A Model for Early Prediction of Diabetes using ANN along with Machine Learning

BY

Minhazul Zannat
1640CSE00466

BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering

Manarat International University

Ashulia, Dhaka, Bangladesh

Approval

This thesis titled "**A model for early prediction of diabetes using ANN along with Machine Learning"** submitted by the following student has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

Minhazul Zannat 1640CSE00466

Examination held on 10th December 2020

BOARD OF EXAMINERS

Sohaib Abdullah
Assistant Professor
Department of Computer Science and Engineering
Manarat International University

Chairman (Supervisor)

Muhammad Sajjad Hussain
Associate Professor & Head
Department of Computer Science and Engineering
Manarat International University

Member (Ex- Officio)

Md. Ali Hossain
Assistant Professor
Department of Computer Science and Engineering
Manarat International University

Member

Zahurul Haque
Lecturer
Department of Computer Science and Engineering
Manarat International University

Member

Dr. Md. Haider Ali
Professor
Department of Computer Science and Engineering
University of Dhaka

Member (External)

DECLARATION

This is to certify that the work presented in this thesis entitled "**A model for early prediction of diabetes using ANN along with Machine Learning**", is the outcome of the research carried out by Minhazul Zannat under the supervision of Sir, Sohaib Abdullah.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma, or other qualifications.

_____

Minhazul Zannat
1640CSE00466

ABSTRACT

With the growth of Information and communication technologies, the health care industry is also producing extensively large data. For managing such large amount of data, an efficient knowledge discovery process is required. This field is developing fast and there is a big scope of early planning towards the treatment of large number of diseases. The planning can be done by developing some strategic solutions based on Data Mining for the treatment of the disease. Classification based on supervised learning is a technique of Data Mining which helps in predicting the label of unknown samples as Class. This is extremely popular technique of Data Mining by which the treatment of a disease could be planned at an early stage. Diabetes is one of the chronic diseases produces metabolism disorder in human bodies. Metabolism refers a chemical process in human body responsible for energy conversion and utilization. Diabetes indicates excess glucose level in the blood could be cured if regular precautions have been taken persistently under certain clinical guidelines. Diabetes mellitus (early stage of diabetes) is a group of diseases characterized by high levels of blood glucose resulting from defects in insulin production, insulin action, or both. The term diabetes mellitus describes a metabolic disorder of multiple an etiology characterized by chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. Prediction of diabetes at an early stage can lead to improved treatment. Data mining techniques are widely used for prediction of disease at an early stage. In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, and association rule mining for diabetes. Significant attributes selection was done via the principal component analysis method. Our findings indicate a strong association of diabetes with body mass index (BMI) and with glucose level, which was extracted via the machine learning approaches. Ridge Classifier, Logistic Regression, Support Vector Classifier, XGB Classifier, Random Forest (RF), Ada-Boost Classifier, K-means clustering and many other techniques along with Artificial Neural Network (ANN) were implemented and tested for the prediction of diabetes. The ANN technique provided a best accuracy of **87.66%,** where combining result of machine learning algorithms and ANN gives **88.31%** accuracy and may be useful to assist medical professionals with treatment decisions. The dataset used in this study, is originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases (publicly available at: UCI ML Repository [20])

Keywords

Machine Learning Classifications, Data mining, Diabetes, Artificial Neural Network (ANN)

ACKNOWLEDGEMENT

This chapter is a study of multiple papers and documents to create a model for early prediction of diabetes. I am extremely grateful to all the papers and documents which are taken into account in this study. This paper is focused on creating an ANN model along with the classification techniques of machine learning approach on basis of the studies and never excuse or complain any of the section of those papers and documents, even though some methods and studies are modified and ignored.

If words are considered as a symbol of approval and token of appreciation, then let the words play the heralding role expressing my gratitude. The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

My heartiest gratitude, profound indebtedness and deep respect go to my supervisor sir Sohaib Abdullah for his constant supervision, affectionate guidance and great encouragement and motivation which helps me to make right choices in implementation phase.

I am thankful to my senior brother Shamìmul Islam Shamim; student of Manarat International University, for helping me in the study of this thesis by giving his valuable suggestions and helping hands.

I am greatly thankful to the Almighty Allah for His blessings to me and providing me the helpful environment for my study and the successful completion of my thesis.

Minhazul Zannat

ID: 1640CSE00466

CONTENTS

APPROVAL, BOARD OF EXAMINERS **i**

DECLARATION **ii**

ABSTRACT **iii**

ACKNOWLEDGEMENT **iV**

LIST OF FIGURES **Vii**

LIST OF TABLES **Viii**

1. **INTRODUCTION** **1**

1.1 Problem Definition **2**

1.2 Researching filed on Diabetes **2**

1.3 Attributes of Diabetes **2**

1.4 Introduction to classifiers of predicting diabetes **2**

1.5 Inspiration Behind This Plan **4**

1.6 Objective of The Thesis **4**

1.7 Overview of the Thesis **5**

1.8 Contributions **5**

1.9 Thesis Outline **6**

2. **RESEARCHES ON THIS FIELD** **7**

2.1 A model for classifying diabetic patient control level using data mining **8**

2.2 Data Mining Models Comparison for Diabetes Prediction **9**

2.3 A model for early prediction of diabetes **10**

2.4 Prediction of Diabetes using Classification Algorithms **11**

3. **METHODOLOGY** **13**

3.1 Dataset Description **13**

3.2 Data Splitting **14**

3.2.1 Actual Split **14**

	3.2.2	Random Split	**14**
	3.3	Data Cleaning	**14**
	3.4	Feature Engineering	**16**
	3.5	Data reduction (Normalization)	**19**
	3.5.1	Logarithm & Log1p()	**20**
	3.5.3	Sqrt()	**21**
	3.5.4	Skewness	**21**
4.	**NETWORK MODELING**		**24**
	4.1	Machine Learning(ML) Procedures	**24**
	4.1.1	Supervised learning(SL)	**24**
	4.1.2	Unsupervised learning(UL)	**25**
	4.1.3	Reinforcement Learning(RL)	**25**
	4.2	Artificial intelligence	**25**
	4.3	Network Architecture	**26**
	4.3.1	Artificial neural network (ANN)	**26**
	4.4	Machine Learning approaches	**30**
	4.4.1	Logistic Regression & Ridge Classifier	**30**
	4.4.2	Support Vector Classifier (SVC)	**31**
	4.4.3	XGBoost & AdaBoost Classifier	**32**
	4.5	Training Procedure	**33**
	4.5.1	ANN model training parameters	**34**
	4.5.2	Machine Learning Algorithms' parameters	**35**
5.	**RESULTS AND DISCUSSION OF THIS STUDY**		**37**
	5.1	Limitations of The Study	**40**
6.	**CONCLUSION & CRITERIA OF STUDY**		**41**
	6.1	Criteria of Future Study	**42**
REFERENCES			**43**

LIST OF FIGURES

3.3.1	Hitmap relationship of Glucose column with other columns	**15**
3.3.2	Accuracy of model on missing data handling	**16**
3.4.1	Preprocessed dataset visualization	**19**
3.5.1	Multi-level normalization on our dataset	**21**
3.5.2	Classes of skewness	**22**
3.5.3	Skewness of our Dataset	**22**
3.5.4	Visualization of the skewness of 3 columns	**23**
3.5.5	Accuracy of models with and without normalization	**23**
4.1	The basic phases of KED	**24**
4.3.1	ANN model of our project	**30**
4.4.1	Support Vector Classifier	**31**
4.4.2	Working model of boosting Classifier	**33**
4.4.3	Use of machine learning algorithm for our best case model	**33**
5.1	Confusion Matrix function	**38**
5.2	Max Voting Code	**38**
5.3	Confusion Matrix	**39**
5.4	Accuracy of Individual Model	**39**
5.5	Best case accuracy	**39**

LIST OF TABLES

1.4.1	TOTAL NUMBER OF CASE OF DIABETES IN DIFFERENT COUNTRIES	**3**
1.4.2	GLOBAL PHENOMENON ABOUT THE DIABETES DISEASE	**3**
3.1.1	Dataset description and characteristics	**13**
3.4.1	Binning of age for class creation (adding new feature)	**17**
3.4.2	Binning of glucose for class creation (adding new feature)	**18**
3.4.3	Binning of diastolic blood pressure (adding new feature)	**18**
3.4.4	Binning of BMI for class creation (adding new feature)	**18**
4.5.1	ANN parameters	**34**
4.5.2	ML algorithms' parameters	**35**
5.1	Result of other research on this tropic	**40**

Chapter 1

INTRODUCTION

1.1 Problem Definition

The disease or condition which is continual or whose effects are permanent is a chronic condition. These types of diseases affected quality of life, which is major adverse effect. Diabetes is one of the most acute diseases, and is present worldwide. A major reason of deaths in adults across the globe includes this chronic condition. Chronic conditions are also cost associated. A major portion of budget is spent on chronic diseases by governments and individuals [1,2]. The worldwide statistics for diabetes in the year 2013 revealed around 382 million individuals had this ailment around the world [3]. It was the fifth leading cause of death in women and eight leading cause of death for both sexes in 2012. Higher income countries have a high probability of diabetes [4]. In 2017, approximately 451 million adults were treated with diabetes worldwide. It is projected that in 2045, almost 693 million patients with diabetes will exist around the globe and half of the population will be undiagnosed. In addition, 850 million USD were spent on patients with diabetes in 2017 [5]. Diabetes can be the reason of many diseases. On highlighting, some of them are cerebrovascular diseases, coronary heart diseases, peripheral neuropathy, eye damage, kidney diseases (even permanent kidney damage), foot diseases (like: diabetic foot) and so on. Prediction of diabetes at an early stage can lead to improved treatment. Even proper treatment at early stage can free a patient completely from this deadly disease.

A recent survey shows that, about 72.4 million people of India and 114.4 million people of China are the patient of Diabetes. This report is the survey of 'International Diabetes Federation, 2017'. Though Bangladesh isn't listed there, but two neighbors of Bangladesh, India & China present the maximum number of diabetes patients and from the presumption, it can be said that Bangladesh is not far from the diabetes prevalence as like it's neighbors. We, also can feel that through the amplitude of the diabetes patients around us.

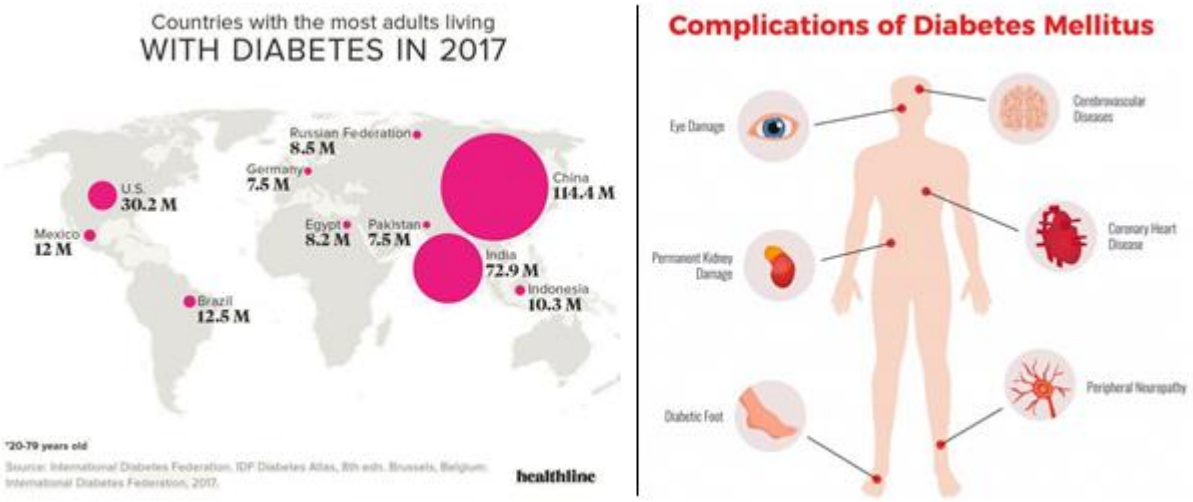


Fig 1.1.1: Survey of diabetes patients in 2017 and category of diseases caused of diabetes

1.2 Researching filed on Diabetes

Research on biological data is limited but with the passage of time enables computational and statistical models to be used for analysis. A sufficient amount of data is also being gathered by healthcare organizations. New knowledge is gathered when models are developed to learn from the observed data using data mining techniques. Data mining is the process of extracting from data and can be utilized to create a decision making process with efficiency in the medical domain [6]. Several data mining techniques have been utilized for disease prediction as well as for knowledge discovery from biomedical data [7,8]. Diagnosis of diabetes is considered a challenging problem for quantitative research. Some parameters like A1c [9], fructosamine, white blood cell count, fibrinogen and hematological indices [10] were shown to be ineffective due to some limitations. Different research studies used these parameters for the diagnosis of diabetes [11–13]. A few treatments have thought to raise A1C including chronic ingestion of liquor, salicylates and narcotics. Ingestion of vitamin C may elevate A1c when estimated by electrophoresis but levels may appear to diminish when estimated by chromatography [14].

