A Framework for Pregnancy-Induced Hypertension Prediction using Machine Learning Approach

By. Dureti Genemo

A Thesis Submitted to the Department of Computing School of Electrical Engineering and Computing.

Presented in Partial Fulfillment for the Degree of Masters of Science in

Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

January, 2020

Adama, Ethiopia

A Framework for Pregnancy Induced Hypertension Prediction using Machine Learning Approach

By: Dureti Genemo

Advisor. Jayant Shekhar(PhD)

A Thesis Submitted to the Department of Computing School of Electrical Engineering and Computing.

Presented in Partial Fulfillment for the Degree of Masters of Science in

Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

January, 2020

Adama, Ethiopia

DECLARATION

I hereby declare that this MSc thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Name: Dureti Genemo

Signature: ----------

This MSc thesis has been submitted for examination with my approval as a thesis advisor.

Name: Pf. Jayant Shekhar (PhD.)

Signature: ----------

Date of Submission: ----------

APPROVAL OF THE BOARD OF EXAMINATION

We, the members of the board of examiners of the final open defense by Dureti Genemo, have read and evaluated his/her thesis entitled "*A Framework for Pregnancy Induced Hypertension Prediction using Machine Learning Approach*" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Computer Science and Engineering.

_____	_____	_____
Supervisor/Advisor	Signature	Date
_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examinor	Signature	Date
_____	_____	_____
External Examinor	Signature	Date

ACKNOWLEDGEMENT

I would like to acknowledge God for his grace, strength, and good health as I undertook this research. My sincere gratitude to my advisor Pf. Jayant Shakhar for his readiness and willingness to advise on the research, his critical comments, commitment to guide, and support greatly improved this thesis work.

I would also like to acknowledge the members of Smart Software SIG of the CSE department. Including the SIG supervisor Dr. Mesifn Abebe for his continued commitment to guide and support this research from its inception to completion. Furthermore, to all members who sat on each phase presentation panels for their inputs which greatly shaped and improved the work. I would like to thanks all my classmates who are working on their master's thesis for their timely help, idea, and support until the completion of my thesis.

My special thanks also go to Efrem M. and Dr. Demu Tesfaye, who is vice president of academics and research of Adama Hospital for helping and allow me to collect data and. Also, I would like to thanks Dr. Tibesso Bulli, Dr.Tolosa, and Dr. Aklilu Hayilu for their major role in the dataset preparation and for medical suggestions and advises and participated in or contribution to this work.

Last but not least, my thanks goes to my wonderful husband and parents for their every support and advice in my life. May God always Bless and keep them safe.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATION

LIST OF ABBREVIATION

ACOG	American congress of obstetrician and gynecologists
BL	Bayesian Learning
BMI	Body Mass Index
CFS	Correlation Based Feature Selection
CV	Cross validation
DBP	Diastolic blood pressure
DHS	Demographic and Health Surveys
DT	Decision Tree
EDHS	Ethiopian Demographic Health survey
ETC	Extra Tree Classifier
F1	F measure
FE	Feature Extraction
FIRF	Feature Importance Using Random Forest
FN	False Negative
FP	False Posetive
FS	Feature Selection
GH	Gestational hypertensi
GTP	Growth Transformation and Plan
ICT	Information communication technology
KDD	Knowledge Discovery in Databases
MDP	Marcov Decision Process
NB	Naïve Bayes
P	Precision
PCA	Principal Component Analaysis
PIH	Pregnancy Induced Hypertension
Preeclam	Preeclampsia
PSO	Particle Swarm Optimization
R	Recall

RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
RMSE	Root Mean Square Error
SBP	Systolic blood pressure
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Posetive
WHO	World Health Officer

ABSTRACT

Pregnancy induced hypertension is a common and very severe medical disorder specific to pregnancy and it is a major health burden and one of the leading causes of maternal and perinatal morbidity and mortality. The goal of this study is to design a framework for pregnancy induced hypertension prediction through integrating healthcare server and machine learning algorithm and techniques. As a solution for this problem, this research proposed pregnancy induced hypertension prediction using machine learning and feature selection/extraction method to build a prediction model and used flask server to integrate prediction model with healthcare server. sample data was collected from three hospitals such as Adama hospital,asella hospital and noah speciality clinic and this hospitals was selected based on the largest services given for pregnant women. the sample data was processed and labeled with the help of health experts and classified into three class. The research employed an experimental approach to select the combination of feature extraction and selection method with a machine learning model using RF,SVM and NB. The models evaluated using 10-fold cross-validation, and classification performance is used to compare the models. The study resulted classification result using random forest with combination of feature extraction using PCA resulted with accuracy of 0.97 than SVM and NB. RF was selected to integrate with flask server and evaluated the framework by health experts.

Keywords: *Pregnancy-Induced hypertension, Feature selection,,feature extraction,Machine learning, Flask server*

CHAPTER ONE

1 Introduction

This chapter introduces the background of the study, motivation of the study, statement of the problems, the significance of the study, research question, objectives of the study, scope and limitation of the study, and application of result, followed by the organization of the research study.

1.1 Background of The Study

Globally, Pregnancy-induced hypertension complicates 10% of all pregnancies in developed and developing countries[1]. Around 40,000 women, mostly from developing countries die each year due to preeclampsia or eclampsia[2]. Pregnancy-induced hypertension is a major cause of pregnancy complications, causing premature delivery, fetal growth retardation, maternal mortality and morbidity and small for gestations. In addition, it has long-term health effects, like, chronic hypertension, kidney failure and nervous system disorders[3]. Pregnancy-induced hypertension is a major health burden and it is one of the leading causes of maternal and perinatal morbidity and mortality[2]. Moreover, it is a common and very severe medical disorder specific to pregnancy[4]. The term pregnancy-induced hypertension is commonly used to describe a wide spectrum of pregnant patients who have elevated blood pressure or severe hypertension with various organ dysfunctions. The indication in pregnant patients may be similar but they may result from different underlying causes such as chronic hypertension, renal disease or preeclampsia[5].

There are three classes of pregnancy-induced hypertension according to the latest classification system described in the National high blood pressure education group those are gestational hypertension, chronic hypertension and preeclampsia[3][6]. Hypertensive category of pregnancy-induced hypertension to call as chronic hypertension is a condition characterized by elevated blood pressure of any causes diagnosed before 20 weeks of gestation and gestational hypertension also an elevated blood pressure of pregnant women after 20 weeks of gestation without protein in urine and the last are preeclampsia, which is unique for pregnant women

which means it is only happening in pregnant women is an elevated blood pressure of pregnant women after 20 weeks of gestation and contains protein in urine[7][5].

There are many risk factor for pregnancy-induced hypertension which leads elevate blood pressure in addition to pregnancy, because pregnancy by itself disorder condition of pregnant women from normal women and it increase blood pressure somehow and this is normal in some extent but there are other factors which lead pregnant women to risk those are medical condition of pregnant women such as socio-demographic condition, obstetric condition and medical condition of regnant women[8]. There is a need to apply machine learning techniques to identify the leading factor of pregnancy induced hypertension. More over ,pregnant women do not know specific features and information about pregnancy induced hypertension and there is no awareness mechanism for pregnant women unless they are going to hospitals. Therefore, there needs to apply machine learning algorithms to identify and recommend the leading cause and features of pregnancy induced hypertension early.

Although, the development of PIH is increasing from time to time. However ,the focal point of many researcher is the prediction of physiological feature of pregnancy induced hypertension but socio-dimographic,obstetric and medical feature are factors associated with pregnancy induced hypertension but not touched and studied in this area. Methods such as ANN,RF,SVM and DT have been used to predict different type of disease like diabetes[9], liver disease[10] and heart disease[11]. However, few researchs have considered the use of machine learning technique to construct prediction model for pregnancy induced hypertension[12][13][14]. To the best of our knowledge, this study is the first study in Ethiopia for developing a framework for pregnancy induced hypertension prediction using machine learning approach. In this study the researcher is motivated to explore the potential applicability of machine learning to predict pregnancy induced hypertension and integrating with healthcare server for users to access and get benefit from this framework.

Machine learning techniques are the most potent and essential techniques now a day, especially in clinical settings and areas, machine learning helps to infer essential and relevant data items from different data items and a lot of data items that may be difficult to correlate. The nature of

clinical data is vast by its nature and complex medical information; therefore, machine learning recognized as a method for diagnosing and predicting clinical outcomes. There are many machine learning techniques used and applied in clinical data analysis and prediction for predicting preeclampsia[15], machine learning algorithm for diabetes prediction[16][17], machine learning algorithm for chronic kidney disease[18][19], machine learning algorithms for heart disease [20][11][21] With higher accuracy than other methods. Since machine learning algorithms trained with a lot of clinical data, it is the most recommended and preferable approach to use in clinical data analysis.

The integration of machine learning with healthcare provides full information and classification of pregnancy induced hypertension both for user and professionals. More over helps to process huge amount of dataset beyond scope of human capability and converts the analaysis of that data into clinical insight that helps physician in planning ,providing care, leading to better outcome, lower costs of care and increased patient satisfaction.

This study aims to develop a framework for pregnancy induced hypertension prediction by integrating machine learning algorithms and healthcare server. The study build predictive model using supervised machine learning algorithm such as SVM,RF and NB with combination of different feature selection and extraction method to get more predictive model and reduce dimension of features. The study aims to integrate the predictive model with healthcare server for deploying the predictive model to production and helps user access the predictive model fro this healthcare server.

Strengthes,Weaknesses,Opportunities, Threats (SWOT) is a tool used to analysis the internal and external factor of ICT in healthcare sectors. Internal factors are weak or strength for applying ICT in health sector, but the external factors are opportunities or threats of applying ICT in health sector.

Strengths, Ethiopia prepared a medium strategic plan entitled”the Growth and Transformation Plan(GTP)”. To improve quality of services thereby achieving health related issue such as reducing child mortality, improve maternal health and combat HIV,AIDS and other

communicable disease. Incorporating ICT in health sector of Ethiopia helps to improve the delivery of health related services such as eHealth practitioners to address their services to communicate based on explicit needs and helps medical specialist to conduct remote health consultation. Moreover, eHealth helps for diagnosis and treatment using the latest health science evidence, facilitates communication among health professional for evidence-based medicine ,strengthens monitoring and control of disease outbreaks, increase administrative and management efficiency within primary, secondary, and tertiary care[22]. Many of the HDSP's components, such as health services delivery, pharmaceutical services, information, education and communication, health management information systems, and monitoring and evaluation can be strengthened through the use of ICT[22]. Weakness, Health facilities still struggle to provide some of the priority services, such as deliveries, in a manner that attracts mothers and other patients[23].

Opportunities, there is a strong commitment from local and national authorities to improve the efficiencies of the health sector, with several projects embracing ICT for healthcare services[22]. Threats, lack of access to electricity, solar power options, and power supply back-ups; insufficient infrastructure and connectivity access; and high costs[23].

1.2 The Motivation of The Study

Many researcher try to identify the cause and effect of pregnancy-induced hypertension globally and locally and the solution was given for how to manage this disorder and the other are prepared documents to manage pregnancy-induced hypertension but it indicates that the whole activity is to health professional or doctors this may increase burden to health worker, and patients may not identify their disorder quickly. WHO recommends and prepared hypertension in pregnancy and how to manage[24] and this document updated regularly, this document helps health professional to diagnose pregnant women with hypertension but there is no mechanism and method for accessing information and recommendation for pregnant women early about pregnancy induced hypertension.

In Ethiopia, also different researchers study the prevalence of pregnancy-induced hypertension, its impact and they suggest some solutions for the healthcare to manage pregnancy-induced hypertension orally in healthcare. Therefore, this case inspired us to develop pregnancy-induced hypertension prediction using a machine learning approach because now a day's machine

learning is the best interactive and preferred approach, especially in health institution for it, is trained by feeding many datasets and will give an efficient result. To contribute by preparing a dataset for pregnancy-induced hypertension for better prediction and analysis by other researchers and future work.

1.3 Statement of The Problem

Population based study conducted in south Africa, the prevalence of pregnancy induced hypertension was 12% and it is the common maternal death of 20.7%[25]. studies in ethiopia show that the prevalence of pregnancy induced hypertension is around 19% of which majority were due to preeclampsia[25]

As Ethiopian Demographic Health survey (EDHS) 2016 reported, maternal mortality ratio is 412 deaths per 100,000 live births and pregnancy induced hypertension has a great role for this maternal death[25]. A review study conducted on the causes of maternal mortality in Ethiopia showed that, the proportion of maternal mortality in Ethiopia due to hypertensive disorders between 1980 and 2012 is increased from 4%-29% at different health facilities[25].

Pregnancy-induced hypertension affects both the mother and the fetus in the womb. Therefore, unless it is identified early, it may complicate pregnancy because one of the causes of maternal mortality during pregnancy is hypertension complication. However, Most women do not have awareness about the complication of hypertension disorder and lack of information, this may lead to many mortality rates for babies and also the mother's kidney failure.

Due to lack of information during pregnancy, there are different problem may happen to women those are:-

- Preterm delivery of the baby
- Fetal health as well as its weight are highly compromised,
- leading to various degrees of fetal morbidity, and fetal damage may be such as to cause fetal death.
- Kidney failure of mother
- Seizures in the mother and Stroke

Research on pregnancy-induced hypertension is studied, but it is not enough for managing and detecting this problem because the nature of pregnancy-induced hypertension is changed from one another and from one area to another area those are developed data analysis for preeclampsia [13][26], blood pressure level prediction[27], rule-based system [28] and the other papers are trying to predict pregnancy-induced hypertension using one feature.

In Ethiopia, there is no study done for prediction of pregnancy-induced hypertension to help a user and professional which can help them early to identify their problem and to get information.

The following research questions formulated to answer the above problem:

- How to integrate the machine learning model and healthcare server ?
- What is the effective tools used to design the framework?
- What is the most effective feature extraction method to identify predictive feature of pregnancy-induced hypertension?
- which machine learning algorithms produces best prediction model for pregnancy-induced hypertension?

1.4 Significance of The Study

The findings of this research will be used to design a framework for pregnancy induced hypertension. It contributes a great deal of benefit to pregnant women, health professionals, policy makers, programmers and researchers.

Accordingly, the study has the following significances:

- It helps pregnant women to get early identification and information of pregnancy induced hypertension. it eradicates complication which might happen due to lack of awareness for the pregnancy induced hypertension during the right time.
- It helps health professionals to use the framework to predict routinely pregnancy induced hypertension.
- Policy makers and programmers can make use of this framework to develop new guidelines and policies and/or modify the existing ones in order to improve achievement of their goals.
- Can also serve as a baseline data reference and initiative other researchers to conduct further studies in the future.

1.5 Objectives of the Study

1.5.1 General Objective

The general objective of this study is to design and develop a framework for pregnancy-induced hypertension prediction using a machine learning approach

1.5.2 Specific Objective

To achieve the general objective of this research, the following specific objectives are envisaged:

- To review literature in area of framework for pregnancy induced hypertension.
- To collect data and preparing data for analysis and model building by cleaning, extracting, and transforming it into a format suitable for machine learning algorithms.
- To prepare dataset and training and testing.
- To identify important features for pregnancy induced hypertension.
- To identify the effective tool for designing a framework.
- To build a framework for pregnancy-induced hypertension.
- To identify effective feature selection and extraction algorithm for pregnancy-induced hypertension.
- To build a prediction model using machine learning algorithms for pregnancy-induced hypertension.
- To compare and evaluate the performance of the resultant prediction models, and select the one which performs the best.
- To recommend the best prediction model based on the evaluation result.
- To evaluate the framework by stakeholders.

1.6 Scope and Limitation of the study

1.6.1 Scope of the study

The scope of the study is applying supervised machine learning algorithms and predicting the status of pregnancy induced hypertension of pregnant women. The study focus on socio-demographic,obstetric and medical factor of pregnancy induced hypertension and data were

collected and dataset prepared. After applying classification algorithm, the highest performance accuracy algorithm will be integrated with healthcare server using flask server.

1.6.2 Limitation of the study

This study comes across this limitations on a different phase of the research process. Due to the lack of an online and open access dataset, the study restricted to collect data from hospitals for prediction of the model. The other limitation of this study due to shortage of time and organized record of patients in the hospitals, the study collected data from only three hospitals for analyzing and prediction model.

Lack of relevant related literature on machine learning approach in the area of pregnancy induced hypertension in Ethiopia, is one of the limitations encountered for experience sharing and comparing the result of this study to undertake the study.

1.7 Beneficiary of the study

The primary beneficiary of this framework is is women whether they are pregnant or not they can get benefit from this framework for identifying and get awareness of their health status early.The other beneficiary of this model is Healthcare professionals for predicting pregnancy induced hypertension, especially when there is a shortage of gynecologists and obstetricians and for small and local healthcare because most of the time there is a shortage of doctors in small healthcare.

The last beneficiary is Different researcher can benefit for a research study this means researchers can replicate the proposed research for pregnancy-induced hypertension in another area of Ethiopia and can serve as a baseline for related work on this issue or use the dataset to improve the research.

1.8 Thesis Organization

The research paper organized under seven chapters in the following ways:

Chapter one discussed above, and It includes the introduction of the study, the motivation, and the statement of problems, the research questions that would be answered in the proposed

solutions, the scope and limitation, the objective of the study, the methodology, and the application results of the study.

Chapter two presents the literature review and related works on pregnancy-induced hypertension, definition of pregnancy-induced hypertension, global impact of pregnancy-induced hypertension and their impact in Ethiopia, technologies and techniques used to manage pregnancy-induced hypertension: mobile health solution for pregnancy-induced hypertension, remote monitoring of pregnancy-induced hypertension, feature selection and machine learning classifier and finally, it discussed the related work.

Chapter three discusses the research methodology used in this research work, methods used to build the dataset, methods used to preprocess data for pregnancy-induced hypertension, feature extraction methods used, and machine learning classifier algorithm's, methods to deploy machine learning model in server and finally, it presents the evaluation methods.

Chapter four discusses the proposed solution for pregnancy-induced hypertension prediction, which is the proposed data preprocessing, feature selection, machine learning training, and testing.

Chapter five discusses the implementation and experimentation of the proposed solution including the working environment, dataset description, also the implementation of preprocessing, feature selection implementation, models implementations, evaluation models and machine learning model deployment in a server.

Chapter six discusses the result of the processed dataset, feature selection, machine learning prediction model and also discusses the significant results obtained by comparing models based on features.

Chapter seven presents a conclusion, recommendation and identifies future works for further research direction on this research topic.

CHAPTER TWO

2 Literature Review and Related works

2.1 Introduction

Literature in the area of pregnancy Induced Hypertension prediction has been reviewed and discussed. The review covers an overview of pregnancy-induced hypertension, existing framework for pregnancy induced hypertension prediction , feature selection and extraction method were reviewed. Review facilitates the comparison of approaches to the specific solutions that have been employed in this study.

