& cidReq=RM

Pastakia, P. H. (2018). Access to Hemoglobin A1c in Rural Africa: A Difficult Reality with Severe Consequences. *Journal of Diabetes Research*, 2018,1-6.

Raymond, M. M., Robert, K., & William, L. (2017). Access to medicines and diagnostic tests integral in the management of diabetes mellitus and cardiovascular diseases in Uganda: insights from the ACCODAD study. *International Journal for Equity in Health, 16*(1), 29-43.

Roy, W. M., David, G., Fredrick, M., Frederick, N. N., Stefan, P., Goran, T., & Claes, G. O. (2013). Diabetes and Pre-Diabetes among Persons Aged 35 to 60 Years in Eastern Uganda: Prevalence and Associated Factors. *Plos One*, 8(8), 29-59.

Roy, T., Al-Mrabeh, A., & Sattar, N. (2019). Understanding the mechanisms of reversal of type 2 diabetes. *Science Direct, 7*(9), 726-736.

Ruth, L. & Coleman, R. J. (2007). Framingham, SCORE, and DECODE Risk Equations Do Not Provide Reliable Cardiovascular Risk Estimates in Type 2 Diabetes. *American Diabetes Association, 30*(5), 1292-1293.

Bolli, G. B., Di Marchi, R. D., Park, G. D., Pramming, S., & Koivisto, V. A. (1999). Insulin analogues and their potential in the management. *Diabetologia*, 42(10), 1151-67.

Shilubane, E. P. (2007). Patients' and family members' knowledge and views regarding diabetes mellitus and its treatment. *Aosis, 30*, 56-62.

Shilubane, H. N. (2003). *Knowledge of patients and family members regarding diabetes mellitus and its treatment*. http://uir.unisa.ac.za/handle/10500/1450

Surabhi, N., Mina S., Argyro S., Ranjit A., & Kypros, H. N. (2011). Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. *An International Journal of Obstetrics & Gynaecology, 31*(2), 135-141.

Worldometer. (2021). *Uganda Population*. https:// www. worldometers. info/ world-population/ uganda- population/

Worldometer. (2021). *Uganda Population*. https:// www. worldometers. info/ world-population/ uganda- population/