1.3 Attributes of Diabetes

Most studies have suggested that a higher white blood cell count is due to chronic inflammation during hypertension [15]. A family history of diabetes has not been associated with BMI and insulin [16]. However, an increased BMI is not always associated with abdominal obesity [17]. A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision making process. There is a need to combine different parameters to effectively predict diabetes at an early stage. Several existing techniques have not provided effective results when different parameters were used for prediction of diabetes [18–19]. In our study, diabetes is predicted with the assistance of significant attributes, and the association of the differing attributes. We examined the diagnosis of diabetes using machine learning approaches along with ANN to make the model for early prediction of this disease.

1.4 Introduction to classifiers of predicting diabetes

There is huge amount of medical data available such as identification of cause and nature of diseases, patient details, resources available for hospitals, availability of doctors and patients. All these data are needed to be analyzed because everybody is seeking knowledge from such vast data with reduced costs. A classifier could be applied on such data for efficient decision making which in turns contribute significant knowledge about a system. Hidden patterns present inside the data, which is required by care givers, patients, and health sector experts could be automatically derived by the application of Data Mining. It enables patients, doctors and everybody in health care industry to make better decisions regarding healthcare and treatment for any disease at early stages. The foremost task of classification is predicting the upcoming behavior of a data sample for instance it can answer whether a person would be having any chances of suffering from diabetes or not, on the basis of hidden patterns on which the classifier is trained. A recent survey of IDF (International Diabetes Federation) shows that more than 70.3 million people in South East Asia are suffering from Diabetes and the number would be increased to 120.9 million by 2030.

The impacts of the disease on adults are increasing in such a way that one out of five people is having diabetes. The following table 1and 2 reveals the situation of diabetes in the countries of South East Asia region as well as worldwide [21].

TABLE 1.4.1. THE TABLE SHOWS THE TOTAL NUMBER OF CASE OF DIABETES IN DIFFERENT COUNTRIES OF SEA REGION

Serial number	**Countries by diabetes cases in SEA**	
	Country	**Cases (in millions)**
1	India	63.0
2	Bangladesh	5.5
3	Sri Lanka	1.1
4	Nepal	0.506
5	Mauritius	0.141
6	Bhutan	0.022
7	Maldives	0.015

TABLE 1.4.2. THE TABLE SHOWS THE GLOBAL PHENOMENON ABOUT THE DIABETES DISEASE

Serial number	**Global figures for diabetes, 2012 (20-79 years)**	
	Global Phenomenon	**Total Count**
1	Prevalence of diabetes in adults	8.3%
2	Number of People with Diabetes	371 million
3	Number of undiagnosed Cases	187 million
4	Deaths due to diabetes	4.8 million
5	Total healthcare expenditures in USD	471.6 billion

The above facts and figures very clearly indicate that the problem of diabetes disease is becoming severe day by day and requires efficient strategic planning specially in Indian scenario. It can be very well achieved with the application of Data Mining. The section II of this paper presents some related works done previously in the field of Data Mining especially in diabetes and section III presents proposed approach of classification followed by results in section IV and in section V concluding remarks with its future perspective has been described. The following table 1and 2 reveals the situation of diabetes in the countries of South East Asia region as.

MOTIVATION

1.5 Inspiration Behind This Plan

DM (Diabetes mellitus, early stage of diabetes) is among the most widespread diseases (World Health Organization, 20202) for the elderly in the country. In 2017, 451 million individuals globally are diabetic as informed by the International Diabetes Federation. Expectations are that this figure will rise to 693 million citizens over the next 26 years. The primary cause of DM remains unclear, but researchers believe that both environmental and genetic factors play an important role in DM. While it is incurable, medications and drugs may be used to control it. Individuals with DM are at danger of having additional health complications, like cardiac arrest and organ damage. Early detection and management with DM will also avoid complications and help to decrease the threat with severe health issues.

Diagnosis of DM may be done either manually by a medical practitioner or by an automatic device. Any of these forms of measurement of DM involve benefits and drawbacks. The main advantage of manual diagnosis is that it does not need any help from the machine for the DM detection procedure, thus allowing the medical professional to be a specialist in the area. Often the symptoms of DM in its initial phase are so low that even an experienced doctor can't fully identify them.

These facts inspire us to do a research on early prediction of diabetes using our computer field. As a result of advances in Machine Learning (ML) and Artificial Intelligence (AI), the disease detection and diagnosis at an initial stage by an automated program is more probable and efficient than the manual DM recognition method [22].

1.6 Objective of The Thesis

This thesis is to design a model for early perdition of diabetes to make better way of increasing awareness, as prediction of diabetes at an early stage can lead to improved treatment. To achieve this objective, we have identified the following specific aims

- ✓ Pre-processing methods that apply to DM datasets, to develop a model of diabetes prediction system which can provide the way of increasing awareness and better understanding of healthcare.
- ✓ Widely utilized ML feature extraction techniques in the area of DM detection and Widely utilized ML-based techniques for detection, classification, and diagnosis of DM to provide a better & friendly system which can help to determine diabetes at early stage so that proper treatment can be ensured in time for anyone.
- ✓ Widely utilized AI-based techniques for intelligent DM assistant for self-management and personalization of DM therapy to develop a better health tracking system which can reduce the time and space complexity of a patient of visiting doctors by providing the health issue about diabetes.
- ✓ Future research directions for research to be resolved by future scientists working in the area of DM detection and diagnosis.

1.7 Overview of the Thesis

Diabetes mellitus(DM) field produces big data based on laboratory valuation, reports about the patient, treatment, follow-ups, medicine, etc. It is difficult to manually assemble all data appropriately. The quality of the organization of data has been affected because of unsuitable data management. Improvement in the data amount requires some suitable way to extract and process data efficiently and effectively. Machines for data collection and inspection are hired in modern and new hospitals to make them able for data collection and share in big information systems. An automated device is capable to identify DM and handle anomalies with far better simplicity and reliability compared to manual detection and diagnosis. Automation of the diagnosis of DM is therefore important. Automated DM systems may be built either by machine learning approaches or by artificial intelligence approaches.

All ML and AI methods have benefits and limitations of their own. Both methods have therefore been used to build automatic DM detection systems. The AI and ML-based techniques need causability and explain ability to perform like a human [23] as many best-performing techniques of Ml and AI are least transparent. The explain ability of AI systems helps to improve the faith of doctors in future AI systems. The causability is based on the causal model which is measured in terms of efficiency, effectiveness related to causal understanding, and its transparency for a user. Several researchers have used ML and AI methods for DM control and self-management and personalization in recent years. Nevertheless, only a few review papers on DM detection and diagnosis procedures have been published [24].

1.8 Contributions

The main contribution of this report is the consideration of both ML and AI-based approaches in DM detection, diagnosis, self-management, and personalization. As we know, this is the first review article that covers both ML and AI for the detection, diagnosis, and self-management of DM and personalization of DM therapy. Review papers are relevant because they summarize the current research in a particular area in a detailed manner. Also, authors have only studied ML procedures, but certain essential facets of ML, like databases, pre-processing methods, and feature extraction and selection approaches used to identify DM and AI solutions to the need for intelligent DM assistants, are not addressed. As a consequence, attempts have been made in the sense of this analysis to examine existing literature on ML and AI approaches to DM studies.

Because of the variety and complexity of DM detection and diagnosis and self-management and personalization systems, a systematic decision-making framework is used for the selection of papers obtained from the Scopus and PubMed database. The purpose of this framework comprises of, (1) Datasets description, (2) Pre-processing techniques, (3) DM feature selection methods, (4) DM detection using ML approaches, (5) Intelligent DM assistant using AI methods and (6) performance matrices.

- We present with the approach of combining neural network & machine learning to determine diabetes at early stage.
- We introduce with a random split approach for neural network at training stage which can help on data training in deep learning approaches and also it increases the performance of model of detecting diabetes.

- We try to introduce with model combining of machine learning and deep learning on max voting, which helps a lot on performance of detecting and this approach is quite popular in machine learning, though it's a new entry in deep learning.

- To the best of my knowledge, it is the first approach of establishing a combine modeling of deep learning and machine learning approaches in this field with the collaboration of these methodologies some of which are so much popular in their implementation in many projects is this field too, such as: **Support Vector Classifier (SVC), Logistic Regression, Random Forest Classifier, K-Neighbors Classifier.**

1.9 Thesis Outline

In the rest of this thesis, we present the details of our approach for building the model for early prediction of diabetes Using ANN along with Machine Learning. Here,

- **Chapter 2:** Presents a survey of some works which have already been taken in action with successful ending & explanation of those works in brief.

- **Chapter 3:** Describes the proposed methodologies used in this thesis with describing about how they collaborate in our model of predicting diabetes at early stage. It also presents the basic information of those methods and implementation of them in it.

- **Chapter 4:** Holds the implementation of our model with the effort of its sections and

- **Chapter 5:** Explain result of this research with comparison of differ models in this field.

Chapter 2

RESEARCHES ON THIS FIELD

Diabetes prediction using Data Mining has been explored by various researchers from time to time and developed encouraging solution for medical expertise and researchers. As a result of all these research, diagnostic and prognostic models have been developed and influenced the existing clinical practices.

In a research work, characteristics of various diagnostic models have been analyzed and some abnormal factors have been identified which may be improved for further clarity in clinical decision making, highlighted by Wyatt and Altman[25].The authors also highlighted the benefits of the Glasgow Coma Scale and specified that confidence, accuracy, effectiveness and interoperability factors, responsible for different situations, are not available up to a certain level which indicates the prime reason of un-usefulness of the approach [26].Apart from the above methodology some authors have specified that the decision tree and C4.5 are also not applicable in every case and for using them, it has to be used in a specific order[27].

The technique of Bayesian approach was also analyzed by authors to show that some amount of reverse engineering was needed to calculate the training proportion of the classification [28]. Some other accurate approaches of Data Mining like Neural Network and Support Vector Machines are supposed to classify appropriately but they fall into categories of "Black Box" [29] [30]. These black box techniques are not suitable since the internal details cannot be understood properly by the researchers and by analyzing the classifier none can understand the core of the problem domain. Among the many statistical approaches available, logistic regression are currently popular used in many medical applications. Although they have solid theoretical foundation, Wyatt and Altman established that one in every five statistical models, the underlying assumptions were violated the integrity of the approach [25]. In sequence, with the research works in the field of diabetes classification, an expert system for predicting the diabetes disease was proposed by working of an expert system that was mechanized on chaining inference depending on backward, forward and forward-backward chaining technique [31] [32]. It suggested an uncertainty principal which calculates the probability of illness and severity of the disease as well as the potential complications of the disease. The dataset of PIMA INDIA was studied by various researchers to develop a certain classification model by using Rapid Miner tool [33]. The author also analyzed how to handle missing values. The impact of preprocessing of diabetes data in artificial neural networking based technique is also examined [34]. The research in the field of diabetes has also been studied based on Association Rules [35] [36]. Some researches in this field are describe bellow with their architecture and outcome.

2.1 A model for classifying diabetic patient control level using data mining [37]

(Assoc. Prof., Department of MIS, Prince Sattam Bin Abdalaziz University, KSA)

In their paper as they say, Diabetes occurs when a blood glucose level is too high and the body couldn't to reduce this level. The diabetic patient doesn't produce any insulin or enough insulin to reduce the high glucose level. Diabetes grows when glucose can't enter the body's cells to be used as energy. Diabetes is an important cause of continued ill health. This disease causes death per year more than HIV SHARDS as one diabetic patient die every 10 seconds. Around 347 million people around the world are suffering from diabetes. In 2010, around 3.4 million people died due to complications of the diabetes and about 80% of diabetes deaths occur in poor countries.

There are two types of diabetes disease which are: Type1 Diabetes is known as insulin dependent. It is categorized by lacking insulin in body and requires insulin. The cause of type 1 is unknown and it is not preventable with current knowledge. Symptoms contain polyuria, polydipsia (thirst), hunger, loss of weight loss, eye problem and weakness. Type2 Diabetes which is formerly called non-insulin-dependent or adult onset. It is as a result of ineffective utilization of the insulin by the cells of the body. 90% of the people with diabetes are categorized type2 patients and mainly due to excess body weight and physical inactivity. In Saudi Arabia, the number of diabetic patients is increasing according to official reports [49]. In 2005, World Health Organization's NCD report of Ministry of Health, Saudi Arabia, identified six types of treatments for diabetes disease: Drug, Diet, Weight reduction, Smoke stop, Exercise and Insulin.

With passage of time, the diabetes can make dangerous complications. Diabetes may lead to heart disease and stroke. About 50% of people with diabetes die of cardiovascular disease. One of the commonly seen complications is foot ulcer and limb amputations, which are mainly due to reduced blood flow and nerve damage (neuropathy) in the feet of diabetic patients. Damage of the small blood vessels of the eye is also common; this may lead to blindness which is due to diabetic retinopathy. Kidney failure of patients, affected by diabetes, is quite common. The overall risk of death among people affected with diabetes is double the risk of their peers without diabetes good measure for that. It should be less than 7%. In this paper a new predicted model has been developed by using data mining techniques. The model aims to classify the diabetic patients into two classes which are: (1) under- control patients who's the HbA1c is less than 7% or out-of-control patients who's the HbA1c is more than 7%. The treatments plan for 10061 diabetic patients were used to build the model. After comprehensive survey for classification techniques, three algorithms have been selected such as NaivaeBayse, Logistic and J48 algorithms. By using WEKA application, the model has been implemented.

