Breaking Barriers in Diabetes Management: Unleashing the Potential of Machine Learning in Diagnoses and Treatment

A DISSERTATION
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

Master of Science

In

Biotechnology

Submitted by:

Deepak Kumar

2K21/MSCBIO/10

Under the supervision of:

Prof. Yasha Hasija

Professor

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi - 110042

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I Deepak Kumar Roll Number: 2K21/MSCBIO/10, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled –**"Breaking Barriers in Diabetes Management: Unleashing the Potential of Machine Learning in Diagnoses and Treatment"** in the fulfillment of the requirementfor the award of the degree of Master of Science in Biotechnology and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, is an authentic record of my own carried out during the period from January- May 2023, under the supervision of Prof. Yasha Hasija.

The matter presented in this report has not been submitted by me for the award for any otherdegree of this or any other Institute/University. The work has been accepted in SCI/SCI expanded /SSCI/Scopus Indexed Journal OR peer reviewed Scopus Index Conference with the following details:

Title of the Paper: Behavior analysis using machine learning algorithms in healthcare sector
Author Names: Deepak Kumar, Anukriti Yadav and Yasha Hasija
Name of Conference: International Conference on Advancement in Computation & Computer Technologies (IEEE InCACCT-2023)
Conference Date and Venue: 06th May at Chandigarh University, Gharuan, Mohali (Punjab) -India
Registration: Done
Status of Paper: Accepted
Date of Paper Communication: 1st March 2023
Date of Paper Acceptance: 5 March 2023

Date: Deepak Kumar

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College Of Engineering)
Bhawana Road, Delhi-110042

Certificate

I hereby certify that the Project Dissertation titled **“Breaking Barriers in Diabetes Management: Unleashing the Potential of Machine Learning in Diagnoses and Treatment”** which is submitted by **Deepak Kumar** (**2K21/MSCBIO/10**), Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or anydiploma to this university or elsewhere.

Place: Delhi
Date :

Prof. Yasha Hasija	**Prof. Pravir Kumar**
(Supervisor)	**Head of Department**
Professor	**Dean (International Affairs)**
Department of Biotechnology	Department of Biotechnology
Delhi Technological University	Delhi Technological University

Acknowledgement

I would like to express my gratitude towards my supervisor, Prof . Yasha Hasija, for giving me the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has motivated to carry out the research and to present my work works as clearly as possible. It was a great privilege and honor to work and study under her guidance. I am extremely grateful for what he has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future.

I would also like to thank to all PhD scholars of complex systems and genome informatics laboratory, Jaishree Meena, and Neha Kumari without them I wouldn’t be able to complete my thesis.

A special thanks goes to my best friends Sandeep Kumar, Anukriti Yadav, Kanhaiya Nayak and Kanica Malhotra for their moral support, tolerance and help from the beginning to the end.

I would also like the institution Delhi Technological University, Delhi for giving me the opportunities throughout the tenure of study.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Deepak Kumar

Abstract

The public's health is seriously threatened by diabetes, a lifelong metabolic disorder that affects people all over the world. Diabetic patients require innovative approaches to care as conventional methods struggle to diagnose them precisely and tailor treatments to each individual. It would be wonderful if a few straightforward tests could determine if someone has diabetes. With an easy, quick test, individuals might be able to receive a diagnosis earlier, enabling them to make healthy lifestyle adjustments and reducing their likelihood of contracting new illnesses. A ML-based approach to diabetes management can reduce barriers to diagnosis and treatment by reducing diagnostic and treatment errors. Machine learning algorithms offer unprecedented proficiency when it comes to predicting diabetes and identifying it based on immense volumes of patient data. By using sophisticated data processing tools, we analyze a variety of variables, such as genetic predisposition, lifestyle choices, and clinical signs. High-risk individuals can be identified using ML models in order to prevent diabetes. The condition threatens a person's life because it can damage the heart, kidneys, and nerves. Further, ML-based diagnostic techniques can assist in identifying and treating various diabetes subtypes, improving treatment effectiveness. ML-driven therapy algorithms optimise insulin doses, food suggestions, and exercise routines by taking into account individual variances and dynamically responding to changing conditions, Consequently, they are less likely to suffer complications and their blood sugar levels are improved. In addition, ML-based decision support systems provide real-time insights and suggestions to healthcare providers, enabling proactive and knowledgeable therapeutic treatments.

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADA American Diabetes Association

ANN Artificial Neural Network

AUC Area Under the Curve

BMI Body Mass Index

BPA Bisphenol-A

CDC Centers for Disease Control and Prevention

CART Classification and Regression Trees

DM Diabetes Mellitus

GDM Gestational Diabetes Mellitus

EHR Electronic Health Record

FBGL Fasting Blood Glucose Levels

LR Logistic Regression

NB Gaussian Naïve Bayes

PCA Principal Component Analysis

RF Random Forest

SVM Support Vector Machine

T1DM Type 1 Diabetes Mellitus

T2DM Type 2 Diabetes Mellitus

Chapter 1

1.1 INTRODUCTION

Hyperglycemia, or elevated blood glucose levels, are a hallmark of diabetes, a long-term metabolizing condition. It is brought about by a lack in the hormone insulin, which the pancreas generates and controls blood glucose concentrations [1]. Only with the help of insulin can the body create and use glucose. Type 1 and type 2 diabetes are the two varieties that are most prevalent.

Type 1 diabetes, or juvenile diabetics, occurs as a result of the body's immune system killing beta cells, which produce insulin. Diabetes type 1 patients require insulin injections for blood sugar control.

Type 2 diabetes occurs when the body is incapable of producing sufficient insulin to maintain a healthy blood glucose balance[2]. Adults more often develop this kind of diabetes and the condition is often linked to obesity, inactivity, and poor eating habits.

Gestational Diabetes: Pregnant women with never-before-existing diabetes develop this type of diabetes. The hormone changes that occur during pregnancy can impair insulin sensitivity, leading to hyperglycemia. After giving birth, pregnancy-related diabetes usually goes away, but it raises the likelihood of acquiring type 2 diabetes later on[3]. For diabetes to be prevented from causing health problems like cardiac arrest, renal dysfunction, blindness, and nerve deterioration, it is necessary to manage it correctly. Treatment for diabetes usually includes medication, insulin therapy, blood sugar monitoring, and improvements in life style, like the frequent exercise and a balanced diet[4]. Blood sugar monitoring is a crucial part of diabetes care since it allows keeping a close eye on blood glycemic levels and improving lifestyle habits and medications as necessary. In conventional blood sugar monitoring methods, glucose levels are measured by self-administered finger-stick assays. However, this approach can be time-consuming and uncomfortable, and it might not provide a reliable picture of an individual's blood sugar over time.