2.1.1 Overview of Pregnancy Induced Hypertension

pregnancy-induced hypertension(PIH) refers to conditions characterized by elevated blood pressure from normal blood pressure this means systolic blood pressure of greater than 120 mmHg and diastolic blood pressure of greater than 80 mmHg and any cause diagnosed before 20 weeks of gestational hypertension is called pregnancy-induced hypertension[29]**.** According to the American college of obstetricians and gynecologists (ACOG), Hypertension in pregnancy is defined as when Systolic blood pressure is > = or equal to 140 mmHg and diastolic blood pressure is > = or equal to 90 mmHg in two occasions at least six hour apart after fifth month of gestation for pregnancy-induced hypertension or before pregnancy/before 20 weeks of gestation for chronic hypertension.

2.2 Machine Learning Approach in Healthcare Sector

Now a days healthcare sector has been adopted and used machine learning technology and machine learning technology plays a great role in healthrelated issue such as developing of new new medical procedure,recording of patients data and treatment of chronic disease. Moreover, machine learning plays a great role in clinical decision support and early identification of disease[30]. Healthcare is a wide area which needs concentration and involves the improvement of medical services in order to serve the medical demands of the people. healthcare providers have begun using tools built from machine learning models that use anomaly detection algorithms to predict heart attacks, strokes, sepsis and other serious complications. These tools

use data from patients' historical records, daily evaluations, and measurements of vital signs in real-time, such as heart rate and blood pressure, to alert staff of imminent patient risks so they can immediately take preventive actions[31].

2.2.1 An Overview of Machine Learning

Machine learning is a type of AI and which is learning of human knowledge from their experience using a computer. Machine learning contains computer science which helps to identify and solve problems and statistics which allows for feature modeling, hypothesis and measures reliability of data. Machine learning learns from past experience to improve future experience and improvement of algorithms without human being involvement[32].

Generally, machine learning approach or techniques pass through the following steps to perform the activity given to it. The first one is fetching or accepting data which is the first and most component or element of machine learning then,cleaning the accepted or fetched data for more understanding data ,the next one is preparing and trainging data ,lastly evaluating the trained data and deploy to production or server for users to access the model[33].

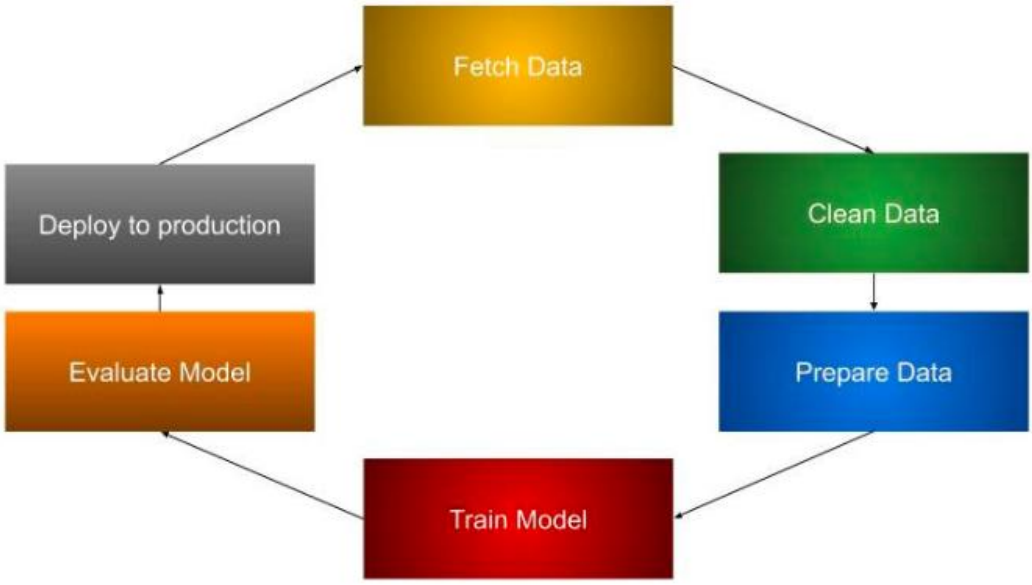


Figure 2.1 Architecture of Machine Learning[33]

2.2.2 Types of Machine Learning

***Supervised Learning Algorithms*: -** This is one of the types of machine learning algorithms that are trained and predicted based on the given training dataset and also helps to predict output value from the given input features. The training dataset contains both input and output value. Choosing supervised algorithms depends on the output value because there are two types of output values, continuous and discrete. The algorithm for continuous features or output value is regression and for discrete features are classifications. There are two types of supervised learning algorithms those are a parametric model based on the combination of a fixed number of parameters such as linear regression and classification. The other is non-parametric models such as support vector machine and nearest neighbor algorithms. The difference between the two types of algorithms is, in parametric types of algorithms, the prediction is based on the learned parameters but in non-parametric types, a prediction is based on the training dataset[32].

Un –Supervised Learning Algorithms:- This type of machine learning algorithm is a prediction of the output based on the similarity and of input features and contains only input variables. One of the types of unsupervised learning algorithms is clustering. It is a method to similar group data based on their distance and their criteria are high intracluster similarity and low inter-cluster similarity. The application area of unsupervised is social network analysis, genetic clustering, market analysis[32]. The authors of this work[34] propose an unsupervised machine learning model that has the ability to identify real-world latent infectious diseases by mining social media data. The study collects data from social media using sentiment analysis techniques and presents a bottom-up approach for latent infectious disease discovery in a given location without prior information, such as disease names and related symptoms.

Semi-Supervised Learning Algorithms:- This is the combination of both supervised and unsupervised learning algorithms, and this helps to solve when there is no labeled dataset or expensive.it works first labeled dataset is used using supervised learning algorithm then unlabeled dataset with previous values finally all values are used[32]. Md. Shahriare Satu,et al [35]explores significant heart disease factor using semi supervised learning algorithms.this study used Cleveland and Hungarian heart disease data which are collected from UCI machine learning repository based on this data, the researcher applied feature selection method like Correlation-

based Feature Subset Selection (CFS), Correlation Attribute Evaluation (CAE), Gain Ratio Attribute Evaluation (GRAE), Info Gain Attribute Evaluation (IGAE), OneR Attribute Evaluation (OneRAE) techniques with semi supervised learning algorithms such as Collective Wrapper (CW), Filtered Collective (FC) and Yet Another Semi Supervised Idea (YATSI) are used to classify heart disease data.

Reinforcement Algorithm:- This is Works in trial and error and delayed reward to get the correct output from training set using evaluative feedback. This algorithm follows Markov Decision Process (MDP) is intimated when the result is wrong and explores the possibility to find the right result[32]. Reinforcement algorithm related models and approaches have been widely applied in healthcare domains these application domains can be categorized into three main types: dynamic treatment regimes in chronic disease or critical care, automated medical diagnosis, and other general domains such as health resources allocation and scheduling, optimal process control[36].

2.2.3 Application of Machine Learning in Health Sector

Machine learning is a tool and techniques which can helps to solve diagnostic and prognostic problem in a different medical domain and also it is used for the analaysis of clinical features with ther combination. Moreover machine learning is used for detection of regularities in the data by dealing with imperfect data. Therefore Implementation of ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to facilitate and enhance the work of medical experts and ultimately to improve the efficiency and quality of medical care. many researcher have worked on machine learning algorithms to solve healthcare problem especially disease diagnosis of patients and researcher accepted machine learning algorithm as a best tool and techniques to diagnose disease. Application of machine learning in healthcare are reviewed and shown below.

Heart disease prediction: Youness et.al [21] proposed heart disease prediction using machine learning algorithm with combination of particle swarm optimization(PSO) and ant colony(ACO). This study used fast correlation-based feature selection(fcfs) for selecting features and for filtering redundant features of heart disease dataset. The study used classification algorithms such as SVM,KNN, NB and ANN optimized by PSO and ACO approach. The study compares

all results and show the highest performance accuracy of 99.65% using FCBF,PSO and ACO. Machine learning algorithms for prevalence of heart disease prediction has been discussed by Divyansh et. al[37]. The employed dataset is UCI heart disease dataset containing 76 features and applied classification algorithms those are SVM,LR and NN. Result shows lesser complex using logistic regression and support vector machines with linear kernel give more accurate results.

Diabetes Disease Prediction: Research on diabete prediction done by Quan et al. [17] to predict diabetes mellitus using DT, RF and NN. The dataset is the hospital physical examination data in Luzhou, China and contains 14 attributes. This study used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) for dimensionality reduction and applied machine learning classification algorithms. The prediction model evaluated using 5-fold cross validation and results showed that prediction with RF produced the highest accuracy(ACC = 0.8084). The other research study done is prediction model for diabetes mellitus using machine learning algorithms. The study used classification algorithms those are DT, KNN, SVM, and RF. Pima Indians diabetes dataset were used to feed the classification algorithms and produced 0.96 accuracy performance using RF[16].

As the researcher applied and used in previous works, the most frequently used machine learning algorithm and approach were used in healthcare sector were identified. The following machine learning algorithms are used frequently in pregnancy-induced hypertension classification and processing and we described here their pros and cons of each algorithm.

Table 2.1: Frequently used Machine Learning Algorithms for Pregnancy Induced Hypertension[38]

No	***Algorithms***	***Advantages***	***Disadvantages***
1	***Decision tree***	handle categorical features, perform well in datasets with a large number of features and high speed of classification.	Lead to overfitting and have a parity problem that means it cannot tolerate interdependent variables.
2	***K-nearest neighbor***	Helps for learning in incrementing distances, and the speed of classification depends on the number of instances and attributes.	The number of k values is defined by a user and low tolerance of interdependent variables,

3	***Support vector machine***	Used for a non-linear solution, tolerance to irrelevant features and have high speed of classification.	Kernel knowledge is needed to get a high performance of classification.
4	***Naïve Bayes***	Helps for small datasets and for independent conditional variables	Performing in assumption for independent variables
5	***Random forest***	Helps to avoid overfitting and helps in estimating important features and rank based on their importance	Time-consuming during modeling classification problem

2.3 Feature Selection and Extraction Method used in Health Sector

There are several feature selection techniques used in the health sector, before any machine learning model is applied, feature selection is an important activity to remove noisy, inconsistent data. Reduce the dimensionality of a dataset to get more efficient and accurate results. If the feature selection algorithm is applied in the dataset, the complexity is decreased.

2.3.1 Types of Feature Selection

Feature extraction:- is a method of reducing features and provides a small feature set that contains relevant and useful information from the original dataset by applying transformation which produces the projection of the original dataset. Popular feature extraction algorithms are principal component analysis and least discriminant analysis algorithms that are frequently used in the clinical dataset for predicting diabetes, heart disease, stroke, hypertension etc[39]**.** Changsheng Zhu et al.[40] proposed and developed improved logistic regression techniques to predict diabetes using PCA and K-means algorithms. The objectives of this study is to improve the performance of logistic regression by comprising PCA and K-means algorithms. The experimental result shows that PCA enhanced the logistic regression classifier accuracy of 1.98% versus the result of other published studies.

Feature selection:- is a method that selects a subset of features based on statistical data analysis criteria from the original dataset. This selection method selects important features and ranks those features based on their statistical score results[39]. In clinical area feature selection helps to

identify and select important variables from patient's data and allows to identify relevant features, which are the most predictor of the disease some area used feature selection are diabetes prediction, hypertension prediction, and heart disease prediction. There are three types of feature selection methods those are filter method, wrapper method and embedded method[39].

Filter Method:- It is a method of selecting and filtering variables before implementation or applying of machine learning model such that classification or regression algorithms. This method mostly used for larger dataset because it is efficient and fast to filter features for execution but the drawback of this method is that it cannot allow identifying the interaction and relationship of dependent and independent variables or feature[41][42]. We can classify the dataset after feature measuring is performed based on filter method some example of filter method feature selection is univariate feature selection, correlation-based feature selection and F1 score based feature selection etc.

Figure 2.2 Filter Method[43]

Huseyin Polat. et al.[44] proposed diagnosis of chronic kidney disease using filter based feature selection method. The study used correlation feature selection subset evaluator with greedy stepwise search engine and filtered subset evaluator with the Best First search engine were used. The results showed that the SVM classifier by using filtered subset evaluator with the Best First search engine feature selection method has higher accuracy rate (98.5%) when compared to other selected methods.

Wrapper Method:- These types of methods are selection features by applying machine learning algorithms and gets the best and relevant features and performs an optimal selection of features. The other aim of this feature selection method is that, it helps to identify the relationship and

dependencies of independent features and dependent features and gives accurate result but the drawback of this method is, leads to overfitting for small dataset. Example of wrapper method is recursive feature elimination, forward-searching, and backward searching algorithm etc.[42]

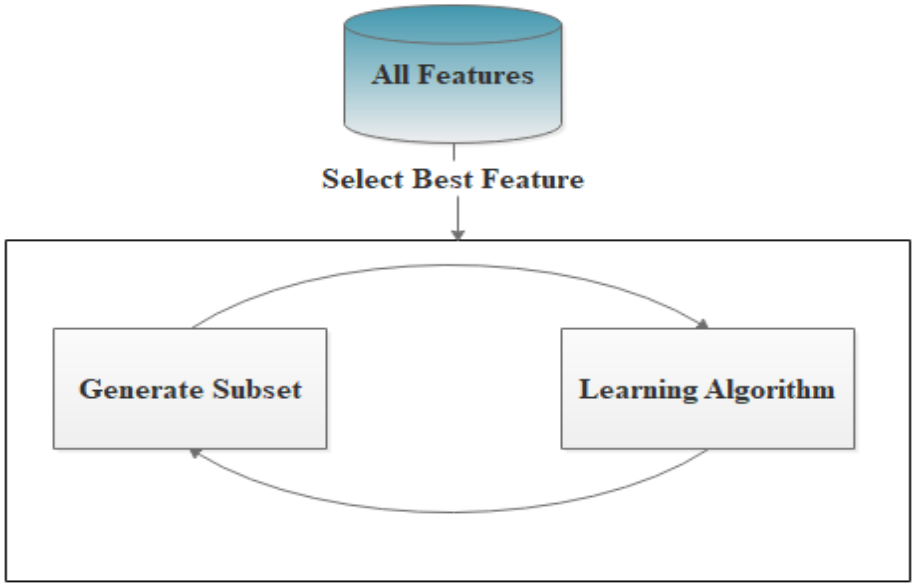


Figure 2.3 Wrapper Method[43]

Prediction of hepatitis prognosis using support vector machines and wrapper method proposed by A.H.Roslina [45]. This study used Wrapper Methods such as Recursive feature elimination and forwad searching to remove noise features before applying classification algorithms. This study compares and shows the difference between before using feature selection and after feature selection and resulted in increase in prediction between data by combining feature selection method prior to classification process.

Embedded Method:- This method selects the best features during the machine learning process and selects essential and relevant features as training and learning algorithms. The advantages of the embedded method are computationally inexpensive but the drawback is, it needs decision-based on classifier[42]

2.3.2 Feature Selection and Extraction Method in PIH

From reviewed literature and experiment , the most frequently used feature selection and extraction method for selecting features are described. The following feature selection and

feature extraction algorithms used in pregnancy-induced hypertension to select and extract essential and relevant features from the original dataset described in tables 2.1

Table 2.2: Frequently used Feature Selection and Extraction Technique for Clinical Data

Feature selection & extraction method	***Algorithms of feature selection***	***Description***
Filter method[46] [41]	Correlation based feature selection (CFS)	Helps to Find the relationship or correlation and uncorrelation of input feature subset with the target features or dependent features
	Univariate feature selection(UFS)	Rank independent features based on a statistical score of each feature and allow users to select the best features.
Wrapper method[47][39][48]	Forward Selections	Evaluates features individually and select the best performing algorithm, then evaluates all possible combination of selected features and select the pair which produces the best algorithm based on pre-test criteria.
	Backward elimination	starts by fitting a model using all features. Then it removes one feature. It will remove the one that produces the highest performing algorithm (least statistically significant) for a specific evaluation criterion. In the second step, it will remove a second feature, the one that again produces the best performing algorithm. Moreover, it proceeds, removing features until a certain criterion is met.
	Recursive Feature Elimination (RFE)	Works recursively by removing attributes and build a model with the left attributes. To identify the attributes or combination of attributes which contribute the best predicting the target attributes used model accuracy result.
Embedded method[39], [42], [47]–[49]	Feature importance with extra tree classifier (ETC)	This approach used to estimate essential features.

	Random Forest Importance	Are a base estimator or decision tree and random extraction of features from the dataset
	Gradient Boosted Trees Importance	Calculates and provides essential features
Feature extraction method [49]	Principal component analysis (PCA)	extract independent features from the original dataset using the linear transformation

2.4 Related work

This section presents a research review of necessary related works to the area of pregnancy-induced hypertension prediction using machine learning algorithms to clearly understand the general technique, method, and result of existing studies. Few studies have carried pregnancy-induced hypertension prediction, however in Ethiopia ,until know there is no research study done in pregnancy induced hypertension but there are some papers done for other disease prediction using data mining and machine learning approach.Therefore papers related to this concept have been reviewed here.

Brook et al[50] proposed predicting skilled women delivery services in Ethiopia using logistic regression and machine learning algorithms. Data were Collected about skilled delivery services and their actors from a Demographic and Health Surveys (DHS) and a standardized women's questionnaire. The study used Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and minimize sampling errors and based on this result ,appled classification algorithms such as NB, J48, ANN, and SVM in WEKA. One of the limitation of this study is used secondary data. Because of this authors were unable to conduct further analysis on key contexts such as maternal deaths. The study suggests that from the classification model developed using J48, ANN, SVM, and Naïve Bayes data mining algorithms, J48 algorithm had a relatively more accurate predictive performance. The results of this study can be useful for developing a web-based application.

Gebremedhin[51] analayze Children's Data Using Machine Learning algorithms. The objectives of this study is for classification of childrens to help outside donors of the love for children organization and to get full information about each child for internal purpose of organization. To achieve this objectives, the researcher used three classification algorithms such as DT, BN and NN within the framework of KDD (Knowledge Discovery in Databases) data mining model. 17044 records or data were collected and cleaned before applying classification algorithms and evaluated using 10-fold cross validation techniques. The research concludes that decision tree (98.83%) should be selected as a model because it gives better results than Bayesian learning (98.32%) and better advantage over Neural Network (98.86) for the classification of organizations children.

Geletaw[52] proposed maternal care data mining in Ethiopia to identify factors which affects postnatal care visit in Ethiopia. The study applied Decision tree (using J48 algorithm) and rule induction techniques on 6558 records of Ethiopian demographic and health survey data. The result shows that J48 algorithm identified the most determinant factors to predict maternal healthcare visit after birth with 93.97 % accuracy and rule induction with accuracy of 93.96%. the researcher suggests that, the result obtained highly supportive to construct, evaluate and update advertising and promotional maternal health policies.