Based on results, Logistic algorithm has been selected as best one with high accuracy rate of 74.8%. The can be used to classify new diabetic patients either under-control or out of control based on their treatment plans. This paper has been distributed into several sections which provide complete detail about all the research phases, for example: The Section No. 2 presents the related works and many researchers conducted on the subject. The Section No. 3 contains the proposed

model components, results as well as full results of discussion. Last Section (No. 4) concludes the research and present recommendations for future works.

Diabetes is considered as one of the diseases which cause more deaths than any other disease in the world. To avoid the dangerous complications of the diabetes, patients should control a blood glucose level as the HbA1c (accumulative blood glucose level for 3 months) should be less than 7%. In this paper a new predicted model has been developed by using data mining techniques. The model aims to classify the diabetic patients into two classes which are: under control (HbA1c < 7%) and out of control (HbA1c > 7%). The treatments plan for 10061 diabetic patients were used to build the model. After comprehensive survey for classification techniques, three algorithms have been selected which were NaivaeBayse, Logistic and J48. By using WEKA application, the model has been implemented. Based on the results of experiment, Logistic algorithm has been selected as best one with high accuracy rate of **74.8%.** To enhance the model accuracy, the nutrition system and exercise need to be added to the dataset as future work.

2.2 Data Mining Models Comparison for Diabetes Prediction [38]

(Amina Azrar, Muhammad Awais, Yasir Ali, Khurram Zaheer, Government College University, Faisalabad, Pakistan)

In their paper they say, Knowledge discovery in databases (KDD) is the system of applying data mining algorithms. Knowledge Discovery in Databases (KDD) is common research area for researchers in machine learning, databases, high performance computing, data visualization and knowledge-based systems. The primary steps for data mining include data selection, data preprocessing, data transformation, data mining, and final evaluation (pattern evaluation and pattern recognition). Data Mining is the process of getting meaningful outcomes from any given dataset. Some of the techniques used for data mining include association rules, classification, clustering, Naïve Bayes, Decision Tree and KNN.

A variety of rules can be generated using data mining techniques. Data Mining is useful for Prediction or Description of a few records. Using prediction, we are expecting unknown values of various variables in dataset whilst description specializes in coming across designs that depict the information translated by means of People. Data mining is useful for predicting diseases. Affected person's history, Hospitals, clinical devices and electronic facts offer a lot of records concerning a selected disease. Those datasets are used for extracting useful information by which we are able to take choices and generate rules. Multiple diseases can be diagnosed using data mining methodologies, for example, AIDS and diabetes. This paper is meant to predict diabetes for pregnant women depending on few given attributes. Some major factors that affect the diabetes or may cause its increase in severity include obesity, weight increase or hypertension.

From the past few years, data mining got a lot of attention for extracting information from large datasets to find patterns and to establish relationships to solve problems. Well known data mining algorithms include classification, association, Naïve Bayes, clustering and decision tree. In medical science field, these algorithms help to predict a disease at early stage for future diagnosis.

Diabetes mellitus is the most growing disease that needs to be predicted at its early stage as it is lifelong disease and there is no cure for it. This research is intended to provide comparison for different data mining algorithms on PID dataset for early prediction of diabetes.

The prevalence of diabetes is increasing among young adults and old age people. This paper focuses that the use of data mining algorithms can be very helpful in early prediction and in consequence early precautions before the diagnosis of disease. The main goal of this paper is to provide a comparison and suggest best algorithm which can be used for the pattern recognition or prediction in healthcare fields. These algorithms are of much importance for medical datasets because these algorithms can be used for automatic classification tools which can help doctors or experts for taking necessary steps for any disease before diagnosis. Each of these algorithms can give high accuracy and efficiency depending upon the type of data and attributes. After the implementations of these algorithms it can be said that for PID dataset Decision Tree gives best accuracy **75.65%.** The tool used for testing and validation is Rapid Miner while all algorithms worked with 70:30 ratios for training and testing.

2.3 A model for early prediction of diabetes [39]

(Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb,Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad Awais Malikb, Muhammad Mehdi Razab,Salman Ibrarb, Zunish Abbasd)

Different classification algorithms were applied on our dataset, and results for all techniques were slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of accuracy and the AUROC curve. The accuracy of models was predicted with the help of a confusion matrix. First, the random forest algorithm was applied. Experiments were done to tune the model with respect to the number of decision tress and the maximum depth of the decision trees. In the first iteration, the number of decision trees was 8 and the depth of the trees were 4. Again while tuning the model and increasing the number of trees, the results were effective as compared to prior results. Increasing the number of decision trees could be used to obtain improved results, but when the number of trees reached 50, performance diminished. We obtained a best accuracy of 74.7% and an AUROC curve value of 0.806 when the number of decision trees was 32 and the depth of the decision trees was 4. The AUROC curve obtained by using the random forest method. The complete results of random forest is described and the confusion matrix.

After the random forest algorithm, the ANN was applied to obtain better results. The model was tuned on the basis of number of hidden neurons, number of learning iterations as well as value of initial learning weights. In first iteration, when number of hidden neurons were 50, number of learning iterations were 100 as well as the value of initial learning weights were 0.1, the model has provided satisfactory results. When the values of the tuned parameters were increased, the results worsened. In the 3rd iteration, the values of tuned parameters were decreased; then better results were obtained as compared to the 1st iteration.

In the 4th iteration, results were obtained which were most effective when the number of hidden neurons was 5, the number of learning iterations was 10, and the value of initial learning weights was 0.4. The AUROC curve of ANN, which has a value of 0.816 and an accuracy of **75.7%,** calculated from confusion matrix.

The K-means clustering method was used after the RF and ANN implementation. To apply K-means clustering in our dataset, we normalized the dataset attributes by using the Min-Max normalization technique. Significant attributes were normalized, having the range of 0–1. K-Means clustering was applied by initially setting the value of K = 2, (as in our dataset only two types of patients exist), one for patients with diabetes and the second for patients without diabetes. When the number of clusters was increased, then accuracy decreased. The KMeans clustering predicted 273 to have a value of 1 (positive) and 495 as 0 (Negative). To evaluate the accuracy of K-means clustering, the results were compared to the target class, which shows 203 instances were classified incorrectly, as noted in the confusion matrix. Both clusters, in which circles in the image show the incorrect instances.

Incorrectly classified instances were 26.43% which show that the accuracy of K-means clustering method was 73.6%. Accuracy of the proposed models has been compared. The random forest method provided an accuracy of 74.7%, ANN gave 75.7% and Kmeans clustering method has given 73.6% accuracy. ANN outperforms other methods. ANN is a nonlinear model that is straightforward and used for comparing statistical methods. It is a nonparametric model, while the majority of statistical techniques are parametric and require a higher foundation of statistics. The main benefit of utilizing ANN over other statistical techniques is its capacity to capture the non-linear relationship among the concerned variables. The primary weakness of the random forest method is that numerous trees can make the algorithm slow and inadequate for prediction in real time. This algorithm is quick to train, yet very moderate to make predictions once it is trained. A gradually more precise prediction requires more trees, which results in a slower model. Hence, these are the main reasons leading to ineffective results in our study.

2.4 Prediction of Diabetes using Classification Algorithms [40]

(Deepti Sisodia, Dilip Singh Sisodiab, National Institute of Technology, G.E Road, Raipur and 492001, India)

In their paper they say, Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and non-ketotic hyperosmolar coma. Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot be controlled.

Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors.

The early identification is the only remedy to stay away from the complications. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms works better in diagnosing different diseases. Data Mining and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study. This research work focuses on pregnant women suffering from diabetes.

In their work, Naive Bayes, SVM, and Decision Tree machine learning classification algorithms are used and evaluated on the PIDD dataset to find the prediction of diabetes in a patient. Experimental performance of all the three algorithms are compared on various measures and achieved good accuracy. The remaining of the research discussion is organized as follows: Section-II briefs Related Work of various classification techniques for prediction of diabetes, Section-III describes the Methodology and brief discussion of Dataset used, Section-IV discusses evaluated Results, and Section-V determines the Conclusion of the research work.

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of **76.30 %** using the Naive Bayes classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

Chapter 3

METHODOLOGY

We have used Python to implement the machine learning and deep learning approaches to get our desire goal. Our methodology consists of four steps which are explained below. Our model is considering the impacts of different attributes, present in dataset, for the severity of the diabetics. The dataset we have used is introduced bellow with its each column's Mean, Standard deviation, Type and description.

3.1 Dataset Description

The dataset used in this study, is originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases (publicly available at: UCI ML Repository [20]). The main Objective of using this dataset was to predict through diagnosis whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification. The Pima Indian Diabetes (PID) dataset having: 9 = 8 + 1 (Class Attribute) attributes, 768 records describing female patients (of which there were 500 negative instances (65.1%) and 268 positive instances (34.9%)). The detailed description of all attributes is given in Table 3.

TABLE 3.1.1. Dataset description and characteristics

Attribute Name	Attribute Description	Mean ± SD	Types
PREGNANCIES	Number of times a woman got pregnant	3.8 ± 3.3	int64
GLUCOSE (MG/DL)	Glucose concentration in oral glucose tolerance test for 120 min	120.8 ± 31.9	int64
BLOODPRESSURE (MMHG)	Diastolic Blood Pressure	69.1 ± 19.3	int64
SKINTHICKNESS (MM)	Fold Thickness of Skin	20.5 ± 15.9	int64
INSULIN (MU U/ML)	Serum Insulin for 2 h	79.7 ± 115.2	int64
BMI (KG/M2)	Body Mass Index (weight/(height)^2)	31.9 ± 7.8	float64
DIABETESPEDIGREEFUNCTION	Diabetes pedigree Function	0.4 ± 0.3	float64
AGE	Age (years)	33.2 ± 11.7	int64
OUTCOME	Class variable (class value 1 for positive 0 for Negative for diabetes)	- - - - - -	int64

3.2 Data Splitting

Our dataset is impacted in a file containing all the entries for each column, that means there is no separate train, test section in it. So, we have to perform some steps to separate the dataset into train, test sections before going to take further more steps. To split the dataset, we have to perform into two different ways, one of which is random split and the other one is actual split. These terms are explained as:

3.2.1 Actual Split

We first split out dataset in 80:20 ratios, where train set contains 80%, and test set contains 20% of full data. We call this actual split as we have made the dataset to split into specific sections which data is also specific by us. We then build all our further steps on the base of this train-test set. We have finished even the modeling using the machine learning algorithms on the base of this set. It gives us 73% combine accuracy of 7 machine learning algorithms which we explained bellow.

3.2.2 Random Split

When we have finished our one kind of project with 73% accuracy using the actual split for train-test set, what we have already explained, we have made the same procedural of data processing model with only the change that this time we use a random split function of python's sklearn library to get randomized train-test set for each run of the model. We have run our initial project on this random splitting procedure for more than 30 times, where we have saved the random splitting set before even getting the accuracy for the set. Using this technique, we get some randomized set gives around 80% accuracy, even the dataset gives 77% accuracy without any data preprocecing. Then we have fixed the best accuracy set and done our further mode steps to increase our model's accuracy, like proper transformation, missing data dandling, adding neural network approach and so on.

We have done all our data processing steps along with data transformation and modeling on the best train-test set finding by random split and taken that as our desire set for our entire modeling as like actual split.

3.3 Data Cleaning

Data Cleaning means the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is considered a foundational element of the basic data science.

Data is the most valuable thing for Analytics and Machine learning. In computing or Business data is needed everywhere. When it comes to the real world data, it is not improbable that data may contain incomplete, inconsistent or missing values. If the data is corrupted, then it may hinder the process or provide inaccurate results. Let's see some examples of the importance of data cleaning.

Suppose are a general manager of a company. our company collects data of different customers who buy products produced by our company. Now we want to know on which products people are interested most and according to that we want to increase the production of that product. But if the data is corrupted or contains missing values then we will be misguided to make the correct decision and we will be in trouble.

At the end of all, Machine Learning is a data-driven AI. In machine learning, if the data is irrelevant or error-prone then it leads to an incorrect model building. As much as we make our data clean, as much as we can make a better model. So, we need to process or clean the data before using it. Without the quality data, it would be foolish to expect anything good outcome.

Data cleaning consists of filling the missing values and removing noisy data. Noisy data contains outliers which are removed to resolve inconsistencies [41]. In our dataset, Glucose, blood Pressure, skin thickness, insulin, BMI have zero (0) values, but they are not true case in those features as no human body can survive without glucose or insulin, any alive person must have blood pressure and there must be no human without skin, so those zero values are representing only the null entries and nothing else.

Thus, all the zero values of those columns were replaced with the median value of that attribute where the median has a hitmap relationship with the corresponding values of other columns. For example: A zero value of (Glucose column & X no. row) is replaced by the result of this formula.

$$\textbf{Value of Glucose [X]} = \frac{\textbf{(median of Glucose Column + Sum of Other Columns' relation)}}{\textbf{2}}$$

$$\textbf{Sum of other Columns' relation = (BMI [X] * 0.20) + (Age [X] * 0.28) + (Insulin [X] * 0.35)}$$

Where Glucose, BMI, Age, Insulin represent the columns in the dataset and [X] represents the row on which the modification is applying and 0.20, 0.28, 0.35 are the corresponding values of the relation between Glucose & BMI, Glucose & Age, Glucose & Insulin in hitmap.