Machine learning can revolutionize diabetes management by offering precise and individualized diagnosis and treatment strategies. Data from a large number of patients could be analyzed by ML algorithms to uncover patterns along with connections missed by humans. The algorithms are also able to predict future results and identify the people who may have difficulty in the near future[5]. Diabetes can lead to blindness due to diabetic retinopathy, a condition that can be predicted with ML algorithms. Algorithms have been developed to analyze retinal scans for signs of diabetic retinopathy, helping clinicians identify patients who may require more testing or treatment. Another study used machine learning techniques to analyze data from continuous glucose monitors (CGMs), which detect glucose levels in real time. In diabetics, these algorithms are highly accurate at predicting hypoglycemic episodes, or low blood sugar occurrences, so that they can take precautions. Diabetes control demands regular assessment in addition to control of

glycemic levels, which is a complex and difficult process[6]. Through the use of machine learning, diabetes can be managed more precisely and individually by providing more precise diagnoses and treatment strategies. It is possible for machine learning algorithms to better identify and avoid complications by analyzing vast databases of patient information in order to identify patterns and connections that humans may not be able to recognize.

1.2 Purpose

This project aims to identify the categorization model or algorithm that offers the highest degree of accuracy. Diabetes can be predicted using the algorithm that has been proven to be the best for determining whether an individual is diabetic or not. Using an optimal classification method or model will prevent any misunderstandings that may result from using a less-than-optimal one. In India as well as abroad, it dominates the chronic disease landscape. Predicting it in advance could make it easier to limit and control with a healthy diet or less aggressive treatment.

1.3 Research Question

1. How may machine learning techniques be used to identify diabetes mellitus?
2. How can the suggested machine learning model for diabetes mellitus detection and diagnosis be compared?
3. What machine learning model would be the most effective for detecting diabetes mellitus?

1.4 Motivation and Problem statement

One of the worst chronic health conditions with preventable outcomes is diabetes. A high blood sugar concentration is primarily responsible for this resulting from inadequate insulin synthesis. About 13 million men and 12.5 million women have type 2 diabetes but are not diagnosed in time. These delays in clinical diagnosis make it difficult to fully exploit the advantages of early therapies, which include addressing hyperglycemia, modifying one's lifestyle, and eliminating cardiovascular risk factors[7]. Taking control of diabetes requires additional assistance and education, which patients must receive in order to improve quality of life. Data from EHRs could be used for better patient health outcomes by anticipating diabetes mellitus more quickly to improve patient outcomes[8]. By doing this, diagnostic turnaround times could be shortened and treatment could be initiated sooner.

In India, the number of persons with diabetes is projected to be 50.9 million. This study evaluates various machine learning techniques in order to determine which one is most effective at identifying diabetes. The primary objective of this work is to create algorithmic-based models using medical data from people with and without diabetes.

The dataset file extension is .csv (Comma Separated Values). Dataframe libraries like Pandas are a built-in part of Python, which is used to import the file into the Python environment. Datasets are divided into:

1) Training set

2) Testing Sets,

followed by analysis of them. There are various machine learning techniques, including Naive Bayes, XgBoost, Logistic Regression, Artificial Neural Networks (ANN), and Random Forest. We aim to determine best and most effective algorithm for predicting diabetes and analyze it. Therefore, early prediction can lead to an earlier diagnosis, resulting in a faster recovery for a patient.

1.5 Scope

Among the leading causes of death, diabetes is one of the most prevalent. The number of cases is projected to triple in India by 2025 from 75 million cases in 2020. The population growth of a nation is exponentially accelerating every year. This presents a severe health threat. The mortality rate from diabetes in Tamil Nadu is the highest in India. Frequently, diabetes leads to long-term health problems and disability[9]. Pregnant mothers with gestational diabetics have a higher incidence of birth malformations, cardiac attacks, kidney damage, blindness, and other health complications. Diabetes treatment will cost the nation around 1.95 lakh crores per year. 34% of the income of the urban poor in India goes toward treating diabetes. This trend suggests a rise in premature deaths, which poses a serious risk to the advancement of the world's economy[10]. Over the last few decades, technological advances have helped lower hyperglycemia. Even with all of these advances in technology, diabetes remains a significant risk to health. The proposed study compares several machine learning methods with DM risk factors to create an ML model to predict diabetes mellitus.

CHAPTER 2

2.1 Background

2.1.1 Diabetes Mellitus

A collection of metabolic conditions known as diabetes cause hyperglycemia, or high blood glucose, due to limitations in the synthesis or action of insulin. In addition to debilitating effects on the eyes, kidneys, nerves, blood vessels, and heart due to diabetes it can also cause malfunctions and failures of a number of other body organs. As stated by CDC 2022(Centers for Disease Control and Prevention), diabetes can take the following types: type 1, type 2, and gestational diabetes[11]. Pancreatic islets of Langerhans produce significant amounts of insulin, a hormone composed of peptides that regulate glucose metabolism. There is also a small amount of insulin produced by some central nervous system neurons[12]. Insulin production and secretion are regulated by blood glucose levels. Insulin is synthesised when blood glucose levels are between 2 mM and 4 mM and secreted when they are greater than 5 nM. Without insulin, blood glucose levels will stay high. When the body has an excessive amount of glucose, hyperglycemia develops[13]. In response to secretion, insulin travels throughout the body and is delivered to skeletal muscle, adipocytes, and hepatocytes, also known as liver cells, where glucose is absorbed, causing blood glucose levels to fall. Also, hyperglycemia can occur when insulin is released but target cells do not absorb the glucose. Diabetes mellitus develops when hyperglycemia lasts for a prolonged length of time[14]. There are serious health consequences to DM, including damage to the nervous system, eye impairment, and kidney impairment, The severity of diabetes may differ based on the variety and length of time the patient has had it; long-term consequences may even be life-threatening.

2.1.2 Classification of Diabetes

There are three distinct forms of diabetes mellitus: type 1 diabetes, type 2 diabetes, and gestational diabetes.