Paper[14] put forward a Boosting Random forest approach for women suffering from maternal hypertensive Disorder. In the proposed model, a boosting technique for maternal disease detection is applied. It is a method used to boost single trees into robust learning algorithms. Finally, it combines the classifiers by letting them vote on a final prediction and achieves accuracy by combining the random features with a method of boosting. Boosted trees try to improve the model fit over different trees by considering past fits. The classifiers focus on the cases which are incorrectly classified in the last round. The primary intent of boosted random forest is to maintain generality and helps attain excellent classification performance.

In paper[53] used Artificial neural network for pregnancy hypertensive disorder classification based on maternal heart rate variability and used the hyperbolic tangent as the activation function in the hidden layer and the SoftMax function for the output layer, for training they applied batch method and the scaled conjugate gradient as an optimization algorithm.

In addition to that, some papers applied machine learning algorithms for diagnosing and detecting preeclampsia which is happening on pregnant women.[28] a real dataset containing data of pregnant women with high risk of preeclampsia from a health center has been analyzed, and a fuzzy linguistic methodology with two main phases is used. Firstly, linguistic transformation is applied to the dataset to increase the interpretability and flexibility in the analysis of preeclampsia. Secondly, knowledge extraction is done by means of inferring rules using decision trees to classify the dataset. The obtained linguistic rules provide monitoring of preeclampsia based on wearable applications and devices.

Moreira. Et al. [13] Proposes the construction of a system to support intelligent decision applied to the diagnosis of preeclampsia using Bayesian networks to help experts in the pregnant care and the proposed decision support system used probabilistic concepts for decision-making. The network structure was obtained from medical references. Noisy-or model was also considered in this work. The operation of the network shows that, in certain cases, this type of modeling can be used profitably, especially when it has a large number of parents, and when parents have characteristics in common.

Neocleous.et al.[26] used a multi-slab neural structure of four slabs connected for parameters obtained from clinical data set, the best results were obtained with a multi-slab neural structure and. In the training set there was a correct classification of the 83.6% cases of preeclampsia and in the test set 93.8%. to improve the medical the diagnosis and the prevention of complex diseases and syndrome machine learning approach is applied in[54] for preeclampsia risk factor and association The aim to use ML techniques for these study case is to detect patterns and relations than a human, or traditional statistics cannot. Models are evaluated by metrics commonly used by the scientific community such as accuracy, the root mean square error

(RMSE), sensitivity or specificity. These metrics allow to effectively evaluate model precision and to compare the results with the research community.

Sharma.[27] studied pregnancy induced hypertension and the level, this paper used 100 data set and based on two parameters that is systolic and diastolic blood pressure then predict the level of pregnancy induced hypertension using machine learning algorithms such as decision tree and support vector machine with accuracy of 90%,97% respectively.[15] Developed models using machine learning to predict preeclampsia of late onset from electronic clinical recorded dataset. They applied machine learning models support vector machine, logistic regression, stochastic gradient boosting, random forest, naïve bays and applied c-statistics to evaluate the performance of the algorithm. To select important features, they used pattern recognition and cluster analysis algorithm and the deploy machine learning model on selected features.

2.5 Summary of Related Work

To summarize the related work, we have described our related work in table which contains title of reviewed paper, methodology they used and gaps. From reviewed paper we have seen that most paper study about data analysis of preeclampsia and some are for pregnancy induced hypertension additionally their data analysis depends on one parameters such as blood pressure measurement, this is difficult to predict this disorder based on one parameters and the other papers are used machine learning model for a few dataset like 15-25 number of dataset , however accuracy of machine learning algorithm depends on the number of dataset and to train machine learning model with high accuracy and it needs more dataset.

Table 2.3: Summary of Related Work

Ref.	***Title***	***Year***	***Methodology***	***Achievement(acc.)***	***Gaps***
[51]	Analyzing Children's Data Using Machine Learning: A Case of Ethiopia	*2018*	DT, BL and NN	DT of 98.83%	• Used few data sanple to generate a rule and It cannot predict pregnancy induced hypertension.
[50]	Predicting skilled delivery service use in Ethiopia: dual application of logistic regression and machine learning algorithms	*2019*	Used SMOTE analysis technique to balance the dataset and multivariable logistic regression to predict	LR of 95%	• Used secondary data • It cannot predict pregnancy induced hypertension
[52]	Ethiopic maternal care data mining: discovering the factors that affect postnatal care visit in Ethiopia	*2016*	Decision tree and rule induction techniques	J48 of 93.97 %	• miss learning capability • It cannot predict pregnancy inducdd hypertension
[53]	Artificial neural network for normal, hypertensive, and preeclampsia	2016	Artificial neural network for classification of hypertension in pregnancy	ANN of 80%	• It cannot predict preeclampsia • Used one feature
[55]	Predicting Hypertensive Disorders in High-risk Pregnancy Using the Random Forest Approach	2017	Tree based random forest classifier for predicting high- risk pregnancy hypertension	RF of 79%	• It cannot identify predicting features and feature extraction

[26]	Neural networks to estimate the risk for preeclampsia occurrence	2018	Feedforward neural structures, both standard multilayer and multi-slab.	FW NN of 83%	• Not used feature extraction • It cannot predicti pregnancy induced hypertension
[54]	Machine Learning Approach for Pre-Eclampsia Risk Factors Association	2018	Applied machine learning for preeclampsia diagnosis.	RF of 84%.	• It cannot predict pregnancy induced hypertension
[27]	Pregnancy induced hypertension level prediction using machine learning approach	2019	Developed machine learning algorithm such as support vector machine, Decision tree, KNN, Random forest	RF of 90%.	• They only focus on one features to predict the disease using. • It cannot identify pregnant women and normal.

CHAPTER THREE

3 Research Methodology

3.1 Overview

This chapter explains the research methodology and the process of developing a framework for pregnancy-induced hypertension prediction using machine learning algorithms. This chapter contains four phases and a given explanation and justification. The first phase explains and justifies data collection methods and techniques to collect data is described.

The second phase explains method and algorithms for experiment such as data preprocessing method which is an essential issue during machine learning model building and feature selection method which is also vital activity to extract and select essential and the most predictor features. The third phase is selecting and explaining machine learning algorithms which will be used in this research study that used to predict and classify pregnancy-induced hypertension based on the provided dataset. The fourth phase is a method and tools to integrate machine learning with healthcare server and finally evaluating the framework.

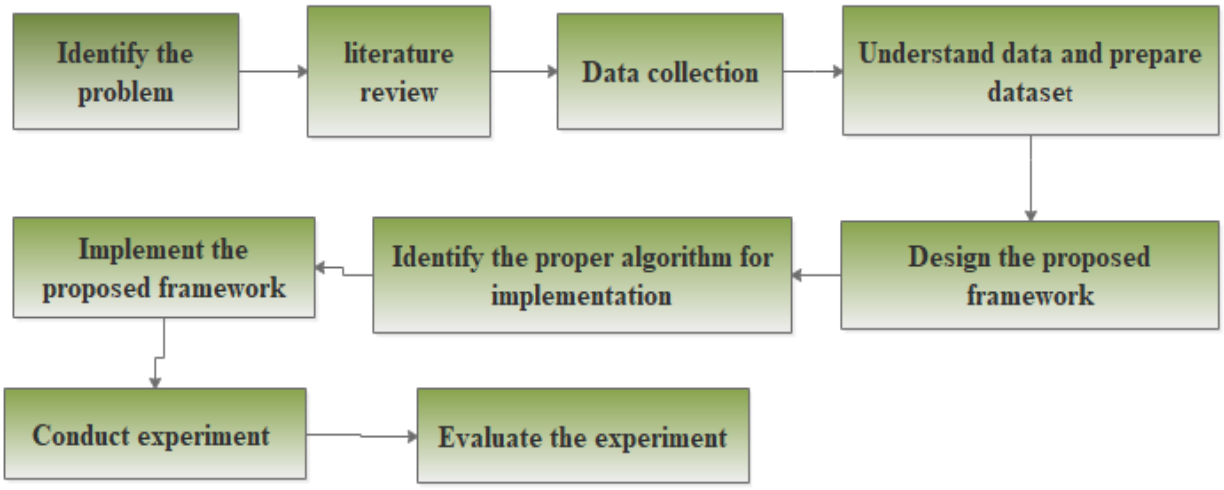


Figure 3.1 Research study workflow

3.2 Literature Review

To attain the objectives of the research, literature review on predicting pregnancy induced hypertension and related works done. Books, articles, thesis papers used to collect the necessary information for this study. Based on the information obtained from the reviewed literature, important features and their representation, algorithms and techniques and tools are selected. The techniques and tools are chosen by considering various factors like their effectiveness in similar previous works. In recent years, many works have conducted; Even if various researches on prediction of pregnancy induced hypertension, no researchers have done on pregnancy induced hypertension prediction by integrating machine learning with healthcare server.

3.3 Data Collection

Various data collection methods are available for clinical data collection such as questionnaires, record reviews of patients data, interviews of patients and interviews of experts in this disorder area[56]. For the proposed solution the selected method of data collection are record review of pregnant patients and interview of experts or gynecology in pregnant patients area because it is the most effective and efficient adaptive method of data collection.

Record Review

Data were collected from Adama Hospital, Noah specialty clinic Adama, and Asella hospital for all pregnant women who were admitted obstetrician and gynecology ward that completed their treatment and delivered or died during a period from Jan1/2018 – Jan1/2019. This three hospitals are selected based on their largest activity for women and and other issues and which are near to our university and noah speciality clininc also the health center for women and children services.

The study focus on pregnant women who are admitted in this hospitals for antenatal care follow up from the first trimester into the third trimester. Moreover, those data were collected by reviewing the document in obstetrics and gynecology ward and required data were extracted from a patient chart, registration book and anesthesia room registration by nine students. Six of them are students who were taking specialization and work in the gynecology and obstetrics department and having work experience in Adama and Asella hospital and three nurses in Noah specialty clinic who work in gynecology and obstetrics department. This department ward

checked the quality of data for incomplete and consistency and one-day training was given for them.

Interview

The interview was taken from obstetrician and gynecology wards about pregnancy-induced hypertension, what are the risk factor and parameters they used to diagnose pregnant women who have pregnancy-induced hypertension disorder.

3.3.1 Sampling Techniques and Sample Size

Purposive sampling techniques are used to collect data in Adama hospital, asella hospital and noah speciality center. The selection criteria of domain experts for understanding the features of pregnancy induced hypertension of the study are based on the profession or expertise recommendation. Data were selected based on the features recommended by obstetrician and gynecology room worker those are 14 features based on this features, records of patients which contains this features are selected and registered. The case of delivery of pregnant women and other disorder which are related to pregnancy was not included and recorded in this data collection.

Table 3.1: Selected Risk Factor of Pregnancy Induced Hypertension

No	***Variables /features***	***Description***	***Options***
1	Gravida	Pregnancy for the first time or more	Primity/multi
2	Systolic blood pressure	Blood pressure value	numeric
3	Diastolic blood pressure	Blood pressure value	numeric
4	Gestational age of onset	Age of pregnancy gestation	numeric
5	Age	Age of pregnant women	numeric
6	PHTN	Previous history of hypertension	Yes/no
7	FHTN	Family history of hypertension	Yes/no
8	HRD	History of renal disease	numeric
9	Body mass index	Weigh/height result of pregnant women	numeric
10	Smoking status	Whether pregnant women used smoking or not	Yes/no
11	History of diabetes	History of pregnant women have diabetes or not	Yes/no
12	Residence	Residence of pregnant women	Urban/rural

13	History of cardiac disease	History of pregnant women have cardiac disease or not	Yes/ no
14	proteinuria	Protein found in urine	Numeric

Using the above features ,data were collected from each three hospitals and the numbers of data from the three hospitals are shown in table 3.2

Table 3.2:Total Number of Collected Data

The Total Number of Collected Data	
Pregnant women records	***Count***
Adama hospital	3100
Asella Hospital	2015
Noah specialty Clinic (Adama)	820
Total	***5935***

3.3.2 Dataset Preparation and Description

From the collected variables, some variables are excluded based on medical experts,' and doctor's recommendations and the construction of the machine learning model was based on features or risk factors. Furthermore, socio-demographic characteristics and gynecological and obstetrics data are combined to get more predictive value based on medical experts' recommendations. Collected data contains socio-demographic variables, obstetric gynecology variable, and medical history and the dependent or diagnosis variable is hypertension disorder of pregnancy (present or absent) and a number of the class of the whole variable is shown in tables below.

Table 3.3:Socio Demographic Variables

Variables	***Description***	***Option***
Age	Age of pregnant women	Numeric value
Residence	Where to live	Urban/rular
Marital status	Status of pregnant women	Married/unmarried
Body mass index (BMI)	Weight classified by height	Numeric
Multiple pregnancies	More than one baby in the womb	numeric

Table 3.4:Obstetric Gynecology Variables

Variables	***Description***	***Option***
Gravidity	Pregnancy for the first time	Prim/multi
Gestational age	Trimester of pregnant women	Numeric
Systolic blood pressure	Blood pressure of pregnant women	Numeric
Diastolic blood pressure	Blood pressure of pregnant women	numeric

Table 3.5: Medical History Variables

Variables	***Description***	***Option***
Previous history of hypertension	History of pregnant women have hypertension or not	Yes/no
Family history of hypertension	History of pregnant women's family have hypertension or not	Yes/no
Cardiovascular disease	Whether pregnant women having cardiovascular disease	Yes/no
Renal factor	Whether pregnant women having renal disease	Yes/no
Diabetes	Whether pregnant women having diabetes disease	Yes/no
proteinuria	Protein found in urine	Numeric

3.3.3 Data Preprocessing

Data processing is the process of transforming raw data into a clean dataset before data analysis or applying machine learning algorithms commonly used as a primary activity during developing a machine learning algorithm that helps to identify the noisy, incomplete and inconsistent features. Cleaning dataset and removing outlier is the most fundamental activity for machine learning especially clinical data most of the time stored in a different format in the record and also contains much missing value.

For preprocessing pregnancy-induced hypertension dataset, primary data preprocessing method will be used, and because our dataset contains a lot of missing values, outlier or noisy data, categorical features, the same features recorded in different format, some data are recorded out of

range since machine learning cannot understand this dataset unless it is encoded in machine learning understandable format. Finally, the following methods used in data preprocessing.

Noisy Filter Method: a well known and useful data preprocessing method for removing outlier, cleaning noisy data variables by variable filter method and in this approach, a variable called as noisy using specific criteria will be removed. A new dataset without noisy and outlier can be used for feature selection and machine learning algorithms for classification[57].

Filling Missing Feature Value: there are several methods for treating and filling missing value such as filling value with the most often the value of dataset, filling missing value with the mean of features value and median value, treating the missing value as particular variables[57]. The last method is encoding categorical features using a one-hot encoding, label encoding method.

Features Normalization: feature Normalization is the process of rescaling original data without changing its behavior or nature and providing new boundary between 0 and 1[58]. This generally means data for a numerical variable are scaled in order to range between a specified set of values, such as 0–1. For example, scaling each patient's severity of illness scores to between 0 and 1 using the known range of that score in order to compare[58]. patients in a multiple regression analysis. There are different types of data normalization such that min-max normalization and standard scalar, a standard scalar is proposed in this study to normalize our dataset. The formula for standard scalar.

$$z = \frac{x - \mu}{\sigma}$$

Equation 3.1; Calculate feature normalization[59]

$\mu = 0$ and $\sigma = 1$

where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples

3.4 Algorithm Selection For Experiment

In order to apply machine learning approach in the preprocessed dataset,there are two steps are there. The first one is applying feature selection and feature extraction algorithms for understanding the features and selecting the best features,the second one is appling machine learning approach on selected and extracted features. The following two section describes this two method of algorithm application.

3.4.1 Feature Selection and Extraction Method

Feature selection and feature extraction is an important step next to preprocessing our dataset when building a machine learning algorithm for pregnancy-induced hypertension datasets. Feature selection method used to select an important and best feature and helps to know and visualize our dataset to identify which features are more predictor of pregnancy-induced hypertension and gives insight about the whole features, to understand the relationship of features, the relationship with the target features and it reduces training time of algorithm and increase the performance of the prediction. Feature extraction also used in pregnancy-induced hypertension for reducing the dimension of our dataset and helps to extract features.

We have selected algorithm for feature extraction is principal component analysis (PCA) and Feature Extraction with Recursive Feature Elimination (RFE), for feature selection is univariate feature selection (UFS) and feature importance with extra tree classifier, this feature selection and extraction algorithms are used for modeling pregnancy-induced hypertension and they are selected based on their popularity in clinical feature selection and extraction process[60].

Feature Selection Using Univariate Feature Selection (UFS); **-** is a popular**,** the fastest and most simplistic selection method used in clinical data analysis. Moreover, univariate feature selection helps to determine the strength of the relationship between each independent feature and the target feature using statistical tests those are chi-square, F-test, mutual information. These approaches rank features according to some score and based on the strength of their relationship with the outcome, and the user selects the best k features accordingly[49].

Feature Extraction Using Principal Component Analysis (PCA):- Is a popular feature extraction method used in clinical data analysis, which creates a linear transformation of input feature and the new variable is the projection of original variable to new variable space which is called principal component analysis[61]. This method is used for dimensionality reduction of features and variables are reduced based on their Eigenvalues and select variable with more significant variance [40][61] PCA, first, normalize data to zero mean then computed covariance matrix which measures joint variability across variables in dataset and contains covariance score for every variable with another variable including itself then lastly finds eigenvectors of covariance and original data was multiplied with Eigenvector provide transformed data.

Formula for PCA, supposing that **x** is an eigenvector of the covariance matrix of PCA, we know that the feature extraction result, with respect to x, of an arbitrary sample vector a is

$$Z = a^T x = \sum_{i=1}^{N} a_i x_i$$

Equation 3.2:Calculate PCA[61]

where $x = [x_{1..}....x_n]^T$, where x is eigen vector

$a = [a_{1.....}a_n]^T$, and a is original data

N is the dimensionality of sample vectors[40][61][17].

Feature Extraction with Recursive Feature Elimination (RFE): is a popular feature extraction method for clinical data analysis and Uses cross-validation to score feature subset and selects the best scoring collection of features and features are ranked by model's coefficient or feature importance. Recursive feature elimination (RFE) selects features by recursively considering smaller and smaller sets of features until the desired number of features to select is eventually reached[60].

3.4.2 Machine Learning Classification Algorithms method

The objective of this study is to predict pregnancy-induced hypertension using a machine learning algorithm on our prepared datasets. We use a multiple supervised machine learning algorithm for comparison and accuracy of the prediction. The algorithm is selected because of their excellent performance for the classification problem, and then the popularity for solving clinical data prediction problems on previous research work for diabetes prediction, chronic

kidney disease, hepatitis, heart disease prediction, cardiovascular disease[62][63][45][9][64][65]. The following three machine learning algorithms used to build a prediction model.