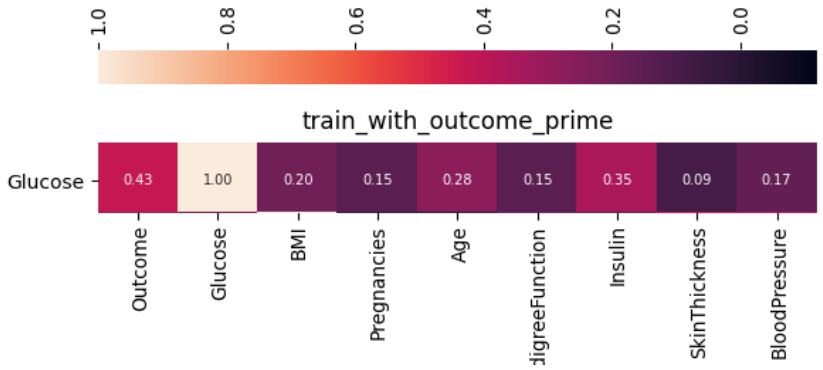


Fig 3.3.1: Hitmap relationship of Glucose column with other columns

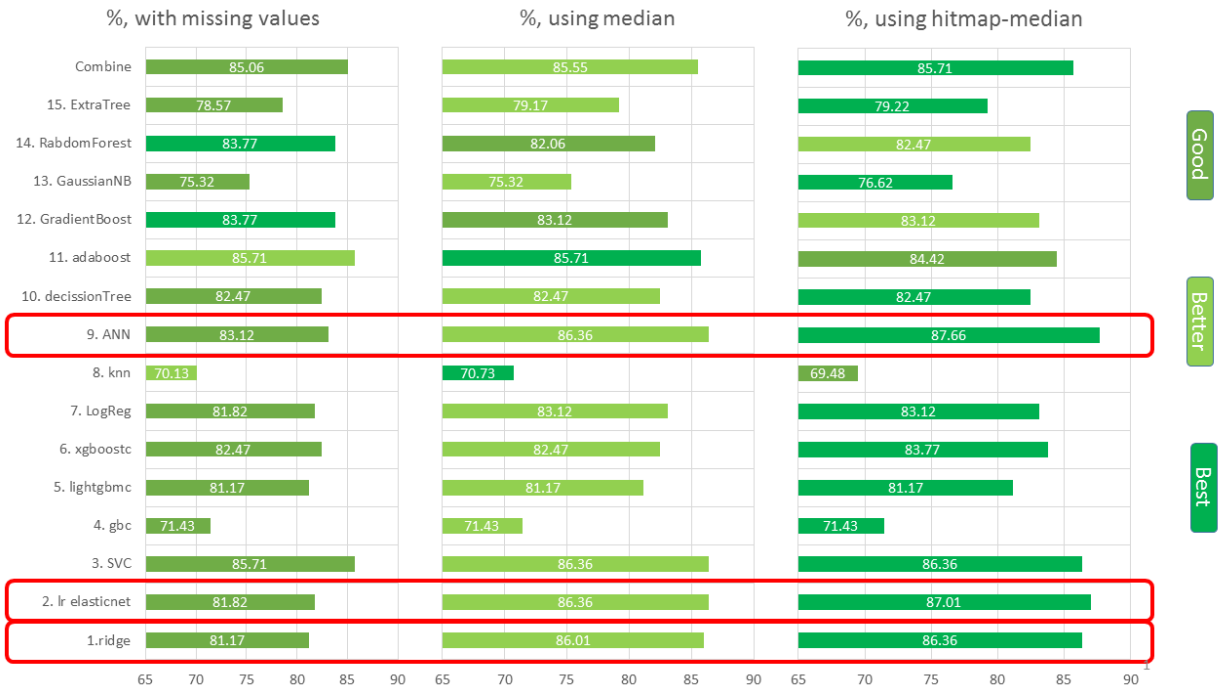


Fig 3.3.2: Accuracy of model on missing data handling

This is the accuracy of models when there is no missing data handling, missing data handling with median only and missing data handling with median and hitmap relationship. It clearly shows that how it is important to handle missing data and sometime it's obliging, like for ANN, Elasticnet, ridge classifier. It also shows that missing data handling with median and hitmap relationship' makes better result on average.

3.4 Feature Engineering

Sometimes the raw data we obtain from various sources won't have the features needed to perform machine learning tasks. When this happens, we must create our own features in order to obtain the desired result. Creating a feature doesn't mean creating data from thin air. we create new features from existing data.

Understanding the need to create features

One great limitation of machine learning algorithms is that it can be impossible to guess a formula that could link our response to the features we're using. Sometimes this inability to guess happens because we can't map the response using the information we have available (meaning that we don't have the right information). In other cases, the information we provided doesn't help the algorithm learn properly.

For instance, if we're modeling the price of real estate properties, the surface of the land is quite predictive because larger properties tend to cost more. But if instead of the surface, we

provide our machine learning algorithm with the length of the sides of the land (the latitude and longitude coordinates of its corners), our algorithm may not figure out what to do with the information we provided. Some algorithms will manage to find the relationship between the features, but most algorithms won't.

The answer to this problem is feature creation. Feature creation is that part of machine learning that is considered more an art than a science because it implies human intervention in creatively mixing the existing features. we perform this task by means of addition, subtraction, multiplication, and ratio to generate new derived features with more predictive power than the originals.

Knowing the problem well and figuring out how a human being would solve it is part of feature creation. So, connecting to the previous example, the fact that land surface connects to the property price is common knowledge. If surface is missing from our features when trying to guess the value of a property, we can recover such information from the existing data — and doing so increases the performance of the predictions.

Data transformation consists of smoothing, normalization, and aggregation of data [42]. For the smoothing of data, the binning method has used. The attribute of age has been useful to classify in five categories, as shown in Table 3.4.1.

Blood glucose concentration in patients who do not have diabetes is different from patients with diabetes. Glucose values have been divided into 5 categories [43] as shown in Table 3.4.2. A strong association has been found between healthy and diabetic patients regarding their blood pressure levels [44]. Blood pressure has been divided into five different categories as shown in Table 3.4.3

The relationship between BMI and diabetes prevalence is consistent. The prevalence of diabetes and obesity is increasing concurrently worldwide. Furthermore, previous studies have shown that BMI is the most important risk factor for type 2 diabetes [45]. BMI values have been categorized into five classes as shown in Table 3.4.4.

For the completion of the preprocessing task, selection of significant attributes and transformation of significant attributes into bins are done after data cleaning. The preprocessed dataset visualization is shown in Fig. 3.4.1.

Table 3.4.1: Binning of age for class creation (adding new feature)

Age (Years)	Age Bins	Class Value
<= 25	Youngest	1
26 - 38	Younger	2
39 - 51	Middle Age	3
52 - 59	Older	4
>= 60	Oldest	5

Table 3.4.2: Binning of glucose for class creation (adding new feature)

Glucose	Glucose Bins	Class Value
<= 60	Very Low	2
61 – 80	Low	1
81 – 140	Normal	3
141 – 180	Early Diabetes	4
>= 181	Diabetes	5

Table 3.4.3: Binning of diastolic blood pressure for class creation (adding new feature)

Blood Pressure	Blood Pressure Bins	Class Value
<= 60	Very Low	1
61 – 75	Low	2
76 – 90	Normal	3
91 – 100	High	4
>= 101	Hypertension	5

Table 3.4.4: Binning of BMI for class creation (adding new feature)

BMI	BMI Bins	Class Value
<= 18.5	Starvation	1
18.6 – 25.0	Normal	2
25.1 – 30.0	Over weight	3
30.1 – 40.0	Obese	4
>= 40.1	Very Obese	5

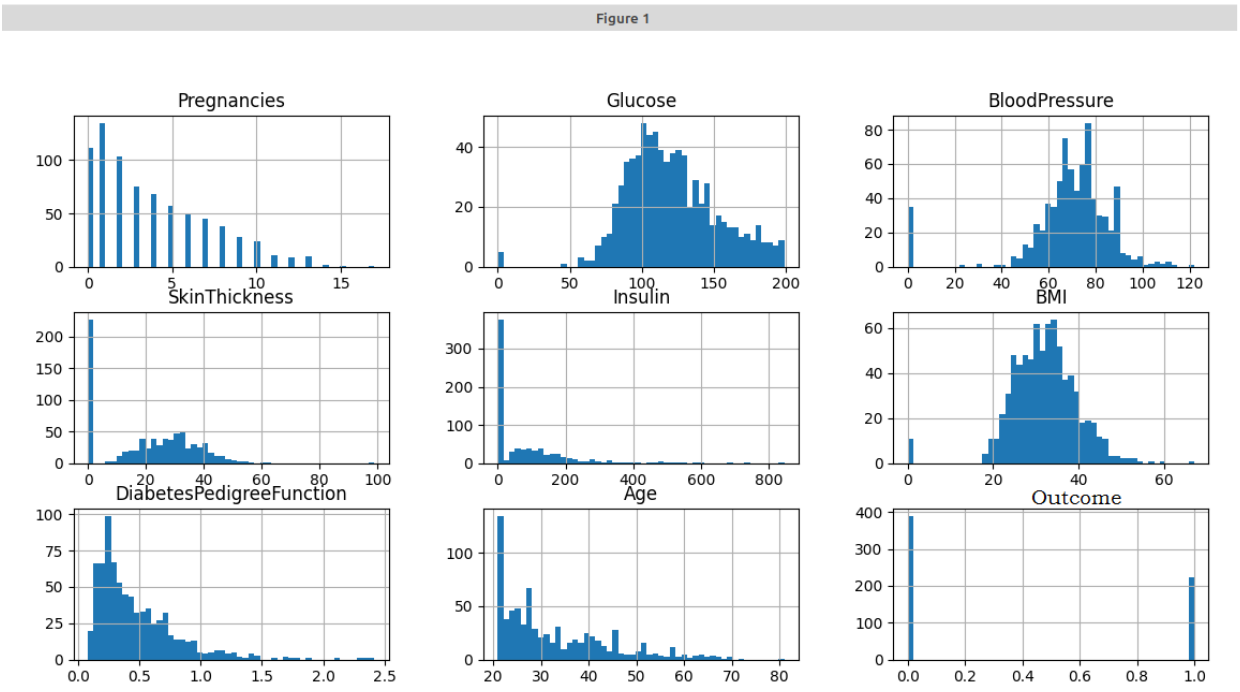


Fig 3.4.1: Preprocessed dataset visualization

3.5 Data reduction (Normalization)

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it. It's also important to apply feature scaling if regularization is used as part of the loss function (so that coefficients are penalized appropriately).

Data reduction obtains a reduced representation of the dataset that is much smaller in volume yet produces the same (or almost the same) result.

Dimensionally reduction has been used to reduce the number of attributes in a dataset [46]. The principal component analysis method was used to extract significant attributes from a complete dataset. Glucose, BMI, diastolic blood pressure and age were significant attributes in the dataset.

We have to do multi-level normalization using log1p and sqrt functions of python's numpy library.

3.5.1 Logarithm (log):

Addition, multiplication, and exponentiation are three of the most fundamental arithmetic operations. Addition, the simplest of these, is undone by subtraction: when we add 5 to x to get x + 5, to reverse this operation we need to subtract 5 from x + 5. Multiplication, the next-simplest operation, is undone by division: if we multiply x by 5 to get 5x, we then can divide 5x by 5 to return to the original expression x. Logarithms also undo a fundamental arithmetic operation, exponentiation. Exponentiation is when we raise a number to a certain power. For example, raising 2 to the power 3 equals 8:

$2^3 = 2 * 2 * 2 = 8$

The general case is when we raise a number b to the power of y to get x:

$b^y = x$

The number b is referred to as the base of this expression. The base is the number that is raised to a particular power—in the above example, the base of the expression $2^3 = 8$ is 2. It is easy to make the base the subject of the expression: all we have to do is take the y-th root of both sides. This gives:

$b = \sqrt[y]{x}$

It is less easy to make y the subject of the expression. Logarithms allow us to do this:

$y = \log_b x$

This expression means that y is equal to the power that we would raise b to, to get x. This operation undoes exponentiation because the logarithm of x tells we the exponent that the base has been raised to.

3.5.2 log1p():

For real-valued input, log1p is accurate also for x so small that 1 + x == 1 in floating-point accuracy. Logarithm is a multivalued function: for each x there is an infinite number of z such that exp(z) = 1 + x. The convention is to return the z whose imaginary part lies in [-pi, pi]. For real-valued input data types, log1p always returns real output. For each value that cannot be expressed as a real number or infinity, it yields nan and sets the invalid floating point error flag.