Type 1 Diabetes

Hyperglycemia and lack of insulin are the hallmarks of T1DM(Type 1 Diabetes Mellitus), In the past, adolescents and children with diabetes were called insulin dependent diabetics. A Innate defense of the body kills the beta-cells of the pancreas unintentionally. A number of autoantibodies cause damage, including insulin antibodies, GAD, GAD65(glutamic acid decarboxylase antibodies) IA2(protein tyrosine phosphatase antibodies), and ZnT8A(zinc transporter protein antibodies). About 10-15% of diabetic individuals worldwide suffer from T1DM. The prevalence of T1DM in children is very high, and it can affect anyone it does not matter how old you are; Nonetheless, children, teens, and young people are most frequently affected by it. Indications might not appear for months or even years after the beta cells are destroyed in the pancreas[15]. A number of T1DM symptoms are described by the International Diabetes Federation (IDF), including polydipsia, polyuria, and enuresis, fatigue, loss of energy, polyphagia, sudden weight loss, slow wound healing, persistent infections, as well as vision impairment caused by diabetes[16]. Patients with T1DM require lifelong insulin replacement medication.

Type 2 Diabetes Mellitus

T2DM (Type 2 Diabetes Mellitus), also referred to as diabetes mellitus with no dependence on insulin, is characterised by hyperglycemia, decreased insulin sensitivity, and relative insulin deficit. Although pancreatic β cells can create enough insulin in T2DM, glucose balance cannot be achieved as a result of the cells' inability to use the insulin effectively[17]. Consequently, pancreatic cell secrete increasing amounts of insulin in order to stimulate a normal response from the body. An increase in blood glucose content is thus responsible for type 2 diabetes and hyperglycemia. T2DM predominates among all types of diabetes. Diabetes type 2 (T2DM) accounts for 95% of all diabetes cases in Indians. Diabetes type 2 is commonly observed in adolescents, teenagers,and youth, but there is a higher prevalence among adults over 45. Hereditary and behavioral factors both contribute to T2DM risk[18]. An example of a lifestyle risk factor is an inactive lifestyle, smoking, and alcohol intake. About fifty five percent T2DM patients have overweight as a risk factor, according to the CDC(Centers for Disease Control and Prevention). Researchers recently found a strong link between urine bisphenol-A (BPA) concentration and type 2 diabetes, and environmental contaminants like bisphenol A may be responsible for the recent spike in cases. Studies indicate a weakly positive association between T2DM and urine bisphenol-A (BPA) content, which may indicate that environmental toxins such as bisphenol A may be contributing to the current rise in T2DM cases. Plastics and epoxy resins are made with BPA, and are used to line food cans and to produce polycarbonate for baby bottles. According to McCarthy , TCF7L2, PPARG, FTO, KCNJ11, NOTCH2, WFS1, CDKAL1, IGF2BP2, SLC30A8, JAZF1, and HHEX all appear to be linked to T2DM[19]. There are several medical diseases that are considered causes of diabetics Type 2, such as overweight, high blood pressure, high fat, Reaven's disease, gigantism, hypercotisolism thyrotoxicosis, pheochromocytes, severe pancreatitis, and cancer.. The ageing process, fatty diets, and inactivity are also risk factors for diabetes.

Delayed Type 2 Diabetes

One-third of individuals with high HbA1c levels remain undiagnosed for over a year following hyperglycemia onset, and most wait four to seven years to receive a diagnosis . Microvascular complications caused by diabetes affect one-quarter of individuals with T2DM when they are diagnosed[20]. A patient document from the Veterans Affairs Medical Centre from 2010 indicates that the average time between the initial evidence of hyperglycemia and a clinical assessment was 3.7 years. In a 2002 study of 1426 participants, only 79% were diagnosed with diabetes mellitus based on hyperglycemia indications in their electronic health records[21]. When the patient goes undiagnosed, there is no chance for early intervention, even if there are no symptoms [22]. In comparison with cases without initial glycemic management, the chances of developing microvascular problems and coronary artery disease were significantly lower when blood sugar levels were managed earlier, a long-lasting decrease in threat upon diagnosis, Insufficient screening and poor access to healthcare can delay diabetes[23]. By using electronic health records (EHRs), diabetes mellitus can be identified and treated early.

Gestational Diabetes Mellitus

During pregnancy, spontaneous hyperglycemia is a defining feature of GDM(gestational diabetes mellitus). If a woman never suffered from diabetes before, it is possible for her to acquire GDM during the course of pregnancy. After delivery, gestational diabetes often disappears from the mother, which is more likely to cause hypoglycemia in the baby[24]. Data from the IDF(International Diabetes Federation) indicate that fifty percent of women with GDM(gestational diabetes) go on to become type two diabetic after their pregnancy. On the basis of CDC(Centers for Disease Control and Prevention) , approximately 7 percent of pregnant women are complicated by gestational diabetes (GDM). During pregnancy, a woman's body generates greater quantities of hormones and encounters a number of modifications, including weight increase, but not enough insulin[25]. A number of physiological changes can contribute to responsiveness to insulin, this exhibits a less effective utilisation of insulin by the body. Pregnant women face some degree of impaired insulin sensitivity throughout their pregnancies,

However sometimes pregnant women might already possess it, placing them at a greater likelihood of developing pregnancy-related diabetes[26]. Some contributing factors for GDM include overweight, inadequate nutrition, nutritional deficiencies, advanced mother's age at delivery And ancestry of diabetes .

Despite the fact that gestational diabetes mellitus often resolves after delivery, it can negatively affect a child's health, increasing likelihood of obesity, type two diabetics, cardiovascular disease, and gestational diabetes mellitus as time goes on[27]. Diabetes can have long-term effects that can be life-threatening. It can lead to renal failure by destroying the kidney's filtration system, blindness by damaging the eyes' blood vessels, foot injury by damaging the nerves' blood supply, erectile dysfunction by damaging the nerve's blood flow, nausea and vomiting by damaging the gastrointestinal tract.

2.1.3 Factor responsible for Diabetes

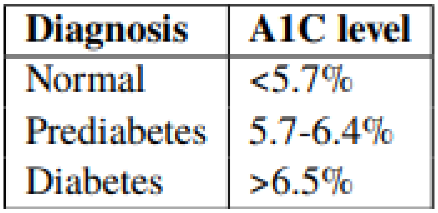
Numerous factors, including genetic predisposition and environmental triggers, might increase one's chance of getting diabetes. It is possible for overweight people to develop diabetes in the future. There is a higher possibility of getting the condition if a parent or sibling has it as well. The risk of diabetes increases with age. There is a correlation between diabetes and blood pressures above 140/90 mm Hg. Furthermore, the risk is increased by low HDL (high-density lipoprotein) levels[28]. Diabetic complications develop gradually. Potential side effects include cardiovascular illness. Neuropathy and cardiovascular issues are significantly increased by diabetes. The term nephropathy refers to kidney damage. Retinopathy is an eye injury. A lack of blood supply to the feet increases the risk of foot injury. acute skin disease.