Support Vector Machine*:* - This is an algorithm that allows for the classification of both linear and non-linear data and applies a non-linear mapping method for transforming training data into a higher dimension. Support vector machine have hyperplane which is a kind of line that separate input variable space into support vector machine either zero or one, therefore, each input variable can be separated by hyperplane line and also the distance between hyperplane and adjacent data coordinates is called margin, so the line which has largest margin have optimal hyper plane which is called support vector[64].

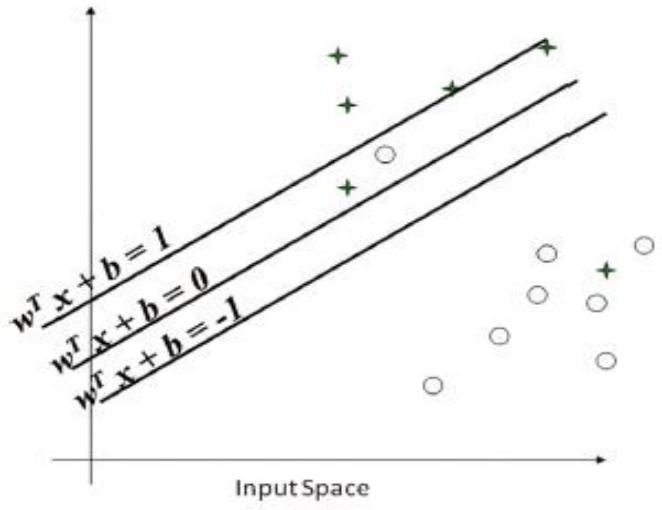


Figure 3.2: Representation of Support Vector Machine[64]

Naive Bayes Classifier;– Naive Bayes classification method is based on Bayes' theorem. It is termed as 'Naive' because it assumes independence between every pair of features in the data and it is called a powerful independent variable. This algorithm considers each feature contributes independently and used for a larger dataset and used for classification problems. [66][10][32]

Bayes theorem:

$$P(M/N) = \frac{(P(N/M)*P(M))}{P(N)}$$

Equation 3.3: Bayes Theorem[32]

Where, P(M|N) = posterior probability this means the probability of hypothesis M given data N

P(N|M) = probability of data N given that the hypothesis M was true

P(M) = probability of hypothesis M being true

P(N) = the probability of the data N

Random Forest Classifier: **-** Is a large collection of decor related decision tree using a bagging technique and creates sub-training sets from the super training set and decision tree is constructed from each sub-training set. During testing, each input vector of a test set is classified by all decision tree in a forest and the forest chooses the classification result using majority votes or average.

3.5 Evaluation of The Experiment

Evaluation of this study has been conducted on multiple phases of the development of the prediction model for pregnancy-induced hypertension. After all, the evaluation conducted the results discussed and concluded by suggesting a model predict pregnancy-induced hypertension.

3.5.1 Prediction Model Evaluation

To evaluate our prediction model, we will use k-fold cross-validation. K- Fold cross-validation helps to identify how well our model will perform when we give a new dataset. We choose a 10-fold cross-validation technique. 10 –fold cross validation techniques divides dataset into ten subsets and one subset used as a test set and the left nine subsets used as a training set. This method used to reduce the problem of overfitting and the problem of underfitting and the other advantages of this validation technique is that it avoids bias and variance problems. Furthermore, K- fold cross-validation avoids overlapping of test sets. It Is always recommended to use 10 fold cross validation techniques[67].

3.5.2 Prediction Model Performance Evaluation

There are different performance Metrics in generally used in medical data classification and diagnosis to analyze the performance of prediction models like sensitivity(recall), specificity (precision), F-measure, accuracy, and confusion matrix. That performance metrics will be used during evaluating our classification or prediction model because they are used generally in all medical data classification model performance

Terms used in prediction performance measure Metrics

- ***True Positives (TP):*** we call this one when our classes are true and also the predicted class also correct or we correctly predicted that they have pregnancy-induced hypertension (PIH)

- ***True Negatives (TN):*** we call this one our actual class is zero and the predicted class also zero, or we correctly predicted that they do not have pregnancy-induced hypertension (no PIH).

- ***False Positives (FP):*** we call this one our actual class is false or zero, and the predicted is 1 or true. False is because the model has predicted and positive because the class predicted was a positive one. (1) or we incorrectly predicted that they do have PIH.

- ***False Negatives (FN):*** we call this one our actual class is True and the predicted is False. False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0) we incorrectly predicted that they do not have PIH

***Average Accuracy*:** Is the result of tests that are predicted among all classes it indicates that our classifier is often correct in predicting of our pregnant patients is having pregnancy-induced hypertension or not.[68]

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Equation 3.4: Accuracy Measure[68]

Confusion Matrix; - Is one of the famous and most accessible Metrics that provides the summary representation of the predicted classes. It provides a summary of true positive, true negative, false positive, and false negative. Furthermore, the confusion matrix used for classification problems of classes having two or more types and it is not a performance measure by itself, but all performance metrics depend on the confusion matrix.[40]

Table 3.6 Sample Format of Confusion Matrix

		Predicted Value		
		No PIH	**GH**	**Preeclam**
Actual Value	**No PIH**	**True no PIH**	**False GH**	**False Preeclam**
	GH	**False no PIH**	**True GH**	**False Preeclam**
	Preeclam	**False no PIH**	**False GH**	**True Preeclam**

Recall (Sensitivity/TP); **-** Is the probability of the test predicting positive cases from the positive cases of respective class and provides us classifier's performance concerning a false negative. Recall concentrate on containing all cases that have "PIH" with the answer as "PIH".

$$Recall = TP/ (TP+TN)$$

Equation 3.5: Recall Calculation For Model[68]

Precision (Specificity /TN); **-** is the probability of the test correctly predicted as positive cases given that the number of cases as positive and precision (specificity provided information about its performance concerning false positives and concentrated on how many did the model caught.

$$Precision = TN/ (TN+FP)$$

Equation 3.6: Precision Calculation For Model[68]

F-measure (F1 Score); - Is the mean of precision and recall, and its result range score is 0 and 1. F1 score tells how many instances are predicted correctly and does not miss an important number of instances. We call our model performance is better when the F1 score is greater or near to 1 and also F1 score tries to find the balance between recall and precision.

$$F_1 = 2.(P.r /p+r)$$

Equation 3.7: F-Measure for Model[68]

Where P is precision and R is a recall

3.6 Integrating the prediction Model With Healthcare Server

The main aim of this study is to develop a framework for pregnancy-induced hypertension using machine learning algorithms, so we have discussed in the previous section about developing machine learning algorithms method because the first activity to develop this framework is preparing the machine learning model and saving the model with best accuracy value. Then after a model is developed and saved, the next step is bringing this model to health care server for deployment through this way model are accessed by professionals for diagnosis and patients for early identification and awareness of pregnancy-induced hypertension.

Python provides several choices of frameworks for web development and other applications such as Django[69], Turbo Gears, Pyramid, Flask, Bottle, Cherry preferences.[69]. For deploy machine learning model into healthcare server, we will use python tools this will provides library to integrate machine learning model with flask server, In this study, we have selected flask framework for integrating the proposed model into healthcare server and Flask Microframework is the most popular and the most extensible and straightforward framework for serving as a server and deploying machine learning model.

Why Flask Server?

Flask Server is a popular and the extreme of python web developers which allows complete flexibility of the user. Flask server helps the developer to control every activity helps the user to add additional functionality like connecting with the database, handle forms, security, and authentication. This makes flask an excellent learning tool if we are new to web development[70][71]**.**

3.7 Tools to Be Used

Design tool: **-** The Edraw Max tool is a type of design tool which enables the software designer to reliably create and publish many kinds of diagrams to represent an idea. Edraw max Is an all

in one diagram software used to create a professional-looking flowchart, organizational chart, web design and more and Edraw max is the best tool to map what we know, includes high-quality shape, gain higher productivity in diagramming with features[72].

Implementation tool; - Anaconda Navigator used for building python machine learning model. Python has large standard libraries and it is easy to develop web services. A web application was built using a flask web server, which is a library of Anaconda Navigator, and also python library is used for performing machine learning models in the server.

Why python?

Python is prioritized the most by those for whom data science is the first profession or field of the study. Python is a general-purpose language which is simple to read and write and it is easier to work with for non-programmer or developer. Furthermore, because it's considered truly universal and used to meet various development needs, it's a language that offers a lot of options to programmers in general. The language is used for system operations, web development, server and administrative tools, deployment, scientific modeling and much more[73]. python rank the first language used by data scientist as described in figure 3.2.

Rank	Change	Language	Share	Trend
1		Python	28.73 %	+4.5 %
2		Java	20.0 %	-2.1 %
3		Javascript	8.35 %	-0.1 %
4		C#	7.43 %	-0.5 %
5		PHP	6.83 %	-1.0 %
6		C/C++	5.87 %	-0.3 %
7		R	3.92 %	-0.2 %
8		Objective-C	2.7 %	-0.6 %
9		Swift	2.41 %	-0.3 %
10		Matlab	1.87 %	-0.3 %
11	↑	TypeScript	1.76 %	+0.2 %
12	↓	Ruby	1.44 %	-0.2 %
13	↑↑↑	Kotlin	1.43 %	+0.4 %
14	↓	VBA	1.41 %	-0.0 %
15	↑↑	Go	1.21 %	+0.3 %

Figure 3.3: Python Rank [74]

CHAPTER FOUR

4 Proposed Framework for PIH Prediction

4.1 Overview

This chapter discusses the proposed Framework for Pregnancy Induced Hypertension Prediction by building pregnant women patients data set and using machine learning techniques to solve the problem for further understanding and usage of the model then deploying this machine learning model into healthcare server which will helps professional to diagnose their patients and helps pregnant women for early identification of their health status of pregnancy-induced hypertension.

The study collected data from Adama hospital, Asella hospital and Noah specialty clinic (Adama) then based on collected data with the help of obstetrician gynecology, we have prepared the dataset for prediction of the problem and we have described the method we used for data collection and preparation in chapter three. This chapter will discuss the overall framework of the proposed framework for pregnancy-induced hypertension using machine learning algorithms and will discuss each subcomponent of our proposed framework.

4.2 The Proposed Framework for PIH Prediction

The proposed framework used to predict pregnant women weather they have pregnancy-induced hypertension or not. The proposed framework described in fig 4.1 takes a dataset as input, preprocessing dataset using preprocessing techniques such that cleaning irrelevant dataset, by transforming raw dataset to machine-understandable format, and also by normalizing to a standard format. After all dataset pass from preprocessing steps, preprocessed features of data set are selected and extracted using feature extraction and feature selection techniques such as principal component analysis(PCA) which is an algorithm for dimensionality reduction, recursive feature elimination(RFE) and univariate feature selection(UFS), After features are extracted and selected, features are split to training and testing set then machine learning modeling is developed using SVM, RF and NB algorithms. The model evaluated using cross-validation, precision, recall, confusion matrix method. Moreover, models with the best accuracy and sensitivity will be selected. The outcome of this model is pregnancy-induced

hypertension(Gestational Hypertension(GH) or Preeclampsia (preeclam) or no pregnancy-induced hypertension.

For building machine learning models, the Python programming language is widely used as it has compelling data science libraries. Flask is Open source Python distribution package, and Anaconda consists of various IDEs such as Spyder and Jupiter, I Python Notebook that is used in our model for building and testing the machine learning codes additionally, anaconda navigator has flask framework which will allow us to integrate machine learning model and healthcare system which is flask server. Flask It is a python based micro-framework with an inbuilt light-weighted web server that needs minimal configuration and can be controlled from python code.

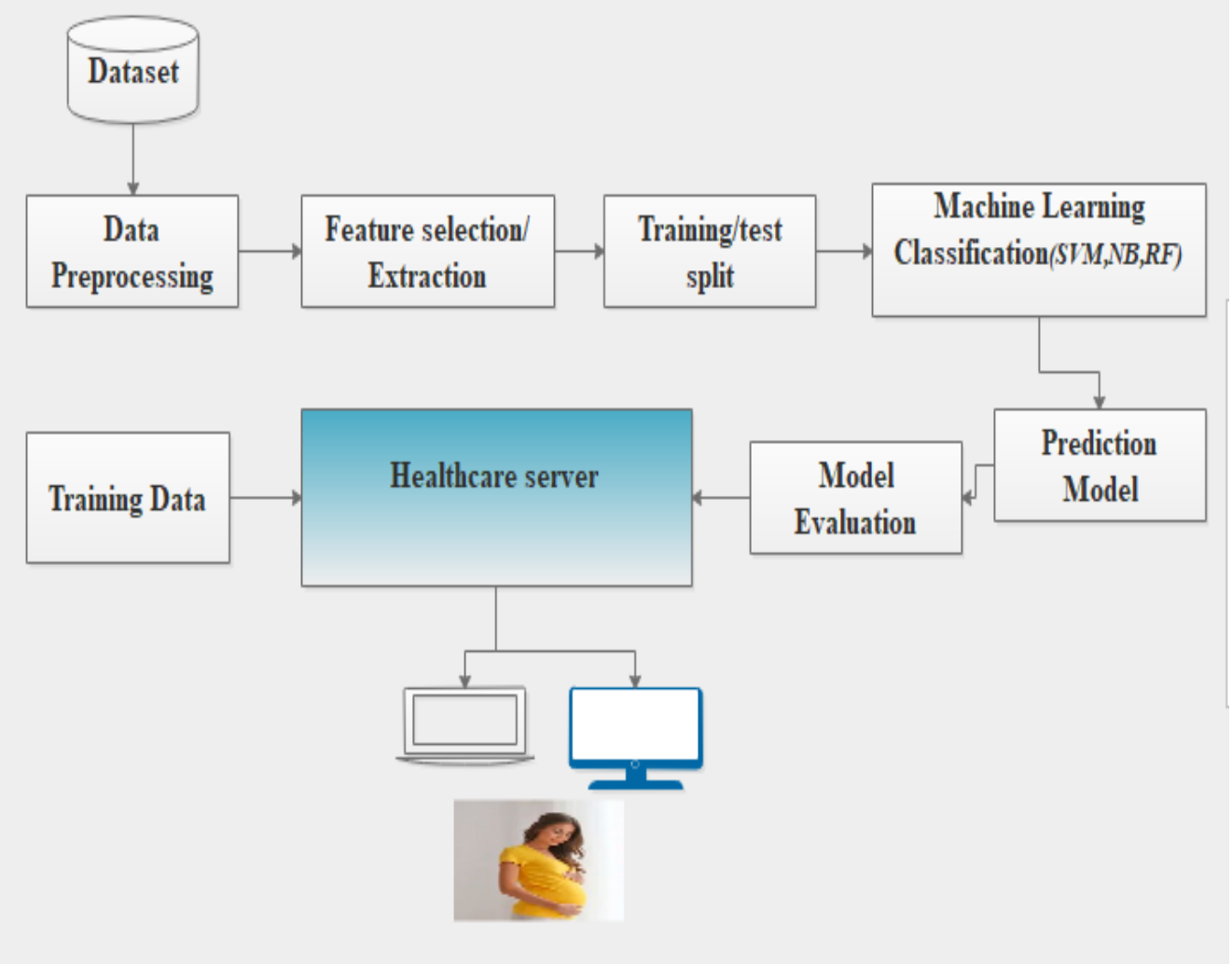


Figure 4.1: Proposed Framework for PIH Prediction Using Machine Learning[75]

4.3 Proposed Data preprocessing

Data preprocessing based on pregnancy-induced hypertension prediction is a critical subcomponent stage. Since the collected data are imperfect, containing inconsistencies and redundancies, this is not directly applicable for a starting machine learning model. Data preprocessing method is proposed for cleaning our dataset because the collected data set contains a lot of inconsistent data and noisy data this will be removed using noisy filter method which is described in chapter three, real data recorder the same feature or variables in different format and also machine learning cannot directly understand data which is recorded in human understandable format this will be treated using feature encoding method. Additionally, the data normalization method is proposed for rescaling the dataset in a standard format.

The other format of raw data records have different category and some are written or recorded in text and the other are numbers so this type of data also needs to be cleaned and converted in machine-understandable format mostly machine learning algorithms understand records in number format and can be converted to this format by using data preprocessing technique such that label encoder used to encode categorical features to label and numeric and also categorical data are different format for example in our dataset nominal categorical data that means features which have values like two categories such as yes/no, therefore, this feature value encoded by one-hot encoder used to encode or convert nominal categorical features to numeric and machine-understandable format.

4.4 Proposed Feature Extraction and Selection Method

Various learning algorithms work efficiently and give more accurate results if the data contains more significant and non-redundant attributes. As the medical datasets contain a large number of redundant & irrelevant features, an efficient feature selection technique is needed to extract exciting features relevant to the disease.

There is many feature selection methods applied in a clinical dataset to identify features as we have reviewed in the literature review and before applying the machine learning model, extracting and selecting essential features from many features in the dataset helps to get more accurate results in less time. In other cases reducing the dimensionality of a dataset, filtering important features or identifying essential features from the dataset helps to reduce complexity.

If a dataset contains more significant and non-redundant attributes, machine learning algorithms work efficiently and provides accurate results. Therefore, our dataset contains a lot of redundant and irrelevant features, it needs an efficient feature selection method that is selected in this study and those are shown in fig.4.2.

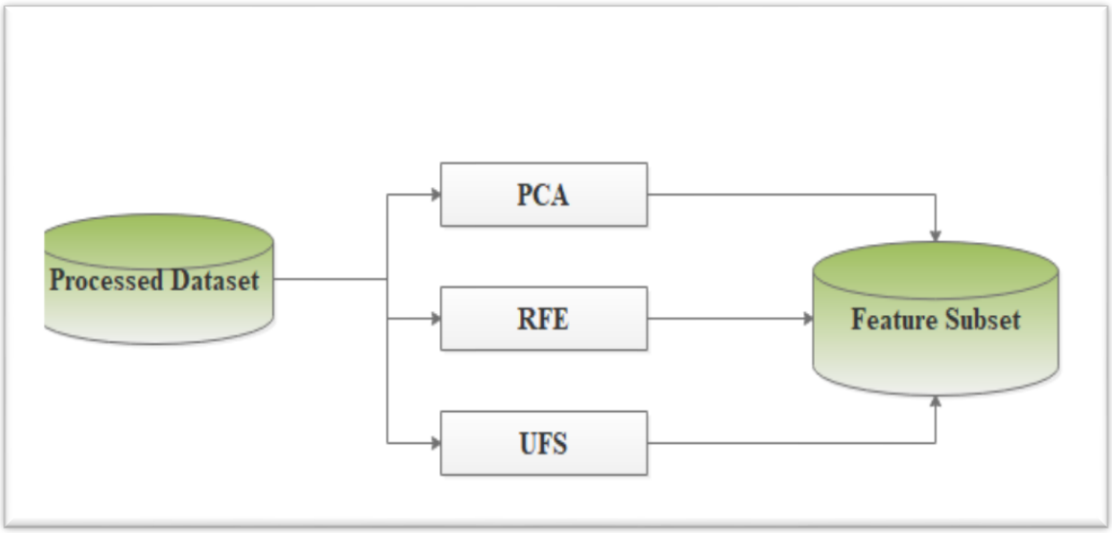


Figure 4.2: Proposed Feature Extraction and Feature Selection Flow

Univariate feature selection is proposed and this approach used to identify the relationship between an independent variable with the target or dependent class using chi-square, which can compute chi-squared statistics between non-negative independent variables and class. Based on this method, the collected features are ranked based on their chi-square result; from this result we select the best k features, and we can apply machine learning algorithms for classification.