For complex-valued input, log1p is a complex analytical function that has a branch cut [-inf, -1] and is continuous from above on it. log1p handles the floating-point negative zero as an infinitesimal negative number, conforming to the C99 standard.

log1p() in Python. numpy. log1p(arr, out = None, *, where = True, casting = 'same_kind', order = 'K', dtype = None, ufunc 'log1p') : This mathematical function helps user to calculate natural logarithmic value of x+1 where x belongs to all the input array elements

3.5.3 sqrt():

sqrt() function is an inbuilt function in Python programming language that returns the square root of any number. Syntax: math. sqrt(x) Parameter: x is any number such that x>=0 Returns: It returns the square root of the number passed in the parameter. Error: When x<0 it does not execute due to a runtime error.
Fig 3.5.1 shows the operations what we have taken to normalize our dataset.

```
all_data['SkinThickness']           = np.log1p(all_data['SkinThickness'])
all_data['PregClass']               = np.log1p(all_data['PregClass'])
all_data['Glucose']                 = np.log1p(all_data['Glucose'])
all_data['BPClass']                 = np.log1p(all_data['BPClass'])
all_data['BMI']                     = np.log1p(all_data['BMI'])

all_data['DiabetesPedigreeFunction']= np.sqrt(np.log1p(np.log1p(all_data['DiabetesPedigreeFunction'])))
all_data['Insulin']                 = np.sqrt(np.log1p(np.log1p(all_data['Insulin'])))
all_data['Age']                     = np.sqrt(np.log1p(np.log1p(all_data['Age'])))
all_data['AgeClass']                = np.log1p(np.log1p(all_data['AgeClass']))
```

Fig 3.5.1 Multi-level normalization on our dataset.

3.5.4 Skewness:

Using these functions, we have reduced the skewness of the columns of our dataset, where skewness refers that how the data in a column of a dataset is distributed. 0 skewness is the best case for any data distribution which refers that the data is perfectly distributed and so no normalization is necessary for that. On the other hand, positive value of skewness refers that the data is lefty distributed and negative value of skewness refers that the data is rightly distributed.

Consider the two distributions in the figure just below. Within each graph, the values on the right side of the distribution taper differently from the values on the left side. These tapering sides are called tails, and they provide a visual means to determine which of the two kinds of skewness a distribution has:

3.5.4.1 Negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left, despite the fact that the curve itself appears to be skewed or leaning to the right; left instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data. A left-skewed distribution usually appears as a right-leaning curve.

3.5.4.2 Positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right, despite the fact that the curve itself appears to be skewed or leaning to the left; right instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a left-leaning curve.

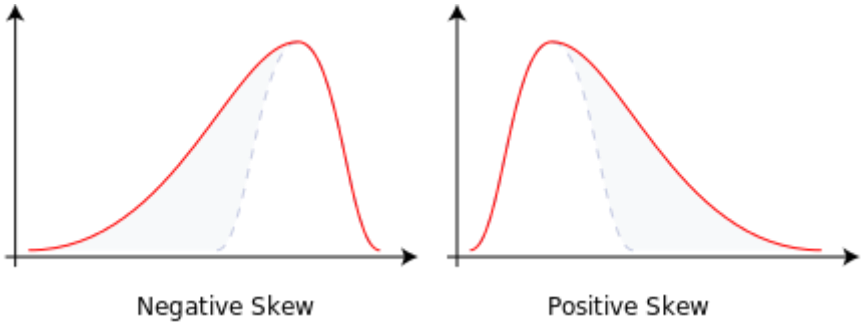


Fig 3.5.2: Classes of skewness

Skewness in a data series may sometimes be observed not only graphically but by simple inspection of the values. For instance, consider the numeric sequence (49, 50, 51), whose values are evenly distributed around a central value of 50. We can transform this sequence into a negatively skewed distribution by adding a value far below the mean, which is probably a negative outlier, e.g. (40, 49, 50, 51). Therefore, the mean of the sequence becomes 47.5, and the median is 49.5. Based on the formula of nonparametric skew, defined as (μ - ν)/σ the skew is negative. Similarly, we can make the sequence positively skewed by adding a value far above the mean, which is probably a positive outlier, e.g. (49, 50, 51, 60), where the mean is 52.5, and the median is 50.5.

As mentioned earlier, a unimodal distribution with zero value of skewness does not imply that this distribution is symmetric necessarily. However, a symmetric unimodal or multimodal distribution always has zero skewness. Fig 3.5.3 shows the skewness of our dataset before normalization as prime skewness and after normalization as final skewness and Fig 3.5.4 shows visualization of the skewness of columns Skin Thickness, BMI & Diabetes Pedigree Function before normalization as all_data Before transformation and after normalization as all_data after transformation. It also shows that we have not done any normalization on Pregnancies column.

```
~~~~~~~~~~ prime skewness ~~~~~~~~~~
Pregnancies                 0.90
Glucose                     0.51
BloodPressure              -0.08
SkinThickness               1.10
Insulin                     2.86
BMI                         0.59
DiabetesPedigreeFunction    1.92
Age                         1.13
AgeClass                    0.97
GlucoClass                 -0.37
BPClass                     0.43
BMIClass                   -0.38
PregClass                   0.60
dtype: float64
~~~~~~~~~~ ~~~~~~~~~~~~~~~ ~~~~~~~~~~

~~~~~~~~~~ final skewness ~~~~~~~~~~
Pregnancies                 0.90
Glucose                    -0.09
BloodPressure              -0.08
SkinThickness              -0.38
Insulin                     0.60
BMI                        -0.03
DiabetesPedigreeFunction    0.27
Age                         0.49
AgeClass                    0.12
GlucoClass                 -0.37
BPClass                    -0.23
BMIClass                   -0.38
PregClass                   0.14
dtype: float64
~~~~~~~~~~ ~~~~~~~~~~~~~~~ ~~~~~~~~~~
```

Fig 3.5.3: Skewness of our Dataset

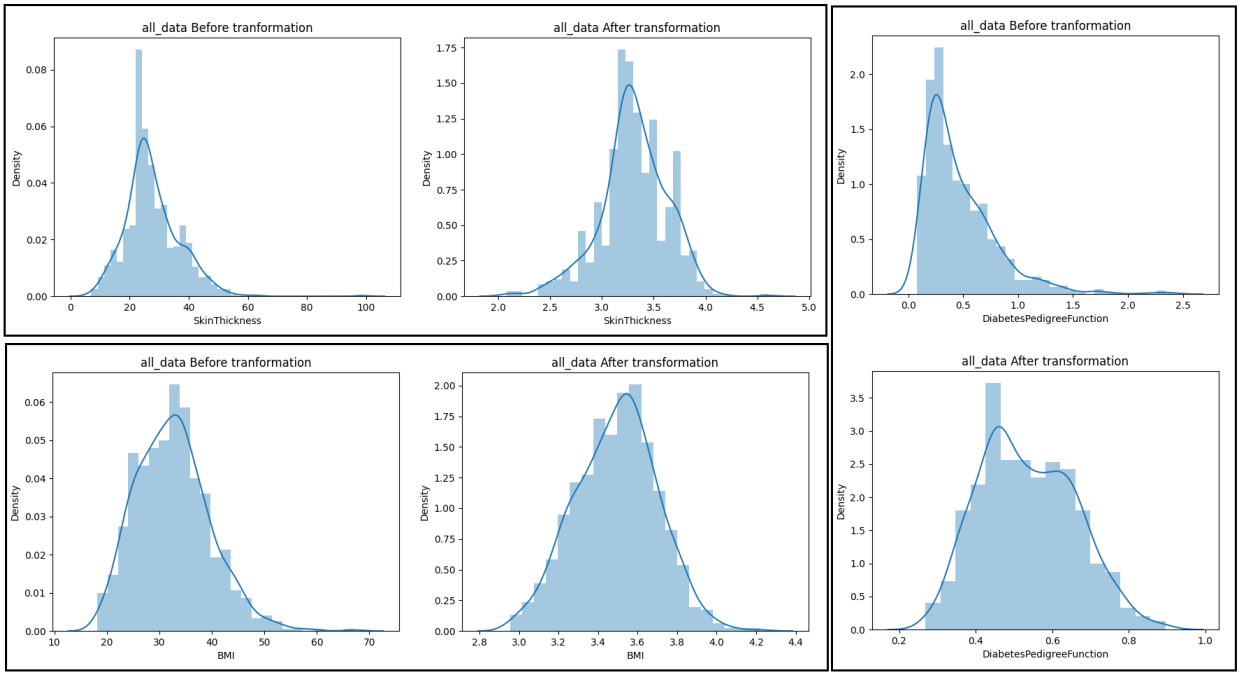


Fig 3.5.4: Visualization of the skewness of 3 columns

Fig 3.5.5: Accuracy of models with and without normalization

This is the accuracy of models without normalization and with normalization. Most of the time it is better to have normalization and even it can make big change in performance, like ANN, but sometime like knn, it can reduce the performance. As I have to decide to keep ANN in my model and most of the other architectures work better with normalization, so I have kept normalization and ignored knn.

Chapter 4

NETWORK MODELING

Machine Learning(ML) and Artificial Intelligence(AI) is a research discipline that deals with the way computers learn from experience. To certain researchers, the phrase "ML" is the part of "AI," provided that the capacity to learn is the coarse attribute of an intellectual individual. The goal of machine learning is to develop computer systems that can learn and respond from their prior observation. The goal of artificial intelligence is to develop an intelligent agent or assistant that use different machine learning techniques-based solution [47].

Knowledge exploration in databases (KEDs) is a discipline that incorporates hypotheses, approaches, and strategies, tries to understand the data and derive valuable facts from it. It is known to be a multi-step method (selection, pre-processing, transformation, ML/AI, understanding/assessment) defined in Fig. 4.1. The most critical phase in the whole KED method is ML/AI, which exemplifies the use of ML algorithms and AI in the data processing.

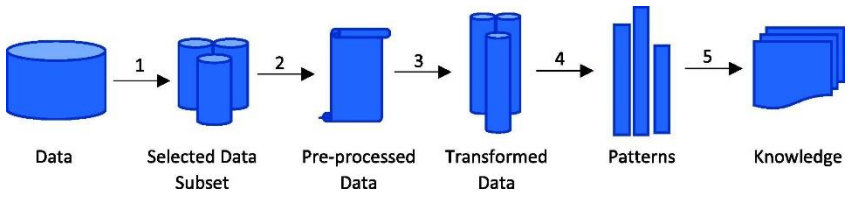


Fig. 4.1: The basic phases of KED: (1) Selection of data, (2) Pre-processing of data, (3) Transformation of data, (4) ML/AI, (5) Understanding/Assessment.

4.1 Machine Learning(ML) Procedures

ML processes are usually categorized into three specific groups. These include 1) Supervised Learning (SL), where the scheme indicates the functionality of the labeled training data; 2) Unsupervised Learning (UL), where the system attempts to deduce the nature of unidentified data; and 3) Reinforcement Learning (RL), where the machine communicates with a dynamic context. Artificial intelligence is used to develop intelligent assistants which will help in self-management and personalization of the disease therapy.

4.1.1 Supervised learning(SL)

SL is where an algorithm is used to learn the mapping function from the input (I) to the output (O)=f(I). The purpose is to determine the mapping function so accurately that the output variables (O) for system can be predicted when a new input data (I) occurs (Taherkhani et al., 2018)

There are two categories of learning processes in SL: classification and regression. Classification methods aim to simulate distinct classes, like genotypes, whereas regression methods forecast real values. Some of the most common methods are KNN, DT, NN, GA, and SVM.

4.1.2 Unsupervised learning(UL)

UL is where there are only input data (I) with no associated output variables. The objective of UL is to simulate the basic nature or distribution of data to understand more about the content. Procedures are left to their devices to explore and display the fascinating content structure (Krotov and Hopfield, 2019).

There are two types of learning processes in supervised learning: association and clustering. The clustering is where the underlying groupings are revealed in the data, while the association rule learning method discovers the rules that define significant portions of the data.

4.1.3 Reinforcement Learning(RL)

RL is a generic term offered to a class of strategies through which the method aims to improve by active contact with the world to optimize any idea of incremental reward. It is necessary to remember that the program has no previous awareness of the actions of the world and the only way to realize is by trial and error (Silver et al., 2018). This type of learning is particularly applicable to autonomous devices, because of their flexibility about their environment.

4.2 Artificial intelligence

R Artificial intelligence (AI), is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals. Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving".

As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect. A quip in Tesler's Theorem says "AI is whatever hasn't been done yet." For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology. Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems (such as chess and Go), autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

4.3 Network Architecture

In our model we have tried to train our model in Supervised learning(SL) way for machine learning part and then combine that with a popular model of Artificial intelligence called Artificial neural network (ANN).

4.3.1 Artificial neural network (ANN)

The Artificial neural network (ANN) is a research area of artificial intelligence and an important technique which is used in data mining. The ANN has three layers: input, hidden, and output layer. The hidden layer consists of units that transform the input layer to the output layer. The output of one neuron works as the input for another layer. ANN detects complex patterns and learns on the basis of these patterns. The human brain contains billions of neurons. These cells are connected to other cells by axons and a single neuron is called as perceptron. Input is accepted by dendrites which is taken as stimuli. Similarly, the ANN is composed of multiple nodes that are connected with each other. The connection between units is represented by a weight. The objective of ANN is to convert input into significant output. Input is the combination of a set of input values that are associated with the weight vector, where the weight can be negative or positive. There is a function that sums the weight and maps the result to the output, such as $\mathbf{y = w_1x_1 + w_2x_2}$. The influence of a unit depends on the weighting; where the input signal of neurons meets is called the synapse. ANN works for both supervised and unsupervised learning techniques. Supervised learning was used in our study because the output is given to the model. In supervised learning, both input and output are known. After processing, the actual output with compared with required outputs. Errors are then back propagated to the system for adjustment. During training, the data is processed many times, so that the network can adjust the weights and refine them [48].

In our model, we configure our ANN model in a sequential model using 8 dense layers, 8 activation layers, 8 batchNormalization layers & 3 dropout layers. We also use one flatten layer in that. Fig. 4.3.1 shows the ANN model which we have used for our project.