APPENDICES

Appendix 1: Questionnaire

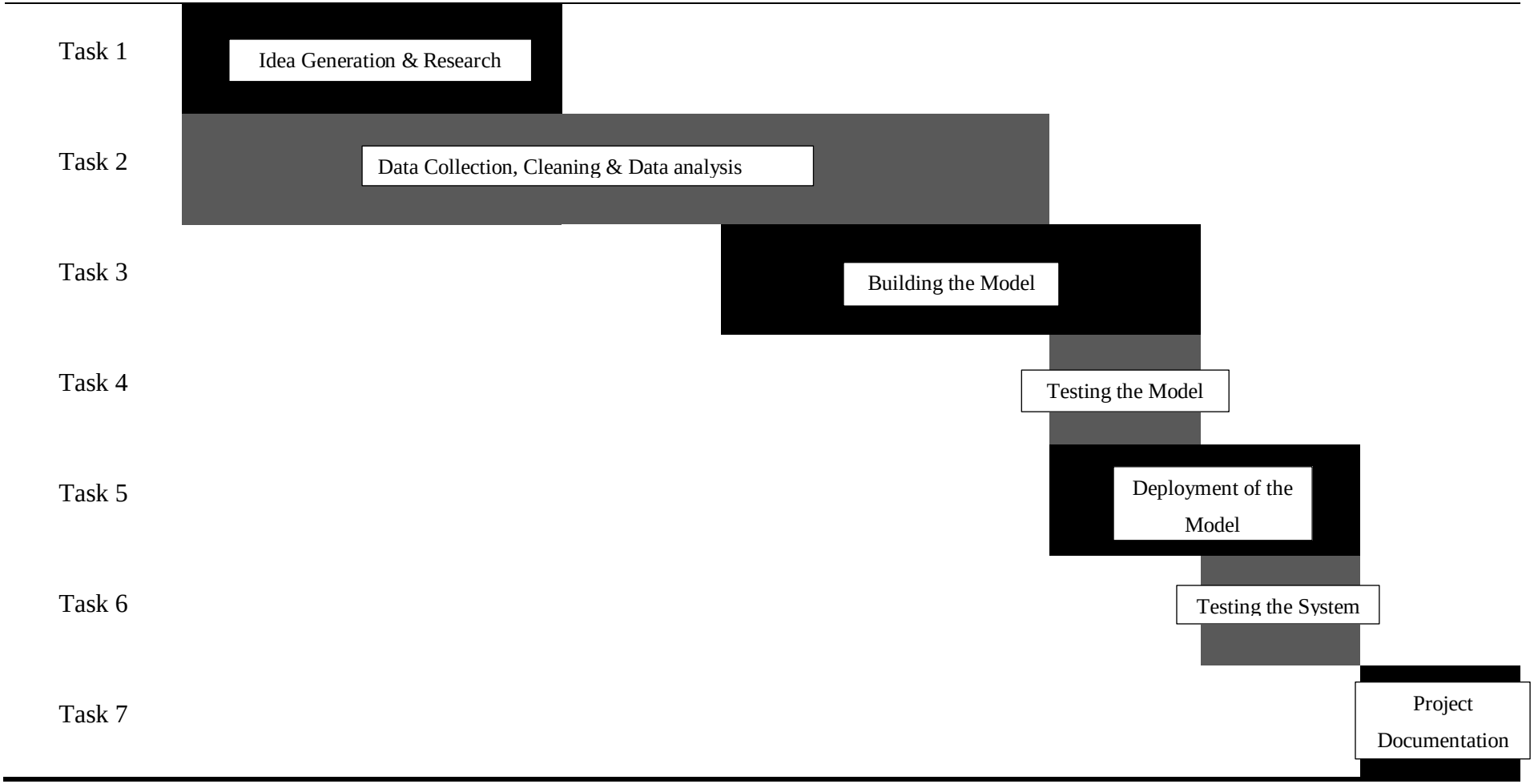
Attributes number	Attributes	Description	Type
1	Age	Age of respondent	numerical
2	Gender	Sex of respondent (0=Male, 1=Female)	categorical
3	Residence	Place of residence (1=central=, 2= eastern, 3= western, 4= northern)	numerical
4	systolic blood pressure	Systolic Blood Pressure (mmHg)	numerical
5	diastolic blood pressure	Diastolic blood pressure (mmHg)	numerical
6	Physical Exercise	Known to do Physical Exercise (0=No, 1=Yes)	categorical
7	BMI	Body mass index (kg/ m^2)	numerical
8	Hypertension	Blood Pressure (0=Normal, 1=Hypertension)	categorical
9	Family History of Diabetes	Have a family member with Diabetes (0=No, 1=Yes)	categorical
10	alcohol intake	alcohol intake (0=No, 1=Yes)	categorical
11	currently a smoker	currently a smoker (0=No, 1=Yes)	categorical
12	Occupation	Occupation (1=subsistence 2=former 3=trading)	Known to have categorical
13	Obesity	obesity (0=No, 1=Yes)	numerical

Appendix 2: The Budget for Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning

Costs for the project	Budget allocation in Euros	Details
Research and Feature Engineering (Data Collection)	**600**	Documentation and Access
Logistics – Travel	**300**	Travel during Data Collection
Project documentation	**400**	Report's printing
Internet (Data Bundle)	**400**	Buying a Modem and Data Bundles
Total	**1700**	

	October 2020	November 2020	December 2020	January 2021	February 2021	March 2021	April 2021	May 2021

Appendix 3: Work Break Structure for a Predictive Model for Early Detection of Diabetes Mellitus using Machine Learning


Task 1
Idea Generation & Research
Task 2
Data Collection, Cleaning & Data analysis
Task 3
Building the Model
Task 4
Testing the Model
Task 5
Deployment of the Model
Task 6
Testing the System
Task 7
Project Documentation

POSTER PRESENTATION

THE NELSON MANDELA AFRICAN INSTITUTION OF SCIENCE AND TECHNOLOGY

Henry Semakula[1], Dr. Michael Kisangirir[2], Dr. Edith Luhanga[3]

A PREDICTIVE MODEL FOR EARLY DETECTION OF DIABETES MELLITUS USING MACHINE LEARNING.