Principal Component Analysis (PCA) is called the unsupervised feature selection method and this approach used to identify the relationship among independent features and used for dimensionality reduction of features based on their Eigenvalue. To compute covariance matrix data are normalized to zero measured joint variability across variable and we will get the covariance score for all independent variables with each other and we will find eigenvectors of covariance and multiplied with an original variable to get transformed data. Through this approach, we can identify variables to be reduced.

Uses cross-validation to score feature subset and selects the best scoring collection of features and features is ranked by model's coefficient or feature importance. Recursive feature

elimination (RFE) selects features by recursively considering smaller and smaller sets of features until the desired number of features to select eventually reached.

4.5 Training and Testing Split

After all, features are identified and selected, those datasets are split into train and test dataset, machine learning model, works in train and test dataset and there is no fixed rule for the percentage of train and test split but most frequently recommended based on our data set such as 80/20 % or 70/30%.

4.6 Machine Learning Model Building

In these subcomponents of the proposed framework for pregnancy-induced hypertension, prediction performs machine learning classifier and training on all feature subset constructed by the feature extraction and selection components methods and on all features. This study builds a classifications model mainly by using a machine learning classifier algorithm such as SVM, RF, and NB on the dataset features which could be selected for the prediction of pregnancy-induced hypertension (GH and Preeclam) or No pregnancy-induced hypertension (no PIH). The process of modeling is to predict the training data that have GH, Preeclam and who do not have PIH. The output of these modeling is a trained model that can be used for predicting pregnancy-induced hypertension, making predictions on new user data input.

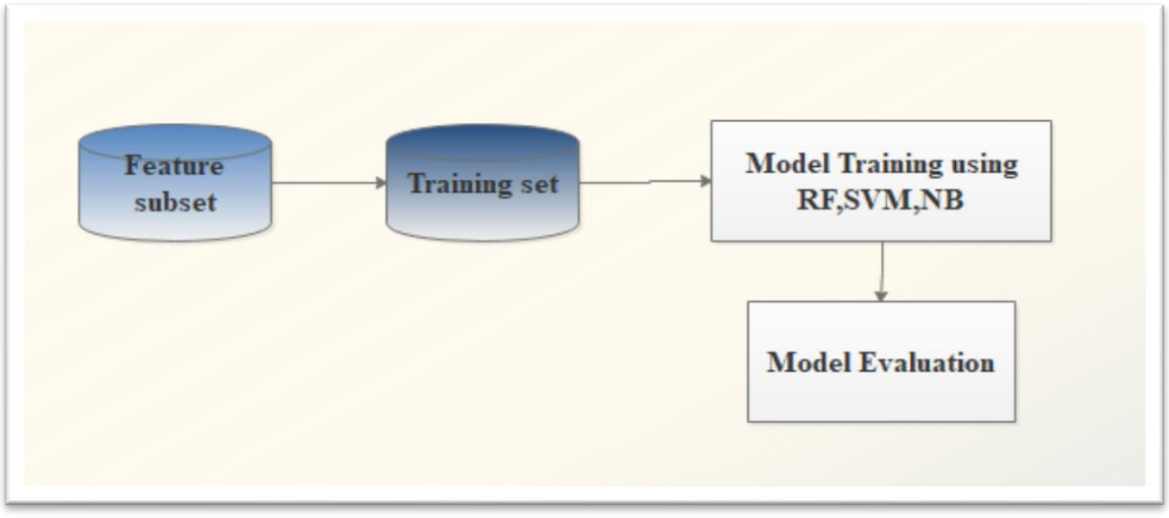


Figure 4.3: Machine Learning Modeling

This study proposed a Support Vector Machine (SVM) machine learning classifier based on a linear kernel algorithm. Linear classifiers use a linear function to score classes by taking the dot product of feature values and feature weights computed during training. And it is One of the most powerful, flexible and ubiquitous learning algorithms,

The proposed classifier, Naïve Bays (NB) classifier is a probabilistic machine learning model that's used for classification. This study uses Gaussian based NB for modeling. NB classifier is a specific instance of the classifier which uses multiple distributions for each of the features in the dataset. The proposed Random Forest (RF) classifier is a meta estimator that fits several decision tree classifiers on a subsample of the dataset. It means RF consists of a large number of individual decision trees that operate as an ensemble by bagging and feature randomness.

4.7 Model Evaluation and Testing

In order to test and estimate the learning ability of machine learning models trained on the pregnancy-induced hypertension dataset. The fundamental concern of the machine learning model is to find an accurate estimation of the generalization error of the trained models on finite datasets. Because of this reason this study used proposed K-fold cross-validation (CV) and different performance evaluation metrics appropriate for models, these metrics are confusion matrix, accuracy, precision, recall, and F-Measure, as discussed in chapter three.

***Pickle:* -** after a model is evaluated and recommended with the best accuracy and sensitivity, the next step is saving the selected model for deploying on the server; when new data is input to the server, the method for saving this model is called pickle. Pickle is a method of serializing and it is a method to write python object to a disk[76]. and de-serialized helps to read serialized model by a python [76]. After the model is trained, it will be saved as a pickle for the reading trained models during the flask server call it.

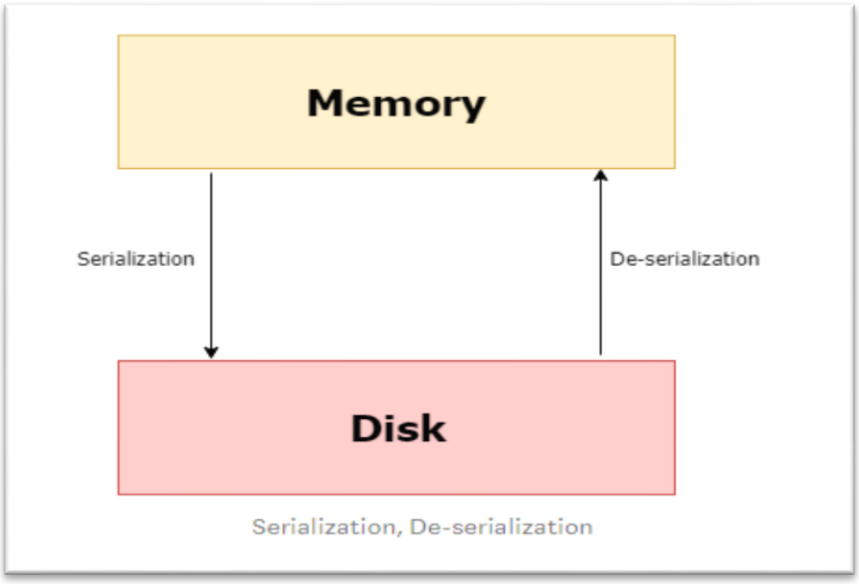


Figure 4.4: Saving Trained Model[76]

4.8 Integrating prediction Model with healthcare server

Machine learning model with the best accuracy is deployed in the healthcare server and this will allow users to get services from machine learning prediction of pregnancy-induced hypertension when they input features using HTML pages that are connected with the system and machine learning model. To access service from the system, Html page is developed and connected to the system, therefore through this page user can input their data and get prediction results from a model. The reason we have selected the flask server for deploying our machine learning model is that, among the server type discussed in chapter three, is flask server is it has built-in development server and performs integrated unit tests and extensively documented. Therefore, we have selected and proposed a flask server for deploying the machine learning model.

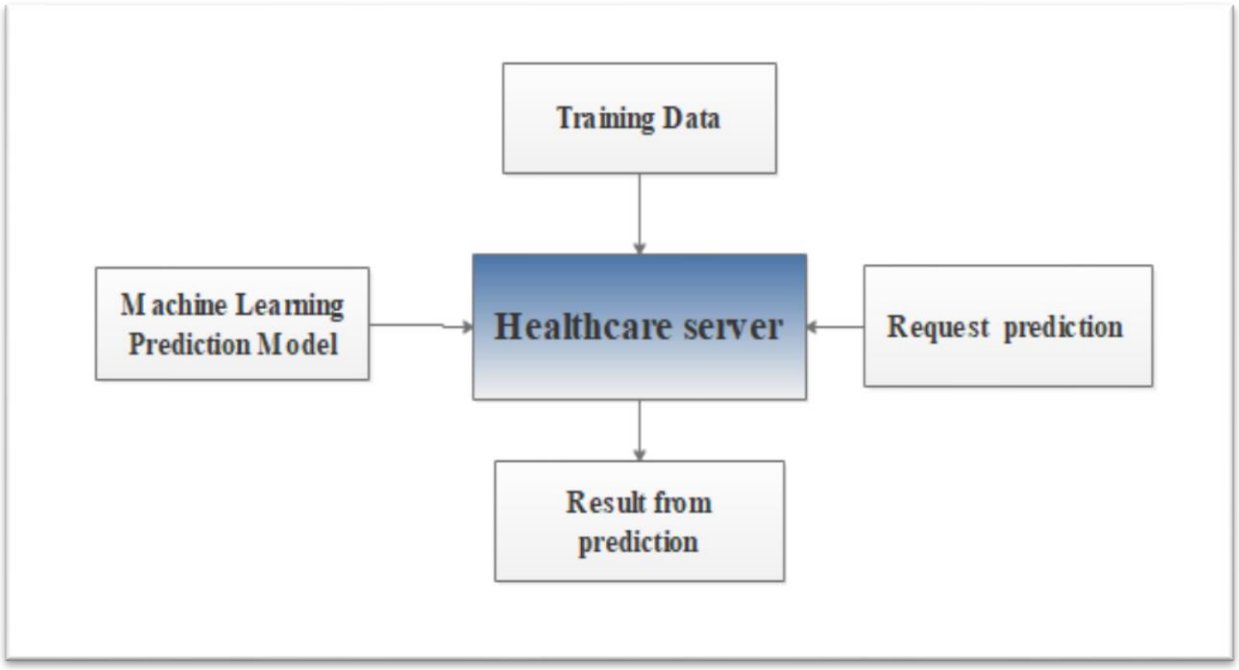


Figure 4.5: Integrating Machine Learning Model with Healthcare Server

CHAPTER FIVE

5 Implementation and Experiment

This chapter describes an implementation of a proposed framework for pregnancy-induced hypertension using prepared dataset and follows an experimental approach to determine the best combination of the machine learning algorithm and features selection for final models' comparison is implemented, and finally, best algorithm with high accuracy will be integrated with flask server for user to access the prediction model.

5.1. Implementation Environment

Table 5.1: Implementation Environment

Tools	
Tools	***Description***
Anaconda Navigator	It allows us to launch development applications and efficiently manage conda packages, environments, and channels without the need to use command-line commands.
Jupyter notebook	An open-source web application that allows us to create and share documents that contain live code, equations, visualizations, and narrative text. Uses include data cleaning and transformation, numerical simulation
Python	powerful programming language to develop a machine learning application and flask server
Microsoft excel	It's used for data preparation tasks in cleaning, filtering and sorting the gathered data.
Python packages	
Scikit learn	It is a python module for machine learning and data mining. This study uses it for feature extraction/selection and training and testing models.
Pandas	It is a data analysis tool and easy to use data structure. This study uses it for data reading, manipulation, writing and handling the data frame.

NumPy		Array processing for number, strings, and objects. This study uses it for handling converting the text to numeric data for features and training and testing the model.
Matplotlib		Publication quality figures in python. This study uses it for data and results visualization.
Flask server		Extensible framework for serving as a server and deploying machine learning model

5.2. Deployment Environment

The tools which is discussed above in section 5.1 have been deployed on a personal computer equipped with a processor Intel® Core™ i5-4310M, CPU 2.70GHz, 2 Core(s), 8 Gigabyte of physical memory, 465 Gigabyte hard disk storage capacities. The operating system is Windows 10 Pro, 64 bits.

5.3. Dataset Description

In order to build a dataset for this study, from Jan.1/2018 to Jan 1/2019, we collected pregnant data from Adama hospital, Asella, Noah Especiality clinic and recorded in an excel sheet. After storing in excel datasheet, the total data we collected is 5935. Moreover, the total filtered dataset using the data preprocessing technique is 3062 and the exact method for building the dataset is discussed in chapter four. The dataset contains 14 columns and the last column is the class of each unique record of the patients or individuals.

Table 5.2: Dataset Description

Column name	***Description***
Systolic blood pressure (SBP)	Systolic blood pressure in mm/hg
Diastolic blood pressure (DBP)	Diastolic blood pressure in mm/hg
Gestational age of onset (GAO)	Age of pregnancy or months in the number
Age	Age of pregnant women in the number
Proteurinea	Protein in urine(yes/no)
Residence	Where they live (urban/rural)
Body mass index (BMI)	Weigh divided by height in number (kg/.)

Gravida	Pregnancy for the first time or more(primi/multi)
Previous history of hypertension (PHHTN)	History of pregnant women with hypertension (yes/no)
Family history of hypertension (FHHTN)	History of a family with hypertension with (yes /no)
History of diabetes (HD)	History of pregnant women with diabetes with (yes /no)
History of cardiac disease (HCD)	History of pregnant women with cardiac disease with (yes /no)
Smoking habit	Whether pregnant women smoking (yes /no)
Class	The Class of PIH (GH, Preeclam or no PIH)

5.4. Implementation of Data Preprocessing

To implement data preprocessing method, the study used python programming language and modules. In order to perform the full implementation of the methods, we perform the following common activities, which are importing important libraries packages and loading the dataset file to the disk using the code in figure5.1. The loaded dataset using panda is used throughout the whole implementation of this study.

```
import numpy as np
import pandas as pd
#Load data into pandas dataframe**
df = pd.read_csv("PIHDataM.csv")
```

Figure 5.1: Python Code for Leading Dataset

After a dataset is loaded data preprocessing method is applied in a loaded dataset for the cleaning dataset, treating missing value using mean(), mode() and most frequent() method. Furthermore, the other is encoding categorical features into machine learning understandable format because the machine learning model understands numerical variables.

5.4.1. Implementation of Cleaning Data

In order to clean and preprocess the dataset, we implement a different method that cleans and remove irrelevant features, filling missing variables or features and also converting or encoding categorical features. To implement this method, a python package contains *preprocessing ()* modules and this package can identify unclean data and removes unwanted data, and also missing value replaced by mean (), median () or mode () method and categorical data encoded using label encoder() and one hot encoder() then cleaned, filled and encoded dataset will be returned. The code for cleaning dataset, filling missing value and encoding categorical value is shown in Appendix B.1.

5.4.2. Implementation of Data Normalization

In this stage all cleaned and encoded dataset is sent to be normalized or rescaled to 0 and 1 ranges using preprocessing () module such as standard scaler () method then normalized and transformed dataset is displayed for the next step. The code for PIH dataset normalization is shown in appendix B.2.

5.5. Feature Selection and Extraction Implementation

To extract and select features from the dataset that is used for training the model. We used the python Scikit-learn module for UFS, PCA, RFE and feature importance using extra tree classifiers in order to implement the feature extractor method. After the dataset was cleaned and normalized using the preprocessing steps, feature selection and extraction method is applied to identify the important feature and extract features for our prediction model.

5.5.1. Implementation of Principal Component Analysis (PCA)

To implement the PCA features extraction method; this study used decomposition of PCA sklearn packages. This method accepts independent components using the PCA () method, then fit independent variables after the whole features are fitted multiplied with variance-ratio and displayer the extracted features.

```
# principal component Analysis(PCA)
from sklearn.decomposition import PCA
names = ['SBP', 'DBP', 'GAO', 'Gravida', 'Age', 'BMI', 'PHHTN', 'FHHTN','Residence','HD','HCD','HRD','smoking' 'Class']
array = df.values
X = array[:,0:13]
Y = array[:,13]
# feature extraction
pca = PCA(n_components=3)
fit = pca.fit(X)
# summarize components
print("Explained Variance: %s" % fit.explained_variance_ratio_)
print(fit.components_)
```

Figure 5.2: Sample code for Principal component analysis

5.5.2. Implementation of Recursive Feature Elimination(RFE)

To implement recursive feature elimination, we used RFE library from sklearn packages and applied logistic regression. In this method, logistic regression with a linear model is applied to the full features and the RFE method displays the selected features based on their rank.

```
# Feature Extraction with RFE(Recursive Feature Elimination)
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
names = ['SBP', 'DBP', 'GAO', 'Gravida', 'Age', 'BMI', 'PHHTN', 'FHHTN','Residence','HD','HCD','HRD','smoking' 'Cla
array = df.values
X = array[:,0:13]
Y = array[:,13]
# feature extraction
model = LogisticRegression(solver='lbfgs')
rfe = RFE(model, 8)
fit = rfe.fit(X,y)
print("Num Features: %d" % fit.n_features_)
print("Selected Features: %s" % fit.support_)
print("Feature Ranking: %s" % fit.ranking_)
# Number of best features
```

Figure 5.3: Sample Code For Recursive Feature Elimination

5.5.3. Implementation of Univariate Feature Selection(UFS)

In order to select best features from the whole column or features, we implemented univariate feature selection with chi-square () method, this method calculates the relationship between independent features from the whole columns with the target-dependent features and gives a score, based on this score it puts in order from high score to low score.

```
# univariate feature selections
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
test = SelectKBest(score_func=chi2, k=4)
fit = test.fit(X, y)
names = ['SBP', 'DBP', 'GAO', 'Gravida', 'Age', 'BMI', 'PHHTN', 'FHHTN','Residence','HD','HCD','HRD','smoking' 'Class']
array = df.values
X = array[:,0:13]
Y = array[:,13]
# summarize scores
numpy.set_printoptions(precision=5)
print(fit.scores_)
features = fit.transform(X)
# summarize selected features
print(features[0:8,:])
```

Figure 5.4: Sample Code For Univariate Feature Selection

5.6. Machine Learning Model Implementation

To build machine learning models using the proposed algorithm, we used a python sci-kit learning library package and followed the conventional method that is importing the important library package for modeling and metrics for model evaluation. This package imported from sklearn library into Jupiter notebook, which allows us to develop an interactive and machine learning model and to develop evaluation Metrics of model performance.

```
:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy as scp
import pickle
import joblib
from sklearn.ensemble import RandomForestClassifier
from sklearn import svm
from sklearn.naive_bayes import GaussianNB
```

Figure 5.5: Importing important packages for modeling

5.7. Model Evaluation And Testing

The training that is performed is validated using the K-fold CV technique implement using the sklearn *cross_val_score ()* methods using K value five. These methods use the models and dataset to validates the learning skill of each model. We use a 5-fold CV, which means that the dataset is divided into five different sets and four set use to train and one set used to test models each 1 to k iteration.

```
# cross validation for classifier model
from sklearn.model_selection import cross_val_score
scores = cross_val_score(Classifier, X_train, y_train, cv=5) #use 5 folds
print(scores)
```

Figure 5.6: Sample Code For Cross-Validation Of Model

The study used a *classification_report (), accuracy_ score ()* and *confusion_ matrix ()* methods, which are used to evaluate the performance of the built models using predict the value of 5-fold CV for the entire dataset. These methods give a result of evaluation metrics such as accuracy, precision, recall, and F1-score value of the model under testing.