4.3.1.1 Dense Layer

A dense layer is just a regular layer of neurons in a neural network. Each neuron receives input from all the neurons in the previous layer, thus densely connected. The layer has a weight matrix W, a bias vector b, and the activations of previous layer a.

The dense layer is a neural network layer that is connected deeply, which means each neuron in the dense layer receives input from all neurons of its previous layer. The dense layer is found to be the most commonly used layer in the models.

In the background, the dense layer performs a matrix-vector multiplication. The values used in the matrix are actually parameters that can be trained and updated with the help of backpropagation.

The output generated by the dense layer is an ‘m’ dimensional vector. Thus, dense layer is basically used for changing the dimensions of the vector. Dense layers also apply operations like rotation, scaling, translation on the vector.

4.3.1.2 Activation layer

Activation layer (function) decides, whether a neuron should be activated or not by calculating weighted sum and further adding bias with it. The purpose of the activation function is to introduce non-linearity into the output of a neuron.

Neural network has neurons that work in correspondence of weight, bias and their respective activation function. In a neural network, we would update the weights and biases of the neurons on the basis of the error at the output. This process is known as back-propagation. Activation functions make the back-propagation possible since the gradients are supplied along with the error to update the weights and biases.

The **activation** function is a mathematical "gate" in between the input feeding the current neuron and its output going to the next **layer**. It can be as simple as a step function that turns the neuron output on and off, depending on a rule or threshold.

In The process of building a neural network, one of the choices you get to make is what activation function to use in the hidden layer as well as at the output layer of the network. This article discusses some of the choices. Elements of a Neural Network: -

Input Layer: - This layer accepts input features. It provides information from the outside world to the network, no computation is performed at this layer, nodes here just pass on the information(features) to the hidden layer.

Hidden Layer: - Nodes of this layer are not exposed to the outer world, they are the part of the abstraction provided by any neural network. Hidden layer performs all sort of computation on the features entered through the input layer and transfer the result to the output layer.

Output Layer: - This layer brings up the information learned by the network to the outer world.

We use softmax as our output layer and LeakyReLU as input and hidden layer for our final model.

4.3.1.3 BatchNormalization layer

To increase the stability of a neural network, batch normalization normalizes the output of a previous activation layer by subtracting the batch mean and dividing by the batch standard deviation. Training deep neural networks with tens of layers is challenging as they can be sensitive to the initial random weights and configuration of the learning algorithm.

One possible reason for this difficulty is the distribution of the inputs to layers deep in the network may change after each mini-batch when the weights are updated. This can cause the learning algorithm to forever chase a moving target. This change in the distribution of inputs to layers in the network is referred to the technical name "internal covariate shifts."

Batch normalization is a technique for training very deep neural networks that standardizes the inputs to a layer for each mini-batch. This has the effect of stabilizing the learning process and dramatically reducing the number of training epochs required to train deep networks. So, batch

normalization is a technique designed to automatically standardize the inputs to a layer in a deep learning neural network.

Once implemented, batch normalization has the effect of dramatically accelerating the training process of a neural network, and in some cases improves the performance of the model via a modest regularization effect.

4.3.1.4 Dropout layer

Dilution (also called Dropout) is a regularization technique for reducing overfitting in artificial neural networks by preventing complex co-adaptations on training data. It is an efficient way of performing model averaging with neural networks. The term dilution refers to the thinning of the weights. The term dropout refers to randomly "dropping out", or omitting, units (both hidden and visible) during the training process of a neural network. Both the thinning of weights and dropping out units trigger the same type of regularization, and often the term dropout is used when referring to the dilution of weights.

Dilution is usually split in weak dilution and strong dilution. Weak dilution describes the process in which the finite fraction of removed connections is small, and strong dilution refers to when this fraction is large. There is no clear distinction on where the limit between strong and weak dilution is, and often the distinction is meaningless, although it has implications for how to solve for exact solutions.

Sometimes dilution is used for adding damping noise to the inputs. In that case, weak dilution refers to adding a small amount of damping noise, while strong dilution refers to adding a greater amount of damping noise. Both can be rewritten as variants of weight dilution.

These techniques are also sometimes referred to as random pruning of weights, but this is usually a non-recurring one-way operation. The network is pruned, and then kept if it is an improvement over the previous model. Dilution and dropout both refer to an iterative process. The pruning of weights typically does not imply that the network continues learning, while in dilution/dropout, the network continues to learn after the technique is applied.

4.3.1.5 Parameter Handling Function

In ANN architecture we also use some functions to maintain & manage parameters in runtime.

- **ModelCheckpoint** The Callback is used in conjunction with training using model.fit() to save a model or weights (in a checkpoint file) at some interval, so the model or weights can be loaded later to continue the training from the state saved. A few options this callback provides include, whether to only keep the model that has achieved the "best performance" so far, or whether to save the model at the end of every epoch regardless of performance. The frequency it should save at. Currently, the callback supports saving at the end of every epoch, or after a fixed number of training batches.

- **Early stopping** : A problem with training neural networks is in the choice of the number of training epochs to use. Too many epochs can lead to overfitting of the training dataset, whereas too few may result in an under fit model. Early stopping is a method that allows you to specify an arbitrary large number of training epochs and stop training once the model performance stops improving on a hold out validation dataset.

- **ReduceLROnPlateau** Is a class Reduce learning rate when a metric has stopped improving. Models often benefit from reducing the learning rate by a factor of 2-10 once learning stagnates. This callback monitors a quantity and if no improvement is seen for a 'patience' number of epochs, the learning rate is reduced.

```
model = Sequential()

model.add(Dense(dns[0],input_shape=(ann_train.shape[1],), kernel_regularizer=regularizer))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())

model.add(Dense(dns[1]))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())
model.add(Dropout(rate = dropout[0]))

model.add(Dense(dns[2]))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())

model.add(Dense(dns[3]))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())
model.add(Dropout(rate = dropout[1]))

model.add(Dense(dns[4]))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())

model.add(Dense(dns[5]))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())

model.add(Dense(dns[6]))
model.add(LeakyReLU(alpha=alpha_lrelu)) if (leakyRelu) else model.add(Activation('relu'))
model.add(BatchNormalization())
model.add(Dropout(rate = dropout[2]))

model.add(Flatten())

model.add(Dense(2))
model.add(Activation('softmax'))
```

.

Fig 4.3.1: ANN model of our project

4.4 Machine Learning approaches

In machine learning section we have used multiple algorithms (12+) to determine which algorithms perform better for our dataset and then conclude with 5 algorithms which give us the best performance among the algorithms of machine learning. Here, we introduce with the 5 algorithms which are taken in final model of our project. Fig. 4.4.3 shows the models we have used and also the models we have used for our best case (Un hashing)

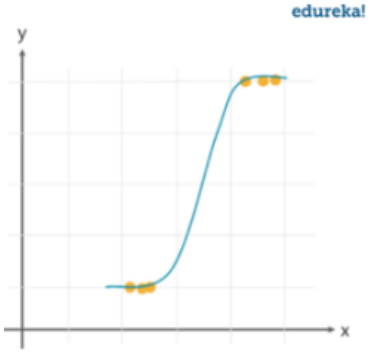
4.4.1 Logistic Regression & Ridge Classifier

The **Ridge Classifier**, based on Ridge regression method, converts the label data into [-1, 1] and solves the problem with regression method. The highest value in prediction is accepted as a target class and for multiclass data multi-output regression is applied.

Ridge regression is a variation of Linear Regression. It belongs a class of regression tools that use L2 regularization. The other type of regularization, L1 regularization, limits the size of the coefficients by adding an L1 penalty equal to the absolute value of the magnitude of coefficients. This sometimes results in the elimination of some coefficients altogether, which can yield sparse models. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will not result in sparse models.

Logistic Regression is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. This is better than other binary classification like nearest neighbor since it also explains quantitatively the factors that lead to classification.

Advantages and Disadvantages



Logistic regression is specifically meant for classification; it is useful in understanding how a set of independent variables affect the outcome of the dependent variable. The main disadvantage of the logistic regression algorithm is that it only works when the predicted variable is binary, it assumes that the data is free of missing values and assumes that the predictors are independent of each other. It's used in identifying risk factors for diseases, word classification, weather prediction, voting applications.

4.4.2 Support Vector Classifier (SVC)

The support vector machine is a classifier that represents the **training data as points in space** separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.

More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outlier's detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

Whereas the original problem may be stated in a finite-dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products of pairs of input data vectors may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function k(x,y) selected to suit the problem. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant, where such a set of vectors is an orthogonal (and thus minimal) set of vectors that defines a hyperplane. The vectors defining the hyperplanes can be chosen to be linear combinations with parameters a_i of images of feature vectors x_i that occur in the data base. With this choice of a hyperplane, the points x in the feature space that are mapped into the hyperplane are defined by the relation $\sum_i \alpha_i k_i(x_i,x)$ = constant. Note that if k(x,y) becomes small y grows further away from x, each term in the sum measures the degree of closeness of the test point x to the corresponding data base point x_i. In this way, the sum of kernels above can be used to measure the relative nearness of each test point to the data points originating in one or the other of the sets to be discriminated. Note the fact that the set of points x_i mapped into any hyperplane can be quite convoluted as a result, allowing much more complex discrimination between sets that are not convex at all in the original space.

Advantages and Disadvantages

It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces. The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

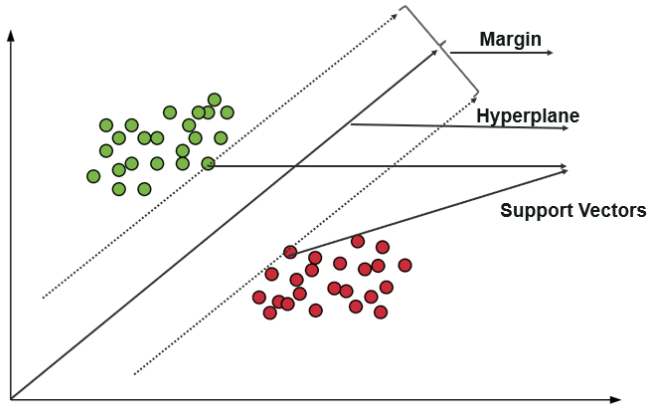


Fig 4.4.1: Support Vector Classifier

4.4.3 XGBoost & AdaBoost Classifier

In recent years, boosting algorithms gained massive popularity in data science or machine learning competitions. Most of the winners of these competitions use boosting algorithms to achieve high accuracy. These Data science competitions provide the global platform for learning, exploring and providing solutions for various business and government problems. Boosting algorithms combine multiple low accuracy (or weak) models to create a high accuracy (or strong) models. It can be utilized in various domains such as credit, insurance, marketing, and sales. Boosting algorithms such as AdaBoost, Gradient Boosting, and XGBoost are widely used machine learning algorithm to win the data science competitions.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

Ada-boost or Adaptive Boosting is one of ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations. Any machine learning algorithm can be used as base classifier if it accepts weights on the training set. Adaboost should meet two conditions:

1. The classifier should be trained interactively on various weighed training examples.

2. In each iteration, it tries to provide an excellent fit for examples by minimizing training error.

It works in the following steps:

4 Initially, Adaboost selects a training subset randomly.
5 It iteratively trains the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training.
6 It assigns the higher weight to wrong classified observations so that in the next iteration these observations will get the high probability for classification.
7 Also, it assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight.
8 This process iterates until the complete training data fits without any error or until reached to the specified maximum number of estimators.
9 To classify, perform a "vote" across all of the learning algorithms you built.