INTRODUCTION

Diabetes is one of the major public health issues, affecting over 425 million people worldwide, mostly adults over the age of 18. The word Diabetes consists of two Greek words “Dia” meaning through, and “betes” meaning pass, which refers to a cycle of heavy thirst and abnormal frequent urination.

OBJECTIVES

MAIN OBJECTIVE

To develop a web-based predictive model for possible cases of Type 2 diabetes based on input symptoms, to support early referral for testing.

SPECFIFIC OBJECTIVES

(i) To collect data for the risk factors that cause diabetes mellitus to be used for building a predictive model for early detection of Diabetes Mellitus.
(ii) To determine the most important features for predicting possible case of Type 2 diabetes.
(iii) To build a prediction model using classification methods.
(iv) To evaluate and validate the prediction model.
(v) To develop a web application based on the best algorithm.

CONCEPTUAL FRAMEWORK

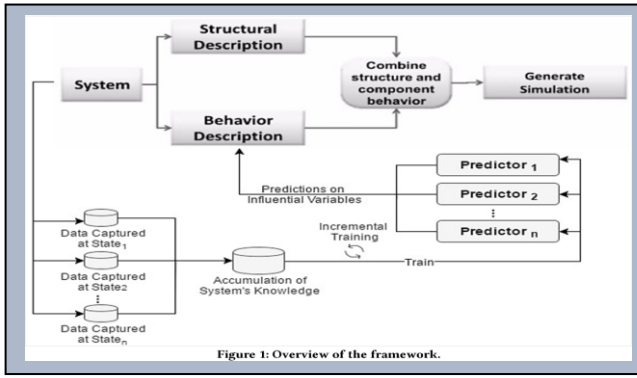
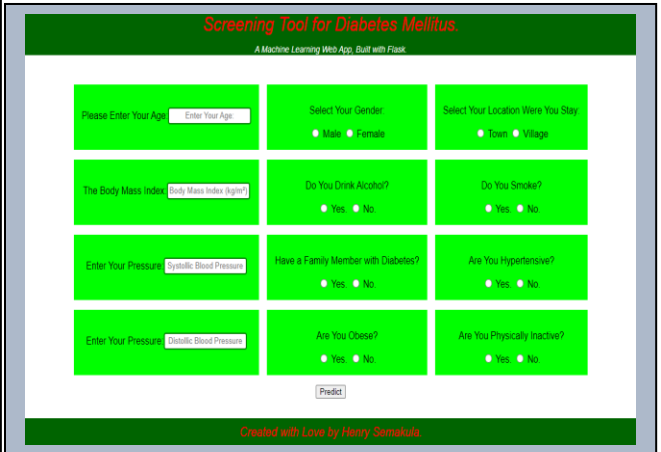


Figure 1: Overview of the framework.

WEB-SCREENING TOOL



PROBLEM STATEMENT

Late detection of diabetes can lead to severe complications such as blindness, impotence in male, kidney failure, cholesterol and heart diseases. There is also both a direct and indirect economic burden brought about by diabetes, particularly in low- and middle-income countries. Although various governments, including Uganda, have scaled up the purchasing and distribution of diabetes diagnostic kits, a majority of the population remains unknowledgeable about symptoms of the disease. which results in late seeking of care, and those in rural areas do not have enough access to diagnosis

METHODOLOGY

The Evolutionary prototyping methodology was chosen for development of the web application due to an application being novel and therefore, end-users did not fully know their requirements, and required multiple chances to provide feedback and revised the system based on their usage experiences and secondly the system was complex and required testing and validation at multiple times.