Finally**,** we have saved our trained model, to save our model, we implemented pickle dump () method, which is accepting the trained model and saving those models and helps to read back from saved disk.

5.8. Integrating Prediction Model With Healthcare Server

To implement a machine learning model or prediction model into a server, we selected and used the flask server. To deploy the model in the flask server, we used python package and flask server and dependencies such as werkzueg, jinja, MarkupSafe, and click are installed through command prompt of pip install flask method(). Werkzueg implements WSGI, the standard Python interface between applications and servers and it is a utility library, Jinja is a template language that renders the pages. our application serves; MarkupSafe comes with Jinja. It escapes untrusted input when rendering templates to avoid injection attacks, Click is a framework for writing command-line applications. It provides the flask command and allows adding custom management commands.

To integrate the machine learning model with flask, the flask is imported from python library and the trained data also imported into python and also the necessary library for template render template(), for pickle() for loading training data is imported in python.

```
from flask import Flask, render_template, url_for, request
from sklearn.externals import joblib
import os
import numpy as np
import pickle

app = Flask(__name__, static_folder='static')

@app.route("/")
def index():
    return render_template('home.html')


@app.route('/result', methods=['POST', 'GET'])
def result():

scaler_path = os.path.join(os.path.dirname(__file__), 'models/scaler.pkl')
scaler = None
with open(scaler_path, 'rb') as f:
    scaler = pickle.load(f)

x = scaler.transform(x)

model_path = os.path.join(os.path.dirname(__file__), 'models/rfc.sav')
clf = joblib.load(model_path)
```

Figure 5.7; Integration of Model and Flask Server

CHAPTER SIX

6 Experiment Evaluation, Result, And Discussion

In this chapter, the result of the experiments of the proposed framework for pregnancy-induced hypertension prediction using machine learning approach is discussed. Here exploratory data analysis result is described to understand the dataset of all features and class distribution of dataset, feature selection and extraction result is described, model classification and performance evaluation of result and finally a framework for integrating machine learning model into the server and accessing the model from the server also described.

6.1 Class Distribution of Dataset

The total size of the dataset is 3062 after preprocessing the whole dataset by removing outlier, redundant and irrelevant data using data preprocessing techniques from the dataset. Out of this dataset, 1726 patients have no pregnancy-induced hypertension, 1132 of the patients with gestational hypertension (GH) and 204 of them are patients with preeclampsia (Preeclam) as described in the figure6.1, result from visualization by the histogram in python tool.

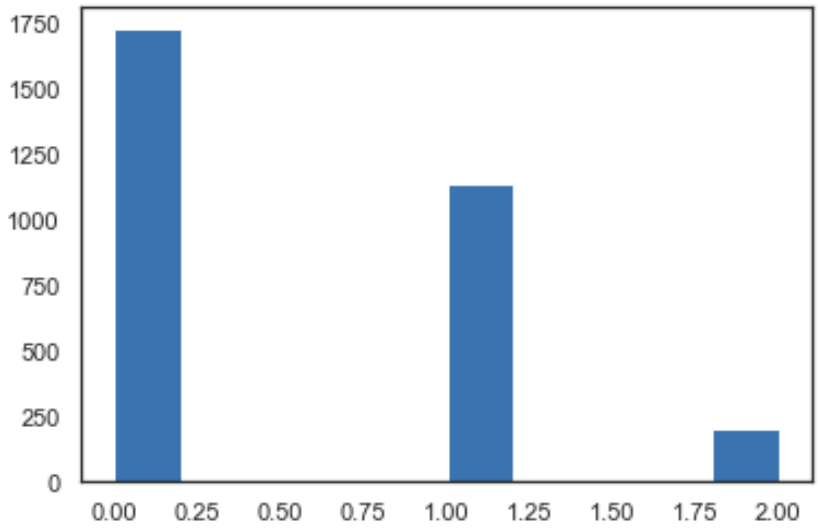


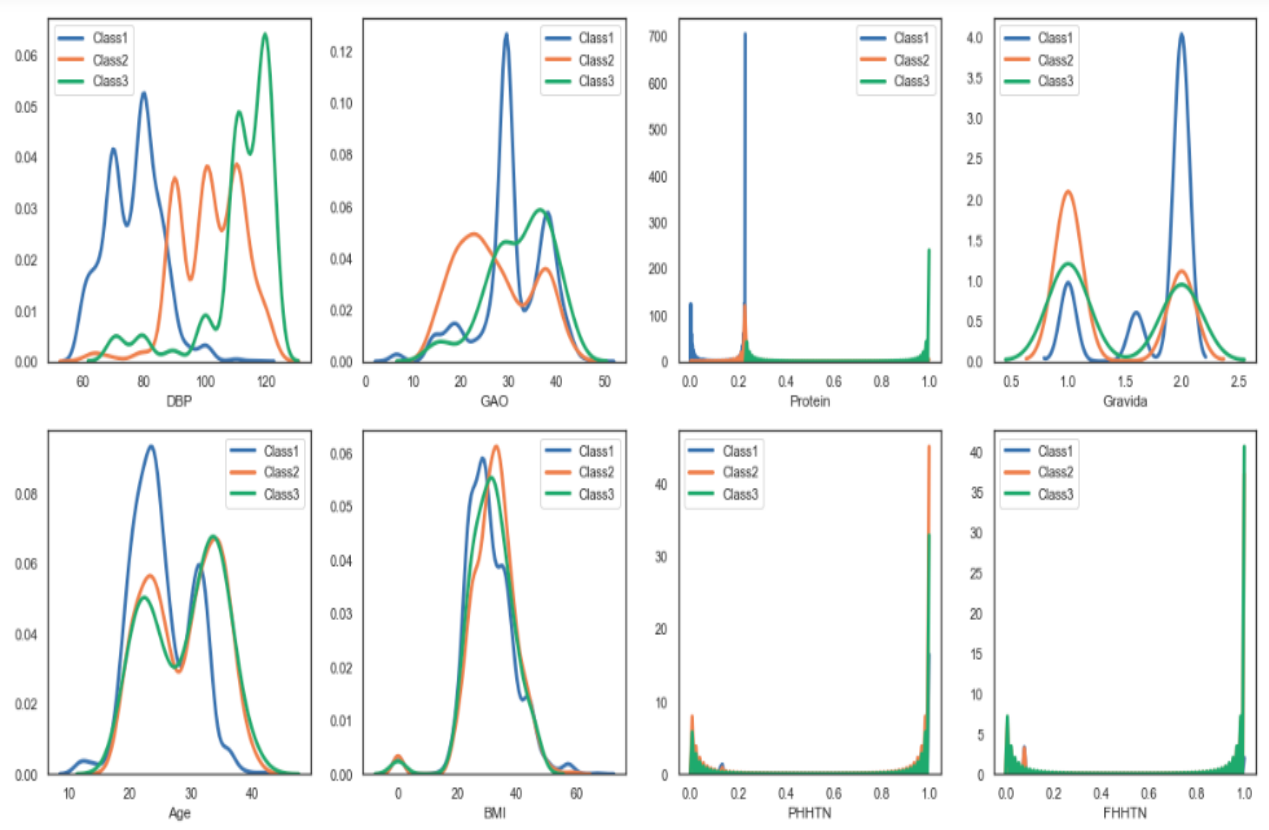
Figure 6.1: Class Distribution Of Dataset

Table 6.1: Class Distribution Of Dataset

Classes	***Count***
No PIH(o)	*1726*
GH(1)	*1132*
Preeclam (2)	*204*

6.1.1 Features Based on Their Classes

Fig.6.2. Shows results from seaborn visualization display the three-class distribution of features from the dataset which are displayed in figure 6.2, and class 1 is for no PIH, class 2 is for GH, and class 3 is for Preeclam class. As a visualization result, most patients with gestational hypertension(GH) and preeclampsia have systolic blood pressure(SBP) from 150 to 200 and with no pregnancy-induced hypertension(no PIH) have systolic blood pressure from 90 to 140. Furthermore, pregnant women with GAO having 20 weeks to 40 having GH and Preeclam. pregnant women who have protein in their urine have preeclampsia than those who do not have.



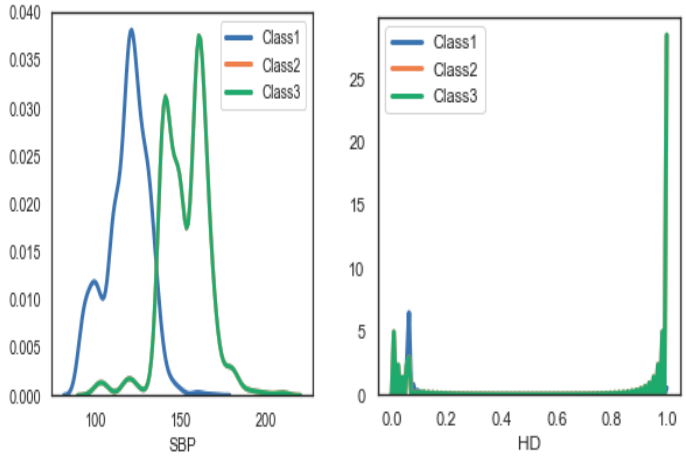


Figure 6.2: Feature Relationship With Classes

6.2 Feature Selection and Extraction Result

The features selection and Extraction process using the three methods PCA, RFE and UFS produced different sets of features sets from the dataset. And important features are produced using a random forest classifier algorithm as shown in fig.6.3.

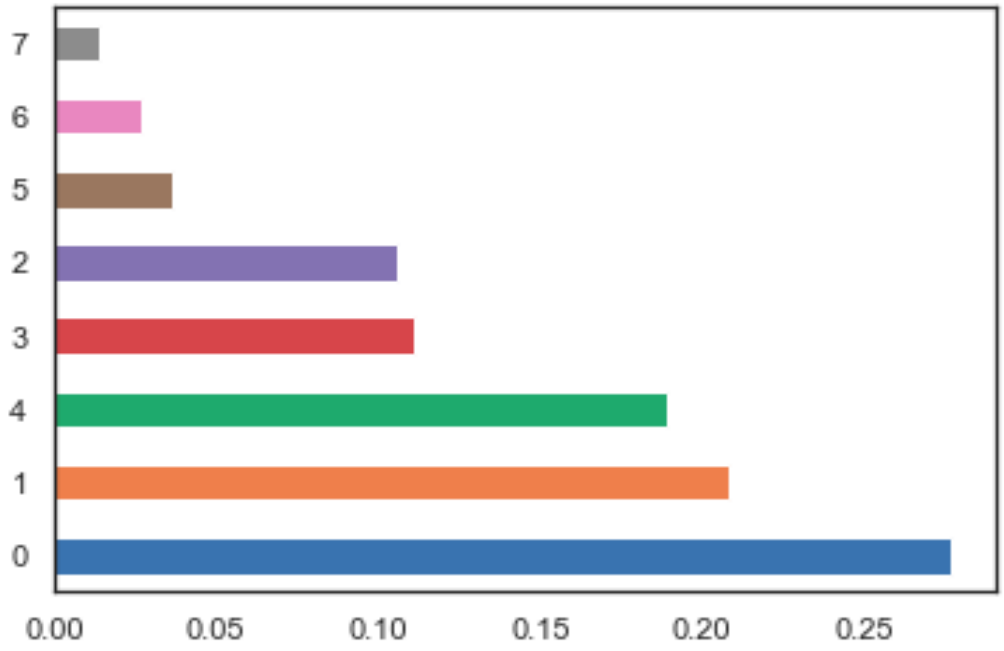


Figure 6.3: Feature Importance Using Random Forest Classifier

6.2.1 Selected Features Result

The features selection and extraction process using the three methods PCA, RFE and UFS produced three different sets of features sets from the dataset. These features sets are used in the training of RF, SVM, NB and models. Table 6.2 shows the selected and extracted features that sets the size of the dataset.

Table 6.2: Feature Selection Result

Feature selection method	***Selected/extracted features***
UFS	8
PCA	5
RFE	10

6.3 Model Evaluation Result

The Experiment resulted in 19 different models based on eight, five, and ten features, with three classification algorithms. Testing these trained models is performed using 5-fold cross-validation. This method randomly split the dataset into five equal size sets. Train the models using 5-1 folds train and test using one remain fold. The process is iteratively for each fold. The obtained Classification models Results described below.

6.3.1 Classification Model Evaluation Result

The trained models is tested using the 10-fold CV. The result of each test accuracy score for SVM, NB, and RF models based on extracted and selected feature sets are presented in tables 6.3,6.4,6.5. In this experiment, we have experimented with three feature selection and extraction methods and applied on machine learning algorithms and we evaluated by cross-validation accuracy to identify the best feature selection method with classification algorithms.

Table 6.3:Cross-Validation For Support Vector Machine

Feature Selection Method	***10-Fold Cross-Validation***										***Average Accuracy***
	1-fold	***2-fold***	***3-fold***	***4-fold***	***5-fold***	***6-fold***	***7-fold***	***8-fold***	***9-fold***	***10-fold***	
PCA	0.93	0.96	0.97	0.95	0.95	0.96	0.97	0.95	0.96	0.97	**0.96**
RFE	0.94	0.94	0.94	0.93	0.99	0.96	0.91	0.95	0.95	0.93	0.95
UFS	0.96	0.95	0.94	0.94	0.96	0.96	0.92	0.93	0.95	0.94	0.95

The results in Table 6.3 shows the CV accuracy scores for the SVM model based on feature with corresponding each fold test. The lower average accuracy of 0.95results recorded on RFE and UFS model and the higher accuracy 0.96 resulted using PCA

Table 6.4: Cross-Validation for Naïve Bayes

Feature Selection Method	***10-Fold Cross-Validation***										***Average Accuracy***
	1-fold	***2-fold***	***3-fold***	***4-fold***	***5-fold***	***6-fold***	***7-fold***	***8-fold***	***9-fold***	***10-fold***	
PCA	0.54	0.55	0.57	0.59	0.60	0.61	0.600	0.568	0.56	0.58	0.58
RFE	0.61	0.60	0.61	0.62	0.61	0.60	0.60	0.60	0.86	0.60	0.64
UFS	0.62	0.61	0.61	0.62	0.61	0.61	0.61	0.62	0.91	0.60	0.65

The results in Table 6.4 shows the CV accuracy scores for the NB model based on feature with corresponding each fold test. The lower average accuracy of 0.58 results recorded on the PCA model and the higher accuracy 0.65 resulted in using the UFS.

Table 6.5: Cross-Validation for Random Forest

Feature Selection Method	***10-Fold Cross-Validation***										***Average Accuracy***
	1-fold	***2-fold***	***3-fold***	***4-fold***	***5-fold***	***6-fold***	***7-fold***	***8-fold***	***9-fold***	***10-fold***	
PCA	0.93	0.96	0.97	0.95	0.97	0.97	0.95	0.96	0.97	0.96	**0.97**
RFE	0.96	0.95	0.98	0.96	0.96	0.96	0.95	0.95	0.96	0.97	0.96
UFS	0.98	0.95	0.98	0.963	0.967	0.98	0.955	0.967	0.967	0.971	0.96

The results in Table 6.5 shows the CV accuracy scores for the NB model based on feature with corresponding each fold test. The lower average accuracy of 0.96 results recorded on the RFE and UFS model and the higher accuracy 0.97 resulted in using PCA.

6.3.2 Classification Performance Result

Besides the result accuracy score obtained by five-fold CV. The study also uses other model's performance evaluation metrics, as discussed in chapters three and four. These metrics are Precision (P), Recall(R), and F1-score (F1). Table 6.6 shows the results of the evaluation metric for each model based on features selected and extracted. Also, it uses a normalized confusion matrix of the models using a five-fold prediction result shown in figures 6.4 below (confusion matrix).

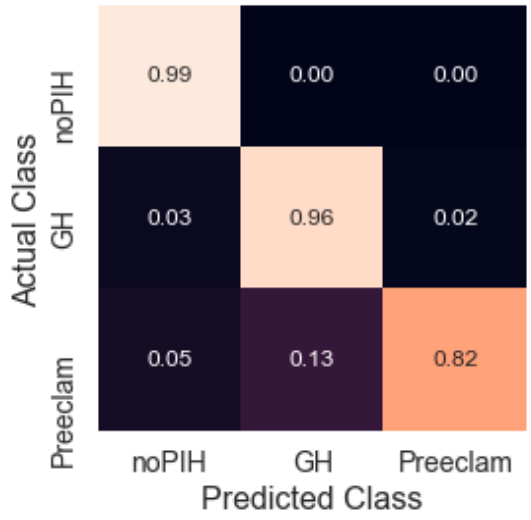
Table 6.6: Classification Performance Based on Feature Selection and Extraction

Feature selection and extraction method	***SVM***				***NB***				***RF***			
	Acc.	***P***	***R***	***F1***	***Acc.***	***P***	***R***	***F1***	***Acc.***	***P***	***R***	***F1***
PCA	0.96	0.96	0.96	0.96	0.61	0.92	0.61	0.69	0.97	0.97	0.97	0.97
RFE	0.94	0.95	0.95	0.95	0.59	0.40	0.59	0.47	0.96	0.96	0.96	0.96
UFS	0.95	0.96	0.96	0.96	0.60	0.42	0.61	0.49	0.96	0.97	0.97	0.97

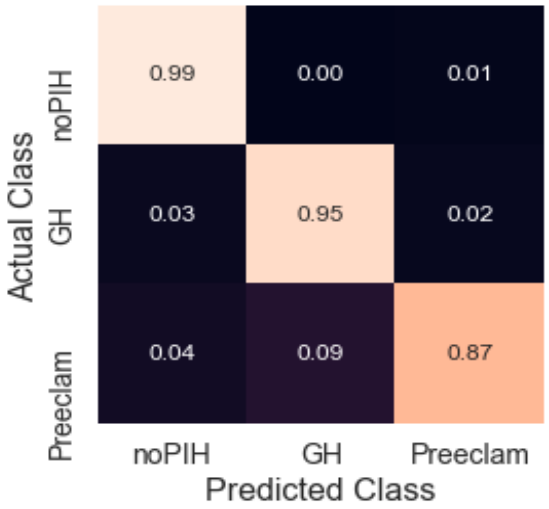
The result in table 6.6 shows, Based on selected and extracted features with feature extraction and feature selection method. The highest accuracy for SVM is 0.96 based on PCA, for NB is 0.61and for RF is 0.97. When comparing F1 measure for all classification model is 0.97for RF using PCA, 0.96 for SVM using PCA. Comparing the accuracy of classifier and F1 measure, RF with 0.97using PCA is the highest than the left model.