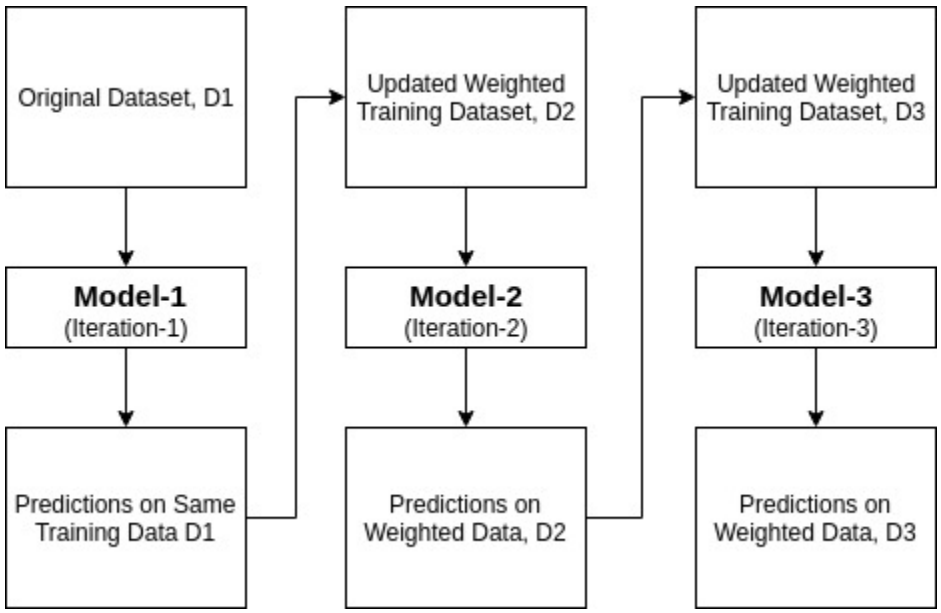


Fig 4.4.2: Working model of boosting Classifier

```
{
  'ridgec'        :   model_database.ridgec,                    # RidgeClassifierCV
  'lr_elasticnet' :   model_database.lr_elasticnet,             # LogisticRegression(penalty = 'elasticnet')
  'svc'           :   model_database.svc,                       # SVC ( 'C': 0.7678, 'penalty': 'l1' )
   'gbc'          :    model_database.gbc,                       # GradientBoostingClassifier
   'lightgbmc'    :    model_database.lightgbmc,                 # LGBMClassifier
  'xgboostc'      :   model_database.xgboostc,                  # XGBClassifier
  'LogReg'        :   model_database.LogisticRegression,        # LogisticRegression
   'knn'          :    model_database.KNeighborsClassifier,      # KNeighborsClassifier
   'SVC2'         :    model_database.SVC2,                      # SVC ( 'C': 1.7, 'kernel': 'rbf' )
   'decissionTree' :   model_database.DecisionTreeClassifier,    # DecisionTreeClassifier
  'adaboost'      :   model_database.AdaBoostClassifier,        # AdaBoostClassifier
   'GradientBoost' :    model_database.GradientBoostingClassifier,  # GradientBoostingClassifier
   'GaussianNB'   :    model_database.GaussianNB,                # GaussianNB
   'RabdomForest' :    model_database.RandomForestClassifier,    # RandomForestClassifier
   'ExtraTree'    :    model_database.ExtraTreesClassifier       # ExtraTreesClassifier
  }
```

Fig 4.4.3: Use of machine learning algorithm for our best case model

4.5 Training Procedure

We have trained our model using ANN and machine learning algorithms and combine their results using max voting approach. On max voting we have kept priority of the best accuracy holder algorithm, which is ANN (87.66% accuracy) for our case. Machine learning algorithms have also done well as ridge and svc classifier give 86.36 % accuracy. On training those models we always change the parameters and weights of the algorithms to get the effectiveness of those on model and to know how the parameters effect on result. Some parameters of those models are given and also the parameters of ANN are introduced bellow.

4.5.1 ANN model training parameters

TABLE 4.5.1: ANN parameters

Functions	Parameters
ModelCheckpoint	monitor=monitor, verbose=2, save_best_only=True, save_weights_only=False, mode='auto', period=1
EarlyStopping	monitor=monitor, patience= 15, verbose=2, mode='auto'
ReduceLROnPlateau	monitor='val_loss', factor=0.2, patience=5, min_lr=0.001
Dense layers (top to down)	[32, 64, 64, 128, 128, 128, 256]
Dropout layers (top to down)	[0.10, 0.15, 0.20]

4.5.2 Machine Learning Algorithms' parameters

A common parameter of most of the ML algorithms is random state which in 42 for our best case model. Some Other parameters are as:

TABLE 4.5.2: ML algorithms' parameters

Algorithm	Parameters
1. RidgeClassifierCV	Kfold(n_splits=10, shuffle=True), alphas
2. LogisticRegression(penalty = 'elasticnet')	solver = 'saga', max_iter=1e7, L1_ratio=0.05,
3. Support Vector Classifier (SVC)	C=20, gamma=0.0003,
4. GradientBoostingClassifier	learning_rate=0.01, loss='exponential'
5. LGBMClassifier	objective='binary', verbose=-1
6. XGBClassifier	learning_rate=0.01, seed=27,
7. LogisticRegression	'penalty': 'L1', 'solver': Liblinear
8. KNeighborsClassifier	'n_neighbors': 4, 'metric': 'minkowski', 'p':2
9. Support Vector Classifier (SVC) 2	'C': 1.7, 'kernel': 'rbf', 'probability':True
10. DecisionTreeClassifier	criterion: gini, max_depth:3, max_features:2
11. AdaBoostClassifier	'learning_rate': 0.04, 'n_estimators': 150
12. GradientBoostingClassifier	'learning_rate': 0.01, 'n_estimators': 100
13. RandomForestClassifier	'n_estimators': 15, 'criterion': 'gini'
14. StackingCVClassifier	Classifiers = MLs', meta_classifier=xgboostc

We have used some machine learning approaches multiple time with different parameters and this gives a proper view of how parameter choosing is important in learning algorithms.

Number 4 and 12 of 'All model accuracy' are using only one approach which is Gradient boosting, but the accuracy of them have a huge difference. Though they are using same approach the parameters of them are different and only that creates the difference between their accuracies. For best case I have used those models which have given more than 83.12% individual accuracy, even ignored some which were performed quite equally.

Chapter 5

RESULTS AND DISCUSSION OF THIS STUDY

Different classification algorithms were applied on our dataset, and results for all techniques were slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of accuracy. The accuracy of models was predicted with the help of confusion matrix which we have implemented as a function for our model and we call it for individual and also combine result to check the accuracy of models which is shown in Fig. 5.3. Formula of calculating accuracy is, ((TP + TN) / (TP + FP + FN + TN)) * 100%.

We first applied machine learning algorithms where we got 67% accuracy on combining models and 72% for RidgeClassifierCV(highest). From second to 15 rounds of model running Support Vector Classifier (SVC), LogisticRegression(penalty = 'elasticnet‘), RidgeClassifierCV, GaussianNB, RandomForestClassifier kept 75%+ accuracy individually. Most of the models worked better in our model and at final moment all gives 75%+ accuracy except GradientBoostingClassifier & KNeighborsClassifier. Fig. 5.2 shows the accuracy of individual algorithm in the final model. We use also ANN model along with those machine learning algorithms. We always truck the parameter store carefully so that no parameter store file can be lost. ANN model gives us 87.66% accuracy at final case and it also holds the best individual accuracy for our model. On combining we get 88.31% accuracy by max voting among ANN, RidgeClassifierCV, LogisticRegression(penalty = 'elasticnet‘), Support Vector Classifier, XGBClassifier, LogisticRegression and AdaBoostClassifier. Fig. 5.3 shows the best accuracy of our model.

The ANN model was tuned on the basis of number of hidden neurons, number of learning iterations as well as value of initial learning weights. In first iteration, when 3 dense layers were used along with 1 batchNormalization, 2 activation layers and 1 dropout as well as the value of initial learning weights were 0.1, the model has provided satisfactory results. When the values of the tuned parameters were increased, the results worsened. In the 3rd iteration, the values of tuned parameters were decreased; then better results were obtained as compared to the 1st iteration. In around 18th iteration, results were obtained which were most effective when the number of dense layer was increased to 7 as well as 5 activation layers were added along with 6 batchNormalization and 3 dropout layer, where the value of initial learning weights was 0.4. Furthermore, research introduces us with LukyReLU as our activation layer and we got 87.66% accuracy from ANN at around 25 attempts.

ANN is a nonlinear model that is straightforward and used for comparing statistical methods. It is a nonparametric model, while the majority of statistical techniques are parametric and require a higher foundation of statistics. The main benefit of utilizing ANN over other statistical techniques is its capacity to capture the non-linear relationship among the concerned variables. This algorithm is quick to train, yet very moderate to make predictions once it is trained. A gradually more precise prediction requires more trees, which results in a slower model. Hence, these are the main reasons leading to ineffective results in our study.

```
#Forming a confusion matrix to check our accuracy
def accuracy_calculator(model_name, y_pred, Y_true):
    pp = []
    for p in y_pred:
        if p>0.5:
            pp.append(1)
        else:
            pp.append(0)

    y_pred_f = pp
    cm=confusion_matrix(Y_true,y_pred_f)
    acc = (cm[0][0]+cm[1][1])/(cm[0][0]+cm[0][1]+cm[1][0]+cm[1][1])*100
    print('{0} model accuracy : {1:.2f} %'.format(model_name, acc))
    print('\n~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~')
    return y_pred_f, acc
```

Fig: 5.1: Confusion Matrix function

```
#print('best_acc_index =',best_acc_index)
# Max voting among predictions
result = np.array([])
for i in range(0, len(m_predict[0])):
    try:
        result = np.append(result, mode([clm[i] for clm in m_predict]))
    except:
        result = np.append(result, m_predict[best_acc_index][i])
return result
```

Fig 5.2: Max Voting Code

We want to mention that, though voting is not a good idea in prediction system, according to some scholars, but as I try to merge many approaches which are closed enough on result accuracy and show better performance among them so that I have to use voting. We know voting in better than fixed selection when the outcome is unknown and the number of voters is quite big, as like our government system.

In this case, individual approach gives individual result where we don't know the actual one. In that situation voting can perform better by making votes using those approaches which are closed enough on result accuracy and show better performance among them and select the result which is voted by maximum. The voting was performed only on result of approaches It surely works better for not only this model but also those which have used multiple approaches and result them independently. It is also ensured that the number of approaches are more than two (more than 5 is better) and the results accuracy of individuals are close enough otherwise, ultimately, max voting can reduce the combine accuracy. and there was no effect of it on any prediction of individual.

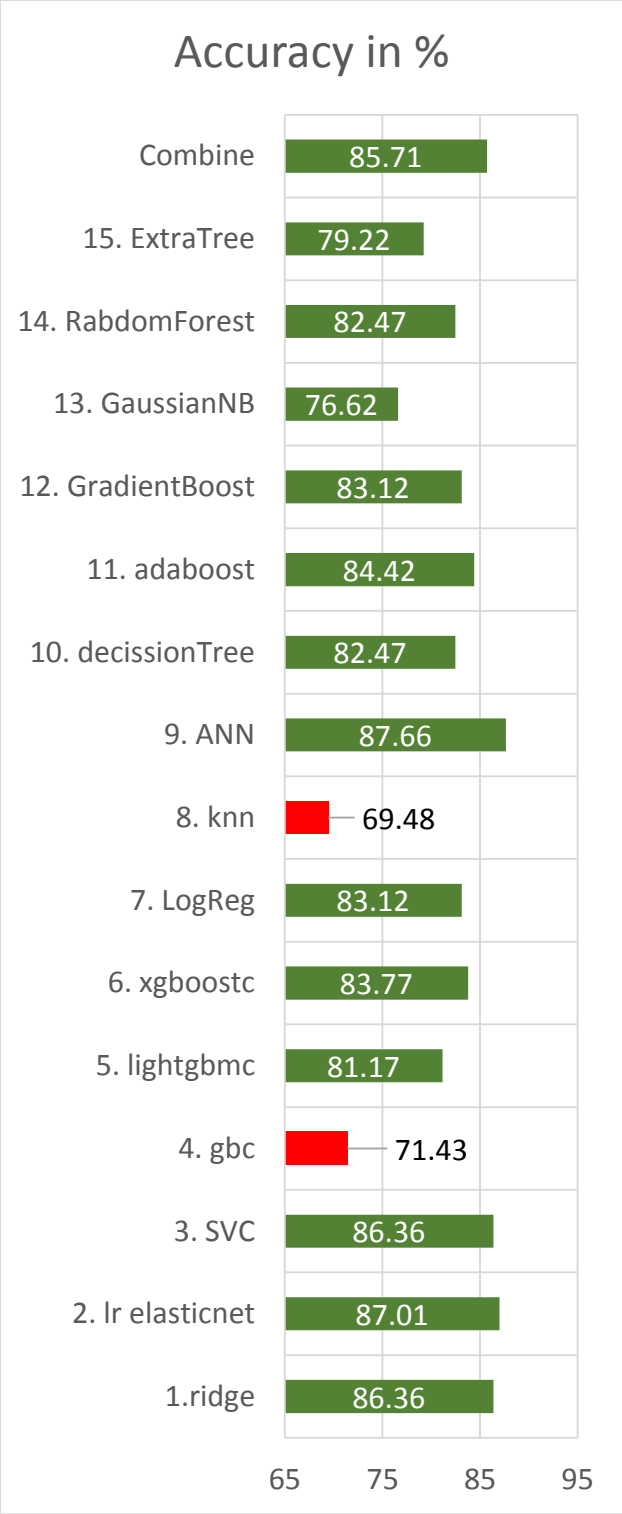


Fig: 5.4: Accuracy of Individual Model

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Fig: 5.3: Confusion Matrix

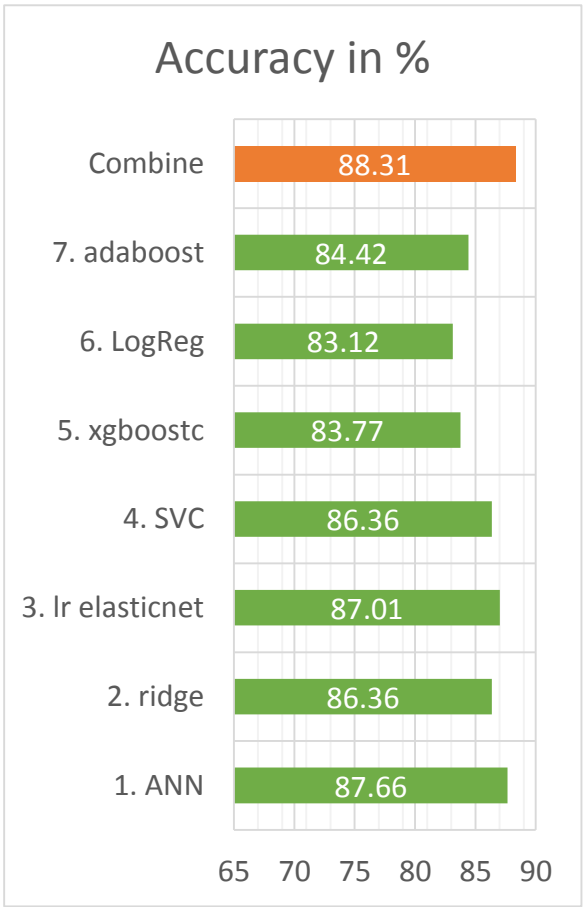


Fig: 5.5: Best case accuracy

Our project gives **88.31%** combine accuracy. On single model perspective ANN gives the best result for us **87.66%** where **87.10%** accuracy is provided by SVC and ridge Classifier algorithms of Machine Learning. Accuracy & architectures of some other papers which we have studied in our research to achieve our goal are shown below:

TABLE 5.1: Result of other research on this tropic

SN #	Paper	Used Approaches	Accuracy
1	USING DATA MINING TO DEVELOP MODEL FOR CLASSIFYING DIABETIC PATIENT CONTROL LEVEL BASED ON HISTORICAL MEDICAL RECORDS	WEKA, Logistic algorithm	74.8%
2	Data Mining Models Comparison for Diabetes Prediction	NaiveBayes, KNN, DecisionTree	75.65%
3	A model for early prediction of diabetes	ANN, RandomForest, K-means	75.70%
4	Prediction of Diabetes using Classification Algorithms	DecisionTree, SVM, NaiveBayes	76.30%

5.1 Limitations of The Study

Although the present study has yielded findings of pedagogical importance, the design of this study was not without flaws. The first limitation concerns the research design. This study was exploratory and lacked a control group, thus limiting the generalization of the results. The second limitation is rooted in the assessment method. Since most of the data were coded and analyzed from qualitative data, the results may reflect in part the way in which the data were collected and analyzed. Some other limitations which can be introduced are:

1. To run ANN with higher epochs & dense size, it requests higher configure environment, but our environment & platform of implementation were not as much high as we needed.
2. Our dataset is not as big as expectation for neural network. Holding only 768 records of which 500 negative instances and 268 positives with only 8 classes of features, our dataset has created a tough challenge for us to study on it for data processing.
3. Another limitation is that we have short time on researching with those models and expect that furthermore study with them can present more than 95% accuracy.
4. The last limitation what we feel is that we have a lacking of the knowledge of parameter handling & choosing them, which can be a huge area of future study for us.

Chapter 6

CONCLUSION & CRITERIA OF FUTURE STUDY

As stated by Martinez (1998), if no mistakes are made, then almost certainly no problem solving is taking place. This project is not out of the context. We may have many issues and bugs in our research, but through our perspective view it is the best achievement of us till now and wish that it also will be same for many learners as well as researchers.

Machine learning and data mining techniques are valuable in disease diagnosis. The capability to predict diabetes early, assumes a vital role for the patient's appropriate treatment procedure. In this paper, a few existing classification methods for medical diagnosis of diabetes patients have been discussed on the basis of accuracy. A classification problem has been detected in the expressions of accuracy. 14 machine learning techniques were applied on the Pima Indians diabetes dataset in total, as well as trained and validated against a test dataset. The results of our model implementations have shown that ANN outperforms the other models. The final result is a combination of ANN and the best 6 algorithms of machine learning for our case.

Considerable part of human population is under the grip of diabetes which is incurable. If not managed well, diabetes can lead to health hazards. Hence, early detection of diabetes is extremely crucial. Nerve damages caused by diabetes, affect the working of the heart. In the proposed work, the dataset of the National Institute of Diabetes and Digestive and Kidney Diseases (publicly available at: UCI ML Repository [20]), is analyzed to diagnose diabetes using deep learning technique (ANN) along with machine learning approaches. The maximum accuracy value of 88.31% was obtained for the combining model of ANN & some machine learning algorithms, where 87.66% accuracy is given by ANN and linear regression of machine learning algorithm holds the second best accuracy 87.01% along with ridge & SVC for their third position on accuracy rank for 86.36%. Our non-invasive, flexible and reproducible system can serve as a reliable tool to clinicians to detect diabetes. Further improvement in accuracy can be obtained using a very large sized input dataset. The potential of deep learning is so tremendous that it can take a big stride in future to the so far challengingly difficult area of anomaly prediction from the anomaly detection if sufficiently large sized input data is available for research. The anomaly prediction can be tried from the input data which may not have anomaly by extracting dynamic characteristics from the input data. The predicted information can serve as a warning signal for the patient as well as the doctor to take sufficient control and precautionary measures.

The limitation of this study is that a structured dataset has been selected but in the future, unstructured data will also be considered, and these methods will be applied to other medical domains for prediction, such as for different types of cancer, psoriasis, and Parkinson's disease. Other attributes including physical inactivity, family history of diabetes, and smoking habit, are also planned to be considered in the future for the diagnosis of diabetes.

6.1 Criteria of Future Study

1. In future study, more neural network approaches can be implemented in it like googlenet, resnet to achieve higher accuracy and we believe that it can be 95%+ accuracy holder if it gets proper research on it.

2. Our dataset is not as much big as neural network desires. In future work, it can be introduced with expected dataset of it which may contain more than 10 thousand entries, then we believe it can present with better performance. It is a flexible model, so any dataset of binary classification can be implemented easily on it and we have this plan for our future work on this topic.

3. Though it gives us a great result from our perspective, which is 88.31% accuracy, but we didn't apply this model in practical field. We have the plan of implementation of this model in practical field as like in some medical institutions and even through online, so that it can reach to many who can be benefited from it.

It also can help doctors along with patients by providing a helping hand of diabetes detection at early stage. We wish, it can help many desires and lives by increasing awareness about the deadly disease before the stage when it says 'Sorry! The time is already gone so far'.

REFERENCES

1. Falvo D, Holland BE. Medical and psychosocial aspects of chronic illness and disability. Jones & Bartlett Learning; 2017.
2. Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes 2017;66:241–55.
3. Tao Z, Shi A, Zhao J. Epidemiological perspectives of diabetes. Cell Biochem Biophys 2015;73:181–5.
4. Organization WH. World health statistics 2016: monitoring health for the SDGs sustainable development goals. World Health Organization; 2016.
5. Cho N, Shaw J, Karuranga S, Huang Y, da Rocha Fernandes J, Ohlrogge A, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018;138:271–81.
6. Diwani S, Mishol S, Kayange DS, Machuve D, Sam A. Overview applications of data mining in health care: the case study of Arusha region‖. Int J Comput Eng Res 2013;3:73–7.
7. Alam TM, Awan MJ. Domain analysis of information ExtractionTechniques. Int J Multidiscip Sci Eng 2018;9:1–9.
8. Alam TM, Khan MMA, Iqbal MA, Wahab A, Mushtaq M. Cervical cancer prediction through different screening methods using data mining. Int J Adv Comput Sci Appl 2019;10:388–96.
9. Cobos L. Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat). Endocr Pract 2018;24:43–4.
10. Dorcely B, Katz K, Jagannathan R, Chiang SS, Oluwadare B, Goldberg IJ, et al. Novel biomarkers for prediabetes, diabetes, and associated complications. Diabetes, Metab Syndrome Obes Targets Ther 2017;10:345.
11. Singh PP, Prasad S, Das B, Poddar U, Choudhury DR. Classification of diabetic patient data using machine learning techniques. Ambient communications and computer systems. Springer; 2018. p. 427–36.
12. Negi A, Jaiswal V. A first attempt to develop a diabetes prediction method based on different global datasets. 2016 fourth international conference on parallel, distributed and grid computing. PDGC); 2016. p. 237–41.
13. Murat N, Dünder E, Cengiz MA, Onger ME. The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels. Biomed Res 2018;29:1370–5.
14. Radin MS. Pitfalls in hemoglobin A1c measurement: when results may be misleading. J Gen Intern Med 2014;29:388–94.
15. Merad-boudia HN, Dali-Sahi M, Kachekouche Y, Dennouni-Medjati N. Hematologic disorders during essential hypertension," diabetes & metabolic syndrome. Clinical Research & Reviews; 2019.
16. Sakurai M, Nakamura K, Miura K, Takamura T, Yoshita K, Sasaki S, et al. Family history of diabetes, lifestyle factors, and the 7-year incident risk of type 2 diabetes mellitus in middle-aged Japanese men and women. J. Diabetes Investig. 2013;4:261–8.
17. Paley CA, Johnson MI. Abdominal obesity and metabolic syndrome: exercise as medicine? BMC Sports Sci. Med. Rehabil. 2018;10:7.

18. Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining. Innovations in information, embedded and communication systems (ICIIECS), 2017 international conference on. 2017. p. 1–5.

19. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Inform. Med. Unlocked 2018;10:100–7.
20. M. Lichman, "Pima Indians diabetes database," ed. Center for machine learning and intelligent systems.: UCI Machine Learning repository.

21. http://www.idf.org/sites/default/files/WP_5E_Update_Country.pdf

22. Sharma and Singh, 2018, Afzali and Yildiz, 2018, Theera-Umpon et al., 2019, Zou et al., 2018, Alghamdi et al., 2017
23. (Holzinger et al., 2019) Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review
24. Gupta and Chhikara, 2018, Cruz-Vega et al., 2019, Choudhury and Gupta, 2019, Sun and Zhang, 2019, Spaggiari et al., 2018, Xiong et al., 2018
25. Altman, Douglas G., et al. "Prognosis and prognostic research: validating a prognostic model." BMJ: British Medical Journal 338.7708 (2009): 1432-1435.
26. Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. Wiley-interscience, 2012
27. Wang, Lipo. Data mining with computational intelligence. SpringerVerlag, 2009.
28. Gil-Jiménez, P., et al. "Shape classification algorithm using support vector machines for traffic sign recognition." Computational Intelligence and Bioinspired Systems (2005): 494-497.
29. Thirugnanam, Mythili, et al. "Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach."Procedia Engineering 38 (2012): 1709-1718.
30. Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." Internet Technology And Secured Transactions, 2012 International Conferece For. IEEE, 2012.
31. Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." Innovations in Information Technology (IIT), 2011 International Conference on. IEEE, 2011.
32. Zhou, Xuezhong, et al. "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support." Artificial Intelligence in Medicine 48.2-3 (2010): 139-152.
33. Jianchao Han; Rodriguez, J.C.; Beheshti, M.; , "Diabetes DataAnalysis and Prediction Model Discovery Using RapidMiner," Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on , vol.3, no., pp.96-99, 13-15 Dec. 2008
34. Jayalakshmi, T.; Santhakumaran, A.; , "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial NeuralNetworks," Data Storage and Data Engineering (DSDE), 2010International Conference on , vol., no., pp.159-163, 9-10 Feb. 2010
35. Patil, B.M.; Joshi, R.C.; Toshniwal, D.; , "Association Rule for Classification of Type-2 Diabetic Patients," Machine Learning and Computing (ICMLC), 2010 Second International Conference on , vol., no., pp.330-334, 9-11 Feb. 2010
36. Nuwangi, S. M., et al. "Usage of association rules and classification techniques in knowledge extraction of diabetes." Advanced Information Management and Service (IMS), 2010 6th International Conference on. IEEE, 2010.
37. Using data mining to develop model for classifying diabetic patient control level based on historical medical records (Assoc. Prof., Department of MIS, Prince Sattam Bin Abdalaziz University, KSA)

38. Data Mining Models Comparison for Diabetes Prediction (Amina Azrar, Muhammad Awais, Yasir Ali, Khurram Zaheer, Government College University, Faisalabad, Pakistan)
39. A model for early prediction of diabetes (Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb,Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad Awais Malikb, Muhammad Mehdi Razab,Salman Ibrarb, Zunish Abbasd)
40. Prediction of Diabetes using Classification Algorithms (Deepti Sisodia Dilip Singh Sisodiab, National Institute of Technology, G.E Road, Raipur and 492001, India)
41. Abidin NZ, Ismail AR, Emran NA. Performance analysis of machine learning algorithms for missing value imputation. Int J Adv Comput Sci Appl 2018;9:442–7.
42. Malley B, Ramazzotti D, Wu J T-y. Data pre-processing. Secondary analysis of electronic health records. Springer; 2016. p. 115–41.
43. Egi M, Bellomo R, Stachowski E, French CJ, Hart GK, Hegarty C, et al. Blood glucose concentration and outcome of critical illness: the impact of diabetes. Crit Care Med 2008;36:2249–55.
44. Brunström M, Carlberg B. Effect of antihypertensive treatment at different blood pressure levels in patients with diabetes mellitus: systematic review and metaanalyses. BMJ 2016;352:i717.
45. Menke A, Rust KF, Fradkin J, Cheng YJ, Cowie CC. Associations between trends in race/ethnicity, aging, and body mass index with diabetes prevalence in the United States: a series of cross-sectional studies. Ann Intern Med 2014;161:328–35
46. Liu H, Motoda H. Feature selection for knowledge discovery and data mining vol. 454. Springer Science & Business Media; 2012.
47. Al-Taee et al., 2016a, Al-Taee et al., 2016b, Thompson and Baranowski, 2019, El-Sappagh et al., 2018, Sosale et al., 2020, Chatrati et al., 2020
48. Schalkoff RJ. Artificial neural networks vol. 1. New York: McGraw-Hill; 1997.
49. http://www.who.int/nmh/countries/sau_en.pdf