2.1.4 Diagnostic

3 distinct methods carried out to detect diabetes.

A1C

Detecting type 2 diabetics and prediabetic can be accomplished through A1C testing.. In this calculator, the typical amounts of blood glucose throughout the previous 3 months are calculated. Because the blood test don't need to be fasted before, any time is suitable for drawing blood. A positive test result does not necessarily establish diabetes as a diagnosis[29]. An additional test may be required several days afterward or performed utilising a different method in order to make an assessment. Tests such as the A1C are affected by blood glucose levels. As glucose levels increase, the A1C score increases. The table below shows the full diagnosis values.

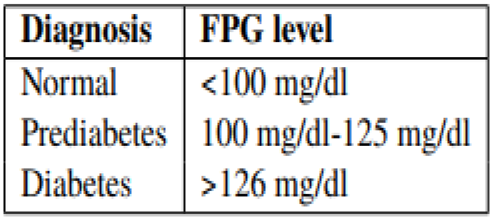


Diagnosis	A1C level
Normal	<5.7%
Prediabetes	5.7-6.4%
Diabetes	>6.5%

TABLE 1. Diagnosing diabetes with A1C readings

Fasting Plasma Glucose (FPG)

Using FPG, we can measure the level of glucose in the body. Fast for a minimum of 8 hours prior to performing the procedure for those who need to have their blood drawn[30]. Thus, the examination is often conducted in the morning. In table 2., values are provided for diagnostic purposes.

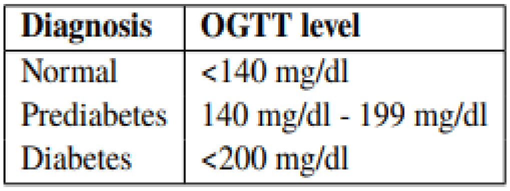


Diagnosis	FPG level
Normal	<100 mg/dl
Prediabetes	100 mg/dl-125 mg/dl
Diabetes	>126 mg/dl

TABLE 2. Diabetic diagnoses made on FPG readings

Oral Glucose Tolerance Test (OGTT)

How the body digests sugar is determined by an OGTT testing. Blood glucose levels are measured prior to and 2 hour after consuming sugary drinks as part of this examination[31]. A diagnosis can be made based on the values in table 3



Diagnosis	OGTT level
Normal	<140 mg/dl
Prediabetes	140 mg/dl - 199 mg/dl
Diabetes	<200 mg/dl

TABLE 3. Diagnosis of diabetes using OGTT levels

2.1.5 Treatment

Insulin shots are important for type 1 diabetic individuals to survive; They require prompt medical intervention after 80% or more of the cells that produce the hormone insulin have been damaged.. Diabetic type 1 is either inherited either inherited or occurs in the early stages of growth. Diabetic type 1 is typically diagnosed within twelve days because of the severity of the condition. However, the individual will be dependent upon drugs throughout their entire lives[32]. In type two diabetes, medication may or may not be needed for all patients. The same goes for pregnancy-related diabetes, which will likely disappear after delivery. Many types of diabetes take the lifestyle of the individual into account. As obesity interferes with insulin absorption, reducing fat in the blood may benefit symptoms. Thus, eating habits and regular physical activity routines are taken into account and altered in order to reduce problems, ensuring that a healthier way of life is sustained, medication no longer needs to be prescribed[33]. A combination of tablets and injections is appropriate in some circumstances to control insulin, while tablets and injections are recommended in other circumstances. A patient's needs may change as his or her level of hyperglycemia changes with time. The need for medication can arise for someone who didn't use to take it, and vice versa.

CHAPTER 3

3.1 Overview of Diabetes Management Challenges

3.1.1 Prevalence and Impact of Diabetes

Uncontrolled diabetes can be dangerous, resulting in a variety of of health problems. Hyperglycemia for a prolonged period of time can cause macrovascular disorders such as peripheral arterial inflammation, strokes and heart disease. A diabetic foot ulcer and lower-body amputation can be caused by diabetic nephropathy, neuropathic diabetics, and retinopathy diseases, which are all blood vessel consequences of diabetes. Furthermore, diabetes is a significant financial burden on society and the healthcare system. Expenditures related to diabetes healthcare of 2023 are expected to reach USD 780 billion by the IDF, or 15% of all medical spending[34]. Diabetes also entails monetary losses due to productivity reductions and a decline in quality of life, which exacerbate the economic burden.

3.1.2 Challenges in Diabetes Management:

1. Compliance with Treatment Plans

Individuals who have diabetes must follow complicated treatment programmes that involve drug schedules, dietary adjustments, physical exercise, and monitoring oneself of blood glucose level[35]. These recommendations are difficult to follow over time, resulting in substandard glycemic control and higher complications rates..

2. Glucose Control Variance

The intrinsic fluctuation of blood glucose levels makes it challenging to achieve and maintain optimum glycemic control. It might be difficult to stabilise and successfully manage glycemic control because of factors including food, exercise, stress, sickness, and adherence to medications.

3. Education and self-management

A person must have knowledge and abilities in order to effectively manage their diabetes. These abilities involve tracking blood glucose levels, giving medicines, knowing food options, and identifying hypo-

or hyperglycemia symptoms. Inadequate access to diabetes education programmes, a lack of knowledge, and problems with healthcare literacy can obstruct effective self-management.

4. Multidisciplinary Care and Comorbidities:

Comorbid conditions including hypertension, dyslipidemia, Weight gain, and heart problems are common ailments associated with diabetics[36]. The management of numerous chronic illnesses at once can be challenging and need for coordinated treatment from a variety of medical professionals with diverse expertise.

5. Health Inequalities and Care Access:

There are differences in the way different populations manage their diabetes, and some people have a harder time getting access to services for diabetes education, excellent treatment, and resources[37]. Uneven diabetes treatment and its consequences are a result of socioeconomic circumstances, cultural norms, geographical distribution, and health system inequities.

3.1.3 Limitations of Current Diagnostic Techniques:

- Fasting Plasma Glucose (FPG) Test:

  Diabetes can be diagnosed with the FPG test, which is often used. When it comes to sensitivity and specificity, it is nonetheless constrained. Transient hyperglycemia and aberrant postprandial glucose levels may go undetected, which might result in a delayed or incorrect diagnosis.

- Oral Glucose Tolerance Test (OGTT):

  Another diagnostic technique that assesses glucose tolerance over an extended period of time is the OGTT[38]. However, it takes a lot of time and calls for patients to consume a glucose load, which can be uncomfortable. Additionally, it is naturally variable and might provide false positives or false-negative readings

  .

- HbA1c Measurement:

  The test of HbA1c is frequently employed to evaluate long-term glycemic management. It does, however, have limits, especially in some groups, such as those with hemoglobinopathies or ailments that impact erythrocyte turnover[39]. Further, in certain clinical circumstances, such as pregnancy or in people with particular medical disorders, HbA1c values could not adequately reflect glycemic management.

- Failure to Detect Early

  When severe beta-cell malfunction or insulin resistance has already materialised, modern diagnostic tools frequently identify diabetes in its latter stages[40]. Early identification is essential for putting therapies in place on time and avoiding consequences, but existing approaches might not be able to accurately detect the disease's initial phases.

- Restricted evaluation for diabetic subtypes:

  Type1, type2, pregnancy-related, and monogenic diabetics are only a few types among the many subtypes. of the diverse disease known as diabetes[41]. The inadequacy of current diagnostic methods to distinguish between these subtypes might result in incorrect categorization and consequent use of subpar therapeutic approaches.

3.1.4 Limitations of Current Treatment Approaches:

- Adverse effects of medications

  There can be negative effects from several diabetes medicines, including insulin and oral antidiabetic therapies. Hypoglycemia, weight gain, gastrointestinal pain, and fluid retention are possible adverse effects. These negative effects may affect the compliance of patients and general therapeutic results.

- Lack of Personalization:

  Although many current treatment strategies adhere to standardised principles, individual reactions to therapies might differ. To improve diabetes management, more individualised strategies are required that take age, comorbidities, lifestyle, and treatment inclinations into account[42]. One-size-fits-all strategies might not be able to meet each patient's particular demands.

- Challenges of Treatment Intensification

  Treatment escalation is required when diabetes worsens in order to meet glycemic goals. However, because to issues such patient resistance, hypoglycemia anxiety, drug load, and expense, escalating therapy might be difficult[43]. These difficulties might cause treatment optimisation to be delayed and result in inadequate control of glucose levels.

- Insufficient glucose control

  Some diabetics may not respond to current treatment methods with the appropriate degree of glycemic control. Some individuals may develop hyperglycemia after therapy, which raises the risk of complications[44]. This emphasises the requirement for cutting-edge therapies and methods to enhance glucose management.

- Lack of Attention to Diabetes Prevention

  Despite the fact that the majority of current medications are geared towards controlling diabetes that has already developed, prevention is becoming increasingly important preventative treatments, such changes in routine, can potentially avoid or postpone the onset of diabetics of type two in patients with elevated risk. However, the application of preventative treatments in medical practise is still in its early stages.

CHAPTER 4

4.1 Machine Learning

Artificial intelligence's subfield called "machine learning" is concerned with computation models and techniques that help computers acquire and process data without being explicitly programmed. Through the analysis and interpretation of data patterns using statistical approaches, optimisation techniques, and computer algorithms, it enables machines to autonomously enhance their performance or behaviour based on past experience. Machine learning emerged as a result of researchers' efforts to understand whether computers might develop ways of imitating the mind of humans. Researchers created machine learning to see if computers might evolve to imitate brain function in humans[45]. ML made its first attempts to acquire the requisite talents to overcome a world championship when Arthur Samuel produced its first checkerboard video game play system in 1952. Later same year, in 1957, Frank Rosenblatt created an electrical device that can replicate the way the mind of humans functions to learn ways to deal with complex problems . The rise of machine learning has altered the use of computer devices in health sectror [46].

In order to make predictions or conclusions on brand-new, unforeseen data, machine learning algorithms utilise prior data, spot patterns, and derive insightful insights. A dataset is used to train a model, which then learns about the input data and related labelled outputs or targets[47]. The model then extrapolates from the training examples to create forecasts or judgements that incorporate brand-new, unobserved data.

The main applications of machine learning in medicine include illness detection and diagnosis, medication development, biomedical analysis of signals, and efficient electronic health records. The implementation of ML systems is viewed in the majority of situations of illness identification and diagnosis as an attempt to replicate the expertise possessed by medical specialists in condition identification[48]. Since ML allows software programs to gain knowledge from records, creating models to identify trends, and allowing it to generate judgments from gained knowledge, the method has no issues with the usage of incomplete medical databases.

4.2 ML in Healthcare

In the field of healthcare, ML has played a vital role in transforming medical practice, research, and decision-making. Machine learning algorithms may provide insightful knowledge, increase the precision of diagnoses, enable personalised therapies, and assist healthcare practitioners in making defensible judgements by utilising the large quantity of patient data. The following are some essential uses of ML in health maintenance. To help in illness assessment and prognosis, ML algorithms are able to analyze patient's information, for example, health histories, imaging scans, lab findings, and genomics details. By recognizing patterns in the data and finding subtle links, these algorithms can aid in identifying and diagnosing disease like cancer, coronary artery disease, diabetes type 2, and neurological disorders. Machine learning algorithms may forecast patient outcomes and determine the likelihood that a disease will emerge[49]. These models are able to detect individuals with a high probability of contracting particular diseases via examining patient data containing clinical traits, genetics, and lifestyle variables. This allows for proactive treatments and individualised preventative measures. Analysing patient features, genetic profiles, and treatment response information enables personalised therapy. These algorithms can aid in determining the best course of therapy, forecast medication reactions, and direct the choice of suitable medicines for certain patients[50]. Using wearable sensors, smartwatches, and heart rate monitors, machine learning algorithms can assess health conditions, spot anomalies, and provide real-time feedback. Using these algorithms, fitness objectives can be tracked, chronic diseases can be managed, and early warning signs of impending health issues can be identified. Radiology and imaging applications of machine learning techniques have showed substantial potential. To assist in the identification and characterisation of illnesses, these algorithms can evaluate and analyse medical imaging such as MRIs, CT scans, and X-rays[51]. They can help radiologists find anomalies, improve diagnostic precision, and speed up interpretation.

4.3 ML(Machine Learning)Algorithms Types

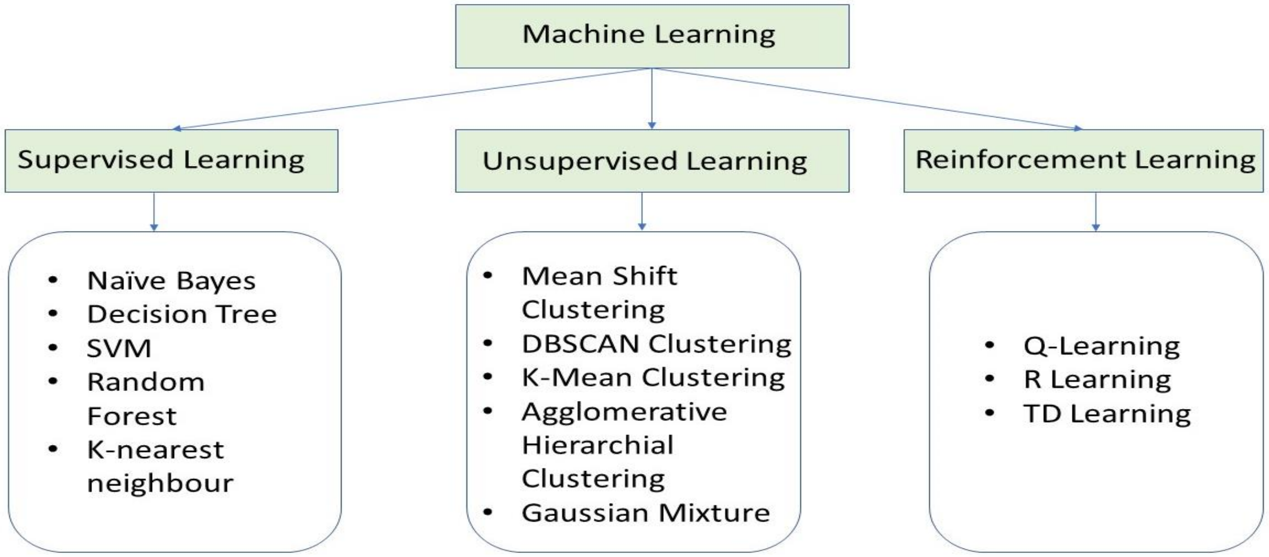


Fig.1. Flow chart showing classification of Machine Learning

4.3.1 Supervised Learning:

Supervisory learning algorithms obtain knowledge from labelled data, in which input data corresponds to well-known output labels or objectives. Several instances of supervised learning strategies involve SVM (Support Vector Machines), decision tree, linear regression and random forests.

The algorithm responds appropriately to every possible input using a training set of examples with appropriate objectives. Learning through examples is sometimes referred to as supervised learning. supervised learning includes regression and classification [52]. It provides a Yes/No prediction, such as " Is this product up to our high quality requirements?". Regression can be used to answer the questions "How much" and "How many".

4.3.2 Unsupervised Learning

Algorithms for unsupervised learning acquire knowledge based on unlabeled data and without precise output labels. In order to recognize patterns, structures, as well as correlations between records, they examine the data. Clustering algorithms (for example, k-means grouping and hierarchical clustering) and techniques for reducing dimensionality (like principal component analysis) fall under nonsupervised learning methods.

No objectives or appropriate answers are given. Unsupervised learning approaches attempt to identify patterns of similarity among the input data and then categorise the data according to these patterns. This also goes by the name of density estimate. Clustering is a component of unsupervised learning [53]. By using similarity, clustering creates groups.

4.3.3 Reinforcement Learning

Reinforcement learning systems pick up knowledge by interacting with their surroundings. The algorithms gradually develop the ability to maximise accumulated rewards after receiving input in the manner of rewards or penalties depending on their behaviour. Algorithms used in reinforcement learning includes Q-learning and deep reinforcement learning, for instance.

The behaviourist psychology supports this learning. Although the algorithm informs you when your response is incorrect, it does not tell you how to change it. Before it discovers the correct solution, it must investigate and test many options. It's also known as "gaining knowledge with a critique."" It does not suggest any upgrades. As opposed to SVM(Support Vector Machine) learning, As a result of reinforcement learning, there is no specific input or output set, nor are there clearly defined suboptimal behaviors. Additionally, it emphasises performance online [54].

4.4 Management of diabetes via Machine Learning

In recent years, machine learning technologies have demonstrated considerable potential in areas related to the management of diabetes, including as diagnosis, risk prediction, glucose monitoring, and therapy optimisation. Following are some ways ML is being used to treat diabetes:

- Early detection and risk assessment:

  Using ML algorithms, clinicians can identify people at high chance of acquiring diabetes or diabetes-related problems based on clinical traits, genetics, lifestyle factors, and biomarkers.. By using these algorithms, medical professionals can identify individuals more likely to acquire diabetes and initiate prevention measures at an early stage.

- Predictive modelling with glucose monitoring:

  Predictive models may be created using machine learning approaches that analyse continuous glucose monitoring (CGM) data, insulin dose records, physical activity data, and food data[55]. These models help with personalised insulin dose and enhance glycemic control by predicting future blood glucose levels and spotting trends.

- Personalised medicine and treatment optimisation.

  By taking into account the specifics of each patient, their reaction to therapy, and their lifestylechoices, machine learning algorithms can help to optimise diabetes treatment strategies. To offer individualised therapy suggestions, these algorithms can analyse information from electronic health records, feedback from patients, and clinical standards.

- Diabetes Complications and Risk Evaluation:

  Machine learning techniques can predict the likelihood of acquiring diabetes-related problems such retinopathy, nephropathy, and cardiovascular illnesses by analyzing a wide range of patient data, comprising medical records, imaging data, and laboratory findings[56]. Early identification, risk classification, and customised therapies can all be facilitated by these algorithms.

- Decision Support System

  Systems that aid in decision-making for healthcare workers can include machine learning models[57]. When making clinical decisions for the treatment of diabetes, such as adjusting insulin dosage, choosing medications, and making lifestyle changes, these systems may analyse patient data, offer real-time recommendations, and support the decision-making process.

Chapter 5

5.1 Proposed Methodology

The report's methodology is described in the following section. The dataset utilised and its constraints are described in Section 5.1.1. The mathematical programmes of the ML techniques are presented in Chapter 6- Random forest Artificial neural network(ANN), Logistic regression(LR), Naïve bayes and XgBoost. The experimental strategies utilised to accomplish the thesis's objectives and respond to the thesis' research questions are described in this chapter. The study compares five machine learning techniques that were developed and examined using the data set. Five ML techniques' F1 scores, recall, accuracy, and precision are contrasted. A description of the experimenting process is provided below.

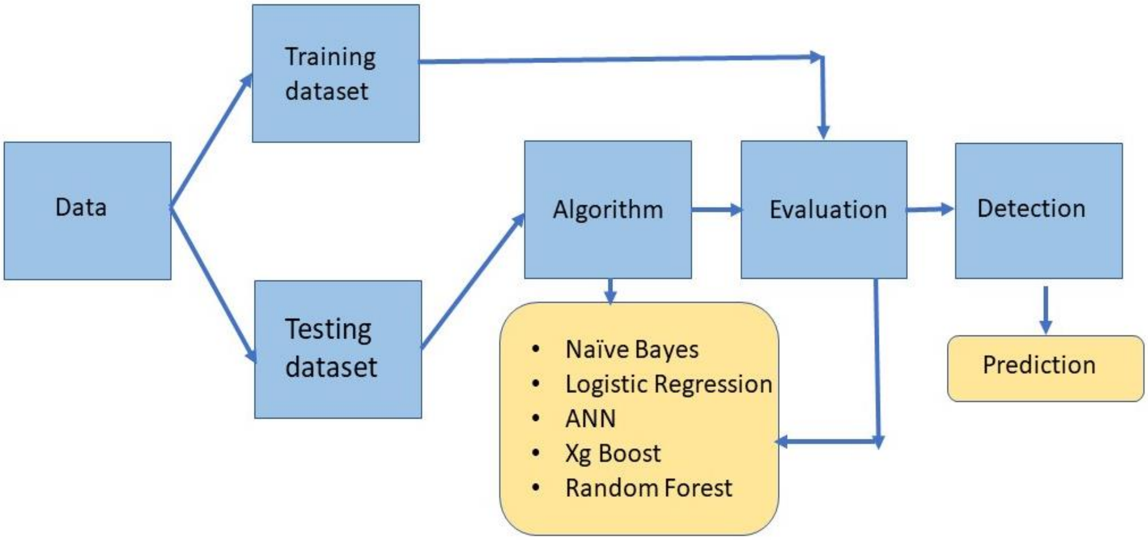


Fig.2. Flow chart of the procedure of the Experiment

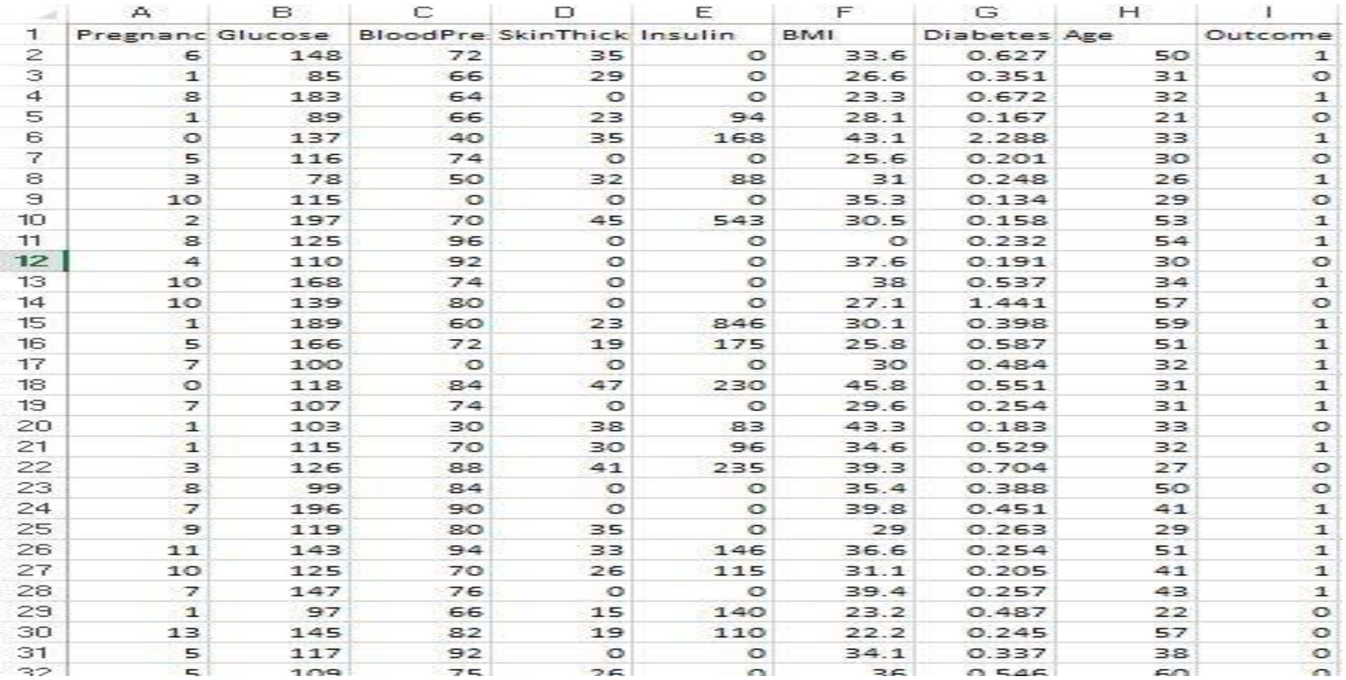
- A dataset is processed for missing or null values, while a label encoding system allows machines to read numerical data.
- A training dataset and a testing dataset are created using the train-test-split method, which divides the raw data into training and testing datasets.
- We trained, tested, and compared the machine learning models using a data set.

5.1.1 Data collection

The process of creating a machine learning recognition system begins with data collecting. For instance, you could want a medical dataset. The information was acquired from Kaggle, the biggest data science network in the world featuring a vast array of tools and services. A larger database of cases was selected under a number of restrictions. All of the patients in this group are Pima Indian women over the age of 21. A total of 8 medical predictive variables and an outcome are included in the dataset. This data set is a part of the UCI ML Repository and was initially obtained from the National Institutes of Diabetes, Digestive, and Kidney Diseases. The goal is to determine if a person has diabetes based on diagnostics parameters. In the case of a patient without diabetes, the result is 0, and in the case of a patient with diabetes, it is 1. A number of factors are used as medical predictors such as; Pregnancy, Glucose, BP, Skin thickness, Insulin, BMI, Diabetes pedigree function and Age, total 8 features of 768 samples are described in the table below

Features	Elucidation
Pregnancy	Count of pregnancies
Glucose:	The plasma glucose levels at two hours after an oral glucose tolerance test (GTIT) are measured
BP	Blood Pressure Diastolic (mm Hg)
Skin Thickness	Skin thickness of the triceps (millimeters)
Insulin	After two hours, serum insulin concentration (μh/ml)
Body mass index	BMI(Kg/m^2)
Diabetes Pedigree Function	Family history of diabetes
Age	Years old
Outcome	Whether a patient has diabetes (1) or not (0) is indicated by a binary number.

TABLE 4. Description of Pima Indian Diabetes Dataset



	A	B	C	D	E	F	G	H	I
1	Pregnanc	Glucose	BloodPre	SkinThick	Insulin	BMI	Diabetes	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1
27	10	125	70	26	115	31.1	0.205	41	1
28	7	147	76	0	0	39.4	0.257	43	1
29	1	97	66	15	140	23.2	0.487	22	0
30	13	145	82	19	110	22.2	0.245	57	0
31	5	117	92	0	0	34.1	0.337	38	0
32	5	109	75	26	0	36	0.546	60	0

Fig.3. Overview of Dataset

5.1.2 Data pre processing

Pre-processing data is one of the crucial steps in machine learning. This stage is the most important phase in developing machine learning models that are more efficient. The management of absent numbers, the eradication of outliers, and any other issues are all a part of this procedure. Categories or ordinal information must be converted in some manner into numerical characteristics since machine learning (ML) methods work best with numerical features, as is well known. Instead of the missing numbers, median values were utilised.

5.1.3 Feature Selection

A model must be trained before choosing which relevant features or attributes will be used. Feature extraction, a method, is carried out by selecting qualities that are relevant to the health situation being diagnosed. At this level, activation mechanisms and neurons are utilised to identify and prioritise the data set's most important features.

5.1.4 Model Selection

Following the feature section and data planning, the machine machine learning platform is selected. Machine learning employs a wide range of methods, including random forests, decision trees, naive bayes models, neural networks with artificial intelligence, and beyond. These models employ deep learning methodology and optimization techniques for improved performance for tasks including photo segmentation, categorization, and illness categorization. The model is executed using Python tools like Numpy, Matplot, Sk Learn, Pandas, etc.

5.1.5 Training and testing the model

Processing datasets using a machine learning algorithm involves two steps. Usually, we split the data by 20% to 80% for the testing and training stages.

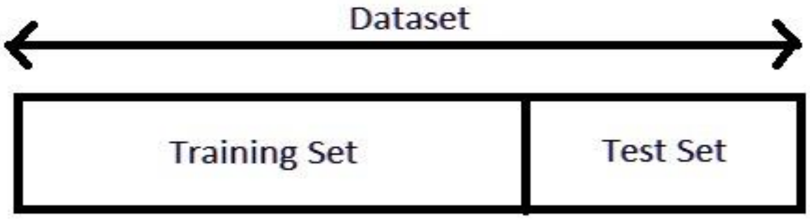
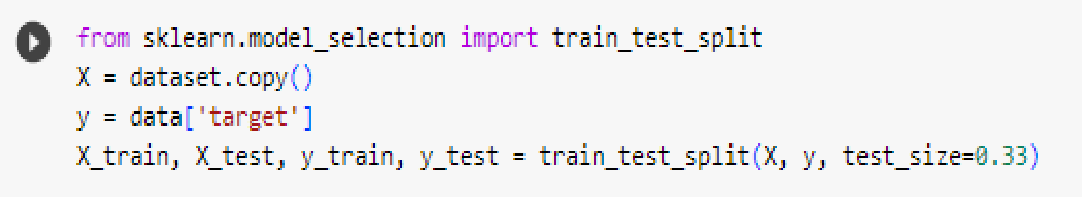


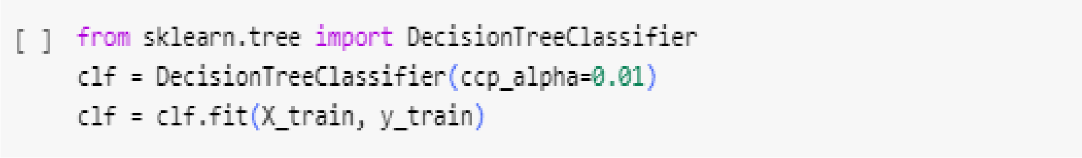
Fig.4. Raw data split into Training and Testing Dataset

To evaluate the effectiveness of the algorithms, we provide the model with fresh data that we've got labels for. The most common method for doing this is to use the Technique train_test_split to split the labelled data that we have collected into two portions. Our machine learning model was built using 75% of the training data, commonly referred to as the training set. The 25% of the information that was gathered that will be used to gauge how well the model works is known as test data or test set. Following the evaluation of the algorithms, we examine the outcomes to select the algorithm that delivers the best accuracy and identify the most reliable model for the detection and management of diabetes.

Code for train-test split



```
from sklearn.model_selection import train_test_split
X = dataset.copy()
y = data['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```



```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(ccp_alpha=0.01)
clf = clf.fit(X_train, y_train)
```

CHAPTER 6

6.1 Technical Approach

6.1.1 Random Forest

The Random Forests algorithm is one of the most powerful classification methods that can accurately categorize large volumes of data. The Random Forest is a kind of closest neighbor predictor, where several decision trees are built during training and an average of their class is derived from the results of each tree. They consist of a mixture of tree predictors, wherein each tree is reliant on the values of a random vector selected independently from the other trees in the forest using the same distribution[58]. By finding the natural equilibrium between the two extremes, Random Forests resolve the issue of high variance and high bias. They also have a method (called the "out of the bag error") for estimating mistake rates.

Each and every tree in a random forest thus relies on the parameters of a randomly chosen vector, whose distribution is the same across all trees. The generalisation error of a forest decreases as the number of trees increases. The robustness of each individual tree in the forest and the relationship between them affect the generalization error of a forest of tree classifiers. In comparison to other ML techniques, RF adheres to specific regulations for tree forming, tree combination, self-assess and the post-processing[59]. It is also robust to overfitting and is thought to be more stable in the presence of outliers and in extremely high-dimensional parameters gaps. A random subspace approach is used by RF to perform an implicit feature selection in order to evaluate variable significance..

The Gini index, which is not-parametric and does not depend on data that belongs to a certain type of distribution, is a measure of the predictive ability of covariates in regression or classification, depending on the impurity reduction concept . The Gini score of a node n for a binary division is determined as follows:

$$Gini\,(n) = 1 - \sum_{j=1}^{2} (p_j)^2$$

where pj is the node n's relative frequency of class j. The Gini index improvement should be maximised for optimal binary node splitting. To put it another way, a low Gini (i.e., a larger reduction in Gini) indicates that a certain predictor characteristic contributes more to the division of the data into the two groups. In order to rate the significance of attributes for a classification task, the Gini index could potentially be utilized .

The outliers in the training set have a significant influence on a number of machine learning models, including logistic and linear regression. In addition to being the result of human or instrument mistake, outliers represent modifications in the system's behaviour. Any given sample might possibly be corrupted. These excessive or outlier numbers have no bearing on the effectiveness or accuracy of the model. This issue is overcome and resolved using the RF Algorithm.