Normalized Confusion matrix for Random Forest on Selected features (PCA, RFE, UFS)

PCA



RFE



UFS

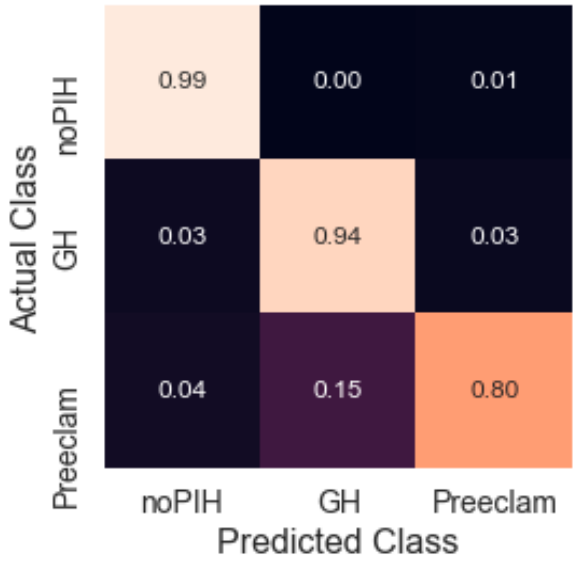


Figure 6.4: Confusion Matrix for Random Forest

Fig 6.4 shows the Confusion matrix for the RF model. RF with PCA 99% of no PIH, 98% of GH and 82% of Preeclampsia class are correctly classified but 3% of GH misclassified as no PIH,

5% of Preeclam misclassified as no PIH, 13% of Preeclam misclassified as GH, 2% of GH misclassified as Preeclam is incorrectly classified.

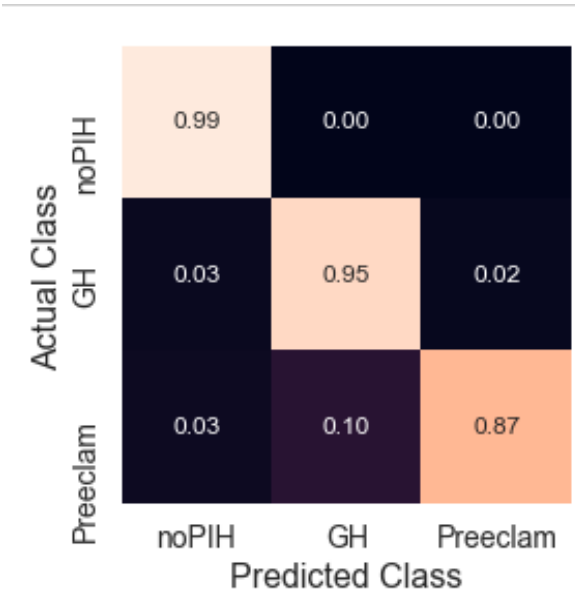
RF with RFE 99% of no PIH, 95% of GH, and 87% of Preeclampsia class are correctly classified. However, 3% GH classified as no PIH, 4% of Preeclam misclassified as no PIH, 9% of Preeclam misclassified as GH, 1% of no PIH misclassified as Preeclampsia 2% of GH misclassified as Preeclam.

RF with UFS 99% of no PIH, 94% of GH, and 80% of Preeclam is correctly classified. However, 3% of GH misclassified as no PIH, 4% of Preeclam misclassified as no PIH, 15% of Preeclam misclassified as GH, 1%of no PIH misclassified as Preeclam and 3% of GH misclassified as Preeclam.

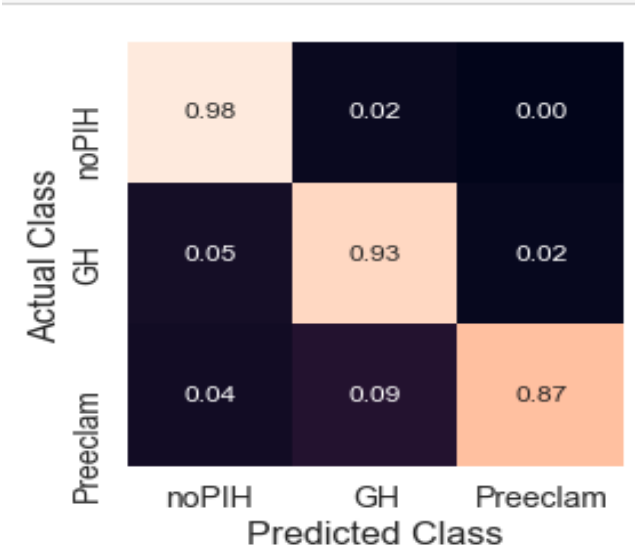
Normalized Confusion matrix for support vector machine on

Selected features (PCA, RFE, UFS)

PCA



RFE



UFS

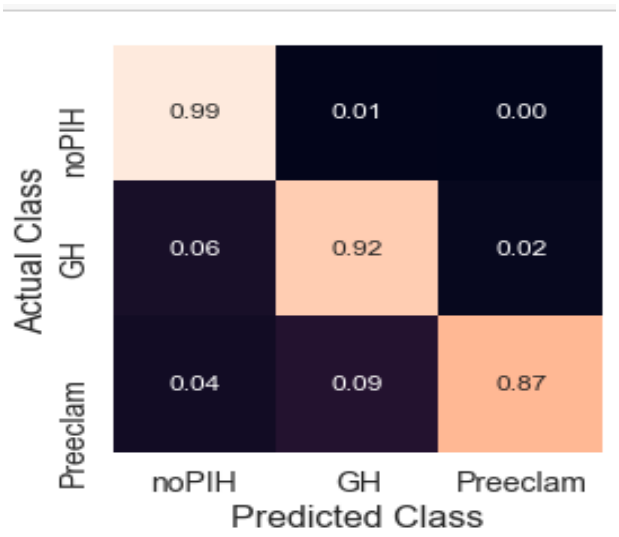


Figure 6.5; confusion matrics for SVM

Fig 6.5 shows the Confusion matrix for the SVM model. SVM with PCA 99% of no PIH, 95% of GH and 97% of Preeclampsia class are correctly classified but 3% of GH misclassified as no PIH, 3% of Preeclam misclassified as no PIH, 10% of Preeclam misclassified as GH, 2% of GH misclassified as Preeclam is incorrectly classified

SVM with RFE 98% of no PIH, 93% of GH and 87% of Preeclampsia class are correctly classified. However, 5% GH misclassified as no PIH, 4% of Preeclam misclassified as no PIH, 1% of no PIH misclassified as GH, 9% of Preeclam misclassified as GH, 2% of GH misclassified as Preeclam.

SVM with UFS 99% of no PIH, 92% of GH, and 87% of Preeclam is correctly classified. However, 6% of GH misclassified as no PIH, 4% of Preeclam misclassified as no PIH, 1% of no PIH misclassified as GH, 9% of Preeclam misclassified as GH and 2% of GH misclassified as Preeclam.

Algorithm Analysis Comparison between the proposed model and Related Approach

There is a few related studies done in pregnancy-induced hypertension using machine learning and other algorithms done for pregnancy-induced hypertension. All of the papers are their accuracy is less than 0.90 and there is one paper with an accuracy of 0.90 and all of them shown in table 6.7. from this result, We have seen that the largest accuracy result for the proposed result highest than papers done in previous.

Table 6.7: Algorithm Analysis Comparison between the proposed model and Related Approach

Author	***Methodology***	***Accuracy Result***
Proposed model	***PCA+ RF***	***0.97***
[12] Predicting Pregnancy Hypertension Using Random Forest	RF,NB,J48	RF with 0.73
[13] Preeclampsia Diagnosis Using Bayesian Network	BN	BN with 0.66
[27] Predicting Pregnancy Induced Hypertension Level Prediction	DT,SVM,LR	DT with 0.90
[54] Machine Learning Approach For Preeclampsia Risk Factor	SVM, RF	RF with 0.85
[26] Neural Network For Preeclampsia	NN	NN with 0.80

6.3.3 Integrating Prediction Model With Healthcare Server And Evaluation

The server is initialized from the command prompt after the server gives that address and the user can access the server using this address. The address contains html page which allows users to enter their status and parameters and get prediction results through this address page.fig.6.6 are users page to enter their parameters using Http address.this is input to flask server.

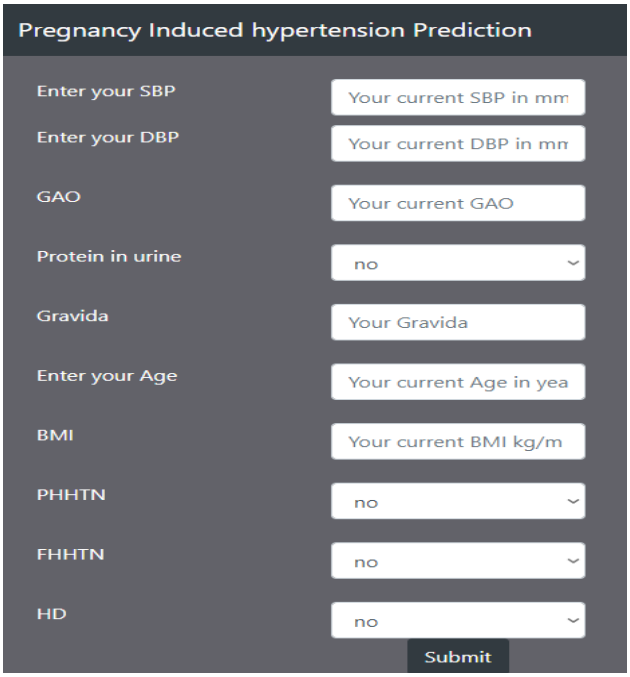


Figure 6.6:GUI for Input Data

Evaluation of Framework

In order to evaluate the framework, there are experiments is done, by giving this framework to experts and patients to evaluate the prediction of the framework. In this cases, experts and pregnant women have evaluated the framework by feeding pregnancy-induced disorder of features and get prediction from this framework. There are four interview was prepared for pregnant women and experts to evaluate the framework. The first one is what is the advantages youget from this framework for both pregnant women and experts. Pregnant women say ,they get a benefit from this framework for early identify the risk factor of pregnancy induced hypertension, get awareness of this disorder and also experts say rhey get benefit for accurately without biased to diagnose their patients moreover,helps for rmote area when there is a shortage of experts and doctors to diagnose their patients. this framework improve the traditional way of diagnosing patients because ,prediction erforms using electronic based system and users and experts get accurate prediction for diagnosing pregnancy induced hypertension. The left interview form is shown in appendix A.1.

6.4 Discussion

Early identification and prediction of pregnancy-induced hypertension can be helpful in improving the lives of pregnant women and increase their life expectancy. Different researchers developed a model for pregnancy-induced hypertension and in this study, supervised machine learning algorithms are used to predict pregnancy-induced hypertension and a flask server is used to help pregnant women to access and get benefit from the developed model. In this study, four feature extraction and selection method is used to identify and select relevant and most predictor features and three supervised machine learning model is used to predict pregnancy-induced hypertension.

Data were collected from Adama hospital, Asella hospital and Noah specialty clinic and after data is collected and registered in excel sheet, data preprocessing done by cleaning irrelevant data, redundant data and treating missing value carefully because clinical data needs careful imputation of missing value and by its nature, it is complex data. After preprocessing of data, the total dataset size is 3062 and out of this dataset 1729 are pregnant patients with normal or no pregnancy-induced hypertension, 1132 of them are pregnant patients with gestational hypertension and 204 of them are pregnant patients with preeclampsia.

The dataset contains three classes of pregnancy-induced hypertension and before applying feature selection and extraction, exploratory data analysis is done and this helps to identify the features and class distribution of a dataset. Results from Exploratory data analysis produced features and their correlation is described in figure 6.2 and figure 6.3.

There are three feature selections and the extraction model is applied in a dataset that helps to select and extract features. PCA, RFE, and UFS selects and extracts different features set and applied machine learning models on them and provides different accuracy results. To evaluate the classification model and classification performance, K-fold cross validation,precision,recall and confusion matrix method is done. In this study feature extraction using PCA is the best model to extract features and later helps to apply the machine learning model and provides the best accuracy as compared to RFE, and UFS in both Cross-validation and classification performance of the model. And random forest with PCA feature extraction gives an accuracy of 0.97 but SVM with PCA produced an accuracy of 0.96 and NB with PCA produced an accuracy of 0.61. Applying feature selection and extraction on the dataset helps to increase the accuracy of the proposed model.

Integrating the machine learning model with the healthcare server helps pregnant women to identify their health status early and helps professionals to diagnose their patients using this model because the developed model has an accuracy of 0.97, this provides confidence to the user to use this model. In this study flask server is used to deploy the developed model into the server. For clients to input their data, the Html page is created and accessed the model through this page.evaluation of this framework is done using prepared interview question to pregnant women and experts.

Pregnancy-induced hypertension prediction developed by Sharma [27], using decision tree and SVM algorithms and the result describe the accuracy of 0.90 by decision tree and they used two features to predict pregnancy-induced hypertension. Those are systolic and diastolic blood pressure and the number or size of the dataset is 50. Martinez. [54] Developed a machine learning model for preeclampsia. The result of the study gives an accuracy of 0.85 using a random forest algorithm. This method extract features using the Gini Index and constructs model using 25 features. The analysis of the previous paper is stated in table 6.7. And the proposed

model is the best accurate result and classification performance of the proposed model also results increasing accuracy.

In this thesis work, there is four research questions are answered based on developing a framework for pregnancy-induced hypertension using a machine learning approach. In the first research question, there is no feature selection or extraction method used for pregnancy-induced hypertension to select and extract features that are the most predictor of features. Even if the existing study predicts pregnancy-induced hypertension by simulation of blood pressure, but there is no included features that are related to this disorder and used feature selection and extraction methods. This study answers the first question by applying the feature selection and extraction method in our dataset and suggesting PCA is the best feature extractor of our dataset.

In the second questions, there is some study which used algorithms for preeclampsia using neural network, artificial neural network, boosted random forest and all of them produced an accuracy of less than 90% and there is one paper applying machine learning algorithm for predicting pregnancy-induced hypertension level using decision tree with an accuracy of 90. In our study, applied three machine learning algorithm those are RF, SVM, and NB and produced the largest accuracy with RF of 0.97 and this is the largest and recommended to the user to use this dataset with this model for diagnosing patients and helps users.

The third research question is increasing the accuracy of our model other than previously developed models because the machine learning model with high accuracy helps to trust this model for use by other researchers and for users to access this model. And the last research question depends on how to integrate our model and a server because there is no study done to integrate machine learning model with a server. Integrating machine learning model with the server contribute user and professional to access the model for diagnosing of their patients and users to identify early their disorder accurately.

6.5 Contribution of the study

The study extends the effort in solving the problem of pregnant women and ccomplication whicha happen during pregnancy. The research contributes to its part in the prediction of pregnancy induced hypertension. The contribution of this research is:

- Predict pregnancy induced hypertension
- Helps professional in decision making process
- Helps pregnant women to get accurate information early and to save them selves early
- Contributes prepared dataset which is not prepared previously
- Increase the availaibility framework for pregnancy induced hypertension prediction in health sector

CHAPTER SEVEN

7 Conclusion, Recommendation And Future Work

In this chapter, the summary of the proposed framework for pregnancy-induced hypertension prediction using a machine learning approach, recommendation, future research directions has been described.

7.1 Conclusion

This research proposed a framework for pregnancy-induced hypertension by integrating machine learning and healthcare server. The study attempted to develop, implement feature selection and extraction methods for collected and preprocessed datasets then applying the machine learning algorithms and compares those algorithms their performance then integrating the largest accurate performed algorithm with healthcare server.

To design this framework, data were collected from three hospitals and study in this area and advice from experts and doctors is done. After data is collected and preprocessed using the python tool which is the most popular machine learning and data science tool and prepared cleaned dataset. The total cleaned dataset size is 3062, machine learning algorithm were applied on this dataset.

In this study, there are four feature selections and extraction methods is used to extract and select features those are PCA, RFE, and UFS for developing our model, and those feature selection and extraction method select features and three machine learning algorithms RF, SVM and NB algorithms are applied. On all feature selection and extraction method, three machines learning algorithms is applied on each of them and produced the largest accurate result is PCA with random forest algorithms produced accuracy of 0.97 and this is the largest accurate result in experiment and the left algorithms are produced slightly low performance than random forest algorithms. Finally, PCA with RF model is used to integrated with healthcare server and the server used in this used is flask micro-framework which is the popular python library, from this server user and professionals can access and benefit early prediction of pregnancy induced hypertension.

7.2 Recommendation

Nowadays pregnancy-induced hypertension is a silent killer, and a lot of pregnant women lost their life by this disease, through this research finding pregnant women can early identify whether they have this disorder or not by the developed framework.

There is no prepared dataset concerning this disorder online and in Ethiopia this limits researcher to more study this area and there is no communication between hospitals research study department and technology department or area if they communicate and do research with each other, there is more improvement of problem in health area both for professional and for researcher.

The last recommendation to this area is, in Ethiopia, there is a shortage of physiological measurement sensors to collect patients' data in realtime from the remote areas for blood pressure sensors, protein measuring sensors. It is better for the user if this sensor was available for measuring their status and deploying on the server for getting prediction results from the server.

7.3 Future Work

The proposed solution collected data from three hospitals and predicting pregnancy-induced hypertension is not enough for generalizing the risk factor of this disorder .Since, there is no prepared dataset online data was collcted and dataset prepared in a few number of features. In the future it is better to increase the size of dataset and collecting from the left area of hospitals and feeding to this prediction model and analaysing the result. The other future work is, IOT based data collection and prediction of pregnancy-induced hypertension is a more effective method for pregnant women to identify their physiological disorder without going to healthcare remotely and using their mobile devices.

REFERENCES

[1] A. K. Berhe, G. M. Kassa, G. A. Fekadu, and A. A. Muche, "Prevalence of hypertensive disorders of pregnancy in Ethiopia : a systemic review and meta-analysis," pp. 1–11, 2018.

[2] K. Hospital, L. Mekonen, Z. Shiferaw, E. Wubshet, and S. Haile, "Journal of Pregnancy and Child Health Pregnancy Induced Hypertension and Associated Factors among Pregnant," vol. 5, no. 3, pp. 3–6, 2018.

[3] T. M, "Hypertensive Disorders Of Pregnancy And Associated Factors," no. 2394, 2016.

[4] G. Ayele, "Factors Associated with Hypertension during Pregnancy in Derashie Woreda South Ethiopia , Case Control," vol. 24, pp. 207–213, 2016.

[5] O.Niem. Article, "Hypertensive Disorders Of Pregnancy In Jimma University Specialized Hospital," 2017.

[6] R. Townsend, P. O. Brien, and A. Khalil, "Current best practice in the management of hypertensive disorders in pregnancy," pp. 79–94, 2016.

[7] J. M. Roberts, G. Pearson, J. Cutler, and M. Lindheimer, "Hypertension During Pregnancy," 2003.

[8] N.Chu."Hypertensive Disorders Of Pregnancy And Associated Factors," No. 2394, 2016.

[9] I. Journal, F. Technological, and T. N. Joshi, "Logistic Regression And Svm Based Diabetes," vol. 5, no. 11, pp. 4347–4350, 2018.

[10] S. Vijayarani and S. Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," vol. 4, no. 4, pp. 816–820, 2015.

[11] D. S. Medhekar, M. P. Bote, and S. D. Deshmukh, "Heart Disease Prediction System using Naive Bayes," vol. 2, no. 3, pp. 1–5, 2013.

[12] W. L. Moreira, J. J. P. C. Rodrigues, A. M. B. Oliveira, K. Saleem, and A. J. Venˆ, "Predicting Hypertensive Disorders in High-risk Pregnancy Using the Random Forest Approach," 2017.

[13] M. W. L. Moreira, J. J. P. C. Rodrigues, A. M. B. Oliveira, R. F. Ramos, and K. Saleem, "A Preeclampsia Diagnosis Approach using Bayesian Networks," 2016.

[14] N. Mathew, "A Boosting Approach for Maternal Hypertensive Disorder Detection," *2018 Second Int. Conf. Inven. Commun. Comput. Technol.*, no. Icicct, pp. 1474–1477, 2018.

[15] J. H. Jhee., "Prediction model development of late-onset preeclampsia using machine learning-based methods," pp. 1–12, 2019.

[16] N. Farooqui, R. Mehra, and A. Tyagi, "Prediction Model for Diabetes Mellitus Using Machine Learning Techniques International Journal of Computer Sciences and Engineering Open Access Prediction Model for Diabetes Mellitus Using Machine Learning Techniques," no. March, 2018.

[17] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," vol. 9, no. November, pp. 1–10, 2018.

[18] S. P. S, "Chronic Kidney Disease Prediction Using Machine Learning," vol. 16, no. 4, pp. 308–311, 2018.

[19] S. Tekale, P. Shingavi, S. Wandhekar, and A. Chatorikar, "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm," vol. 7, no. 10, pp. 92–96, 2018.

[20] A. F. Otoom, E. E. Abdallah, Y. Kilani, and A. Kefaye, "Effective diagnosis and monitoring of heart disease Effective Diagnosis and Monitoring of Heart Disease," no. January, 2015.

[21] Y. Khourdifi, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization," vol. 12, no. 1, 2019.

[22] M. Lixi, "ICT as an Enabler of Transformation in Ethiopia," 2014.

[23] H. Sector and T. Plan, "Health Sector Transformation Plan," 2019.

[24] J. M, "Prevention and treatment of pre-eclampsia and eclampsia," 2017.

[25] T. A. Gudeta, "Prevalence of Pregnancy Induced Hypertension and Its Bad Birth Outcome Journal of Pregnancy and Child Health Prevalence of Pregnancy Induced Hypertension

and Its Bad Birth Outcome among Women Attending Delivery Service," no. January, 2017.

[26] C. K. Neocleous, P. Anastasopoulos, K. H. Nikolaides, C. N. Schizas, and K. C. Neokleous, "Neural networks to estimate the risk for preeclampsia occurrence," pp. 8–11.

[27] P. T. and R. P. Anuja Hiwale, *Prediction of Pregnancy-Induced Hypertension Levels Using Machine Learning Algorithms*. 2019.

[28] M. Espinilla, J. Medina, S. Campaña, and J. Londoño, "Fuzzy Intelligent System for Patients with Preeclampsia in Wearable Devices," vol. 13, 2017.

[29] H. B. Kahsay, F. E. Gashe, and W. M. Ayele, "Risk factors for hypertensive disorders of pregnancy among mothers in Tigray region , Ethiopia : matched case-control study," vol. 5, pp. 1–10, 2018.

[30] G. D. Magoulas and A. Prentza, "Machine Learning in Medical Applications," no. September, 2001.

[31] A. Dhillon and A. Singh, "Machine Learning in Healthcare Data Analysis : A Survey," vol. 8, no. July, pp. 1–10, 2019.

[32] S. Report, "A study on machine learning algorithm in medical diagnosis," vol. 9, no. 4, pp. 42–46, 2018.

[33] I. H, "Basic Architecture of Any Machine Learning/Deep Learning Project." [Online]. Available: https://medium.com/analytics-vidhya/basic-architecture-of-any-machine-learning-deep-learning-project-21e687250f22.

[34] S. Lim, C. S. Tucker, and S. Kumara, "An unsupervised machine learning model for discovering latent infectious diseases using social media data," *J. Biomed. Inform.*, vol. 66, pp. 82–94, 2017.

[35] S. Satu, F. Tasnim, and T. Akter, "Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms," *2018 Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng.*, vol. 234, pp. 1–4, 2018.

[36] C. Yu, J. Liu, and S. Nemati, "Reinforcement Learning in Healthcare : A Survey."1019

[37] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, “Comparative Study of Classification Techniques (SVM , Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease,” vol. 5, no. 5, 2015.

[38] L. E. Juarez-orozco, O. Martinez-manzanera, S. V Nesterov, S. Kajander, and J. Knuuti, “The machine learning horizon in cardiac hybrid imaging,” 2018.

[39] S. E. Awan, M. Bennamoun, F. S. Id, F. M. Sanfilippo, B. J. Chow, and G. D. Id, “Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death,” pp. 1–13, 2019.

[40] C. Zhu, C. Uwa, and W. Feng, “Informatics in Medicine Unlocked Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques,” *Informatics Med. Unlocked*, no. April, p. 100179, 2019.

[41] S. Paul, “Beginner’s Guide to Feature Selection in Python.” [Online]. Available: https://www.datacamp.com/community/tutorials/feature-selection-python.

[42] A. Jovi, K. Brki, and N. Bogunovi, “A review of feature selection methods with applications,” vol. 19.

[43] “Beginner’s Guide to Feature Selection in Python (article) - DataCamp.” . [Online]. Available: https://www.datacamp.com/community/tutorials/feature-selection-python.

[44] H. Polat, H. D. Mehr, and A. Cetin, “Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods,” vol. 567, no. April, pp. 3–11, 2017.

[45] A. H. Roslina, “Prediction Of Hepatitis Prognosis Using Support Vector Machines And Wrapper Method,” no. Fskd, pp. 2209–2211, 2010.

[46] G. Dong, “A Quick Guide to Feature Engineering,” 2019. [Online]. Available: https://www.kdnuggets.com/2019/02/quick-guide-feature-engineering.html.

[47] J. Tang and S. Alelyani, “Chapter 2 Feature Selection for Classification : A Review,” 2015.

[48] S. S. Senthil *et al.*, “Comparison Of Feature Selection Methods For Chronic Kidney Data

Set Using Data Mining Classification Analytical Model," pp. 459–469, 2019.

[49] E. Hemphill, J. Lindsay, C. Lee, I. M. Ion, and C. E. Nelson, "Feature selection and classifier performance on diverse bio- logical datasets," vol. 15, no. Suppl 13, pp. 1–14, 2014.

[50] B. Tesfaye, S. Atique, T. Azim, and M. M. Kebede, "Predicting skilled delivery service use in Ethiopia : dual application of logistic regression and machine learning algorithms," vol. 5, pp. 1–10, 2019.

[51] G. G. Teklehaymanot and N. Gandhi, "Analyzing Children ' s Data Using Machine Learning : A Case Study in Ethiopia," no. November, 2018.

[52] G. Sahle, "Ethiopic maternal care data mining : discovering the factors that affect postnatal care visit in Ethiopia," *Heal. Inf. Sci. Syst.*, no. May, 2016.

[53] E. Tejera, M. J. Areias, A. N. A. Rodrigues, A. N. A. Ramo, J. M. Nieto-villar, and I. Rebelo, "Artificial neural network for normal , hypertensive , and preeclamptic pregnancy classification using maternal heart rate variability indexes," vol. 24, no. 164, pp. 1147–1151, 2011.

[54] A. Martinez-velasco and L. Miralles, "Machine Learning Approach for Pre-Eclampsia Risk Factors Association Machine Learning Approach for Pre-Eclampsia Risk Factors Association," no. January 2019, 2018.

[55] J. Rodrigues, "Predicting Hypertensive Disorders in High-risk Pregnancy Using the Random Forest Approach Predicting Hypertensive Disorders in High-risk Pregnancy Using the Random Forest Approach," no. May, 2017.

[56] J. Ferencz and C. Serban, "Patient Data Collection Methods . Retrospective Insights . Patient Data Collection Methods . Retrospective Insights .," no. July 2017, 2018.

[57] M. N. Vrahatis, *Data preprocessing in predictive data mining*, vol. 34. 2019.

[58] Jason Brownlee, "Rescaling Data for Machine Learning in Python with Scikit-Learn." [Online]. Available: https://machinelearningmastery.com/rescaling-data-for-machine-learning-in-python-with-scikit-learn/.

[59] kraj, “Easy Explanation of Data Normalization/Standardization in Machine Learning(using Python).” [Online]. Available: https://kraj3.com.np/blog/2019/06/easy-explanation-of-normalization-standardization-in-machine-learningusing-python/.

[60] J. Brownlee, “Feature Selection For Machine Learning in Python.” [Online]. Available: https://machinelearningmastery.com/feature-selection-machine-learning-python/.

[61] F. Song, Z. Guo, and D. Mei, “Feature selection using principal component analysis,” 2010.

[62] C. Technology, “Classification Of Heart Disease Using Svm And ANN,” vol. 2, no. 9, pp. 694–701, 2013.

[63] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, “A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL),” vol. 68, no. 16, pp. 11–15, 2013.

[64] H. Kaur and V. Kumari, “Applied Computing and Informatics Predictive modelling and analytics for diabetes using a machine learning approach,” *Appl. Comput. Informatics*, no. xxxx, pp. 0–5, 2018.

[65] S. B. Choi *et al.*, “Screening for Prediabetes Using Machine Learning Models,” vol. 2014, 2014.

[66] S. Dhamodharan, “Liver Disease Prediction Using Bayesian Classification,” no. May, pp. 1–3, 2014.

[67] D. Singh, “Validating Machine Learning Models with scikit-learn.” [Online]. Available: https://www.pluralsight.com/guides/validating-machine-learning-models-scikit-learn.

[68] D. Jain and V. Singh, “Feature selection and classification systems for chronic disease prediction : A review,” *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, 2018.

[69] B. Adibhatla, “Django Vs Flask: A Look At The Two Popular Python Web Frameworks,” 2019. [Online]. Available: https://analyticsindiamag.com/django-vs-flask-a-look-at-the-two-popular-python-web-frameworks/.

[70] W. Vincent, “Flask vs Django (2020).” [Online]. Available: https://wsvincent.com/flask-

vs-django/.

[71] Aspher Brothers, “Django vs Flask in 2019 – Comparison of most popular Python Frameworks.” [Online]. Available: https://asperbrothers.com/blog/django-vs-fla.

[72] M, “Edraw Max, All-in-One Diagram Software,” 2019. [Online]. Available: https://www.edrawsoft.com/edraw-max/.

[73] Kayla Matthews, “6 Reasons Why Python Is Suddenly Super Popular,” *www.kdnuggets*, 2017. [Online]. Available: https://www.kdnuggets.com/2017/07/6-reasons-python-suddenly-super-popular.html.

[74] N. Mertens, “Top 4 Reasons Python is So Popular in 2020,” 2019. [Online]. Available: https://www.goskills.com/Development/Articles/Why-is-Python-so-popular.

[75] B. Trstenjak, “Adaptable Web Prediction Framework for Disease Prediction Based on the Hybrid Case Based Reasoning Model,” vol. 6, no. 6, pp. 1212–1216, 2016.

[76] V. Kumar, “Deploying machine learning models for free,” *Deploy Machine Learning Models for Free*, 2019. [Online]. Available: https://medium.com/analytics-vidhya/how-to-deploy-simple-machine-learning-models-for-free-56cdccc62b8d.

APPENDIXES

Appendix A : Interview Question

Appendix A.1 : Interview Question for Collecting Data

1. What is pregnancy induced hypertension induced hypertension?
2. What is the risk factor for pregnancy induced hypertension?
3. What is the socio-dimographic characteristics of pregnancy induced hypertension?
4. What is the medical characteristics of pregnancy induced hypertension?
5. What is the Obstetric Gynecology characteristics of pregnancy induced hypertension?
6. What is the effect of pregnancy induced hypertension?

Appendix A.2 : Interview Question for Evaluating Framework

1. What is the advantages you get from this framework?
2. How this framework improves the manual way of diagnosing pregnancy induced hypertension?
3. How you can trust this framework?
4. How this framework helps healthcare which have small number of experts or doctors?

Appendix B : A sample source code

Appendix B.1: A Sample Code for Preprocessing

Data preprocessing

```
# Outputs the columns Independent Variable
X = df.iloc[:, :-1].values

# Outputs the last column + all its values - Dependent Variable
y = df.iloc[:,14].values
```

```
# Import library and class
from sklearn.preprocessing import Imputer

# Initalize Imputer
imputer = Imputer(missing_values="NaN", strategy="mean", axis=0)

# Grab only the columns with the missing data
imputer = imputer.fit(X[:, 1:14])

# Replace the missing fields of data with the mean of the column
X[:, 1:14] = imputer.transform(X[:, 1:14])
```

```
# Import library and classes
from sklearn.preprocessing import LabelEncoder, OneHotEncoder

# Initalise LabelEncoder for IV
labelEncoder_X = LabelEncoder()

# Change Values to an array of 'label numbers' & add them to X
X[:, 0] = labelEncoder_X.fit_transform(X[:, 0])

# Use Dummy Encoding:  - only first column
oneHotEncoder = OneHotEncoder(categorical_features=[0])

# Dummy Encode & convert to array
X = oneHotEncoder.fit_transform(X).toarray()

# Initalise LabelEncoder for DV
labelEncoder_y = LabelEncoder()

# As yes & no responses - don't need to Dummy Encode DV (binary responses)
y = labelEncoder_y.fit_transform(y)
```

```
# Avoiding the Dummy Variable Trap (Remove first column from X)
X = X[:, 1:]
```

Appendix B.2 A Sample Code for Normalization

```
# Import library and function
from sklearn.cross_validation import train_test_split

# Test set: 20%; Training Set: 80%
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
# Import library and classes
from sklearn.preprocessing import StandardScaler

# Initialise Feature Scaling
sc_X = StandardScaler()

# Scale X Training set by fitting it to X & transforming it
# X Train must be done first before X Test. This ensures that both sets are on the same scale
X_train = sc_X.fit_transform(X_train)

# Scale X Test set by transforming it
X_test = sc_X.transform(X_test)
```

Appendix B.3 A sample code for feature extraction using PCA

```
                # 2 principal component Analysis(PCA)
from sklearn.decomposition import PCA
names = ['SBP', 'DBP', 'GAO','Protein' 'Gravida', 'Age', 'BMI', 'PHHTN', 'FHHTN','Residence','HD','HCD','HRD','smoking' 'Class']
array = df.values
X = array[:,0:13]
Y = array[:,13]
# feature extraction
pca = PCA(n_components=5)
fit = pca.fit(X)
# summarize components
print("Explained Variance: %s" % fit.explained_variance_ratio_)
print(fit.components_)
plt.show()
```

Appendix B.4: Sample Code for Building and Evaluating Models

```
# Random Forest Classifier
from sklearn import model_selection
# random forest model creation
rfc = RandomForestClassifier()
rfc.fit(X_train, y_train)
rfcP_predict = rfc.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score,make_scorer
print(confusion_matrix(y_test, rfcP_predict ))
print(classification_report(y_test, rfcP_predict ))
print('accuracy for rfc: %s' % accuracy_score(y_test, rfc_predict))
# cross validation for naive bayes
from sklearn.model_selection import cross_val_score
scores = cross_val_score(svm, X_train, y_train, cv=5) #use 5 folds
print(scores)
print("Average Cross Validation Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

```
#  Support Vector Machine
from sklearn import svm
svm = svm.SVC(kernel = 'linear')
svm.fit(X_train, y_train)
svm_predict = svm.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score,make_scorer
print(confusion_matrix(y_test, svm_predict ))
print(classification_report(y_test, svm_predict ))
print('accuracy for svm: %s' % accuracy_score(y_test, svm_predict))
# cross validation for naive bayes
from sklearn.model_selection import cross_val_score
scores = cross_val_score(svm, X_train, y_train, cv=5) #use 5 folds
print(scores)
print("Average Cross Validation Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

```
# Naive Bayes
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB().fit(X_train, y_train)
gnb_predictions = gnb.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score,make_scorer
print(confusion_matrix(y_test, gnb_predictions ))
print(classification_report(y_test, gnb_predictions ))
print('accuracy for nb: %s' % accuracy_score(y_test, gnb_predictions))
# cross validation for naive bayes
from sklearn.model_selection import cross_val_score
scores = cross_val_score(GaussianNB(), X_train, y_train, cv=5) #use 5 folds
print(scores)
print("Average Cross Validation Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

Appendix B.5 :Sample code for integrating model and flask

```
from flask import Flask, render_template, url_for, request
from sklearn.externals import joblib
import os
import numpy as np
import pickle

app = Flask(__name__, static_folder='static')

@app.route("/")
def index():
    return render_template('home.html')


@app.route('/result', methods=['POST', 'GET'])
def result():


 scaler_path = os.path.join(os.path.dirname(__file__), 'models/scaler.pkl')
 scaler = None
 with open(scaler_path, 'rb') as f:
     scaler = pickle.load(f)

 x = scaler.transform(x)

 model_path = os.path.join(os.path.dirname(__file__), 'models/rfc.sav')
 clf = joblib.load(model_path)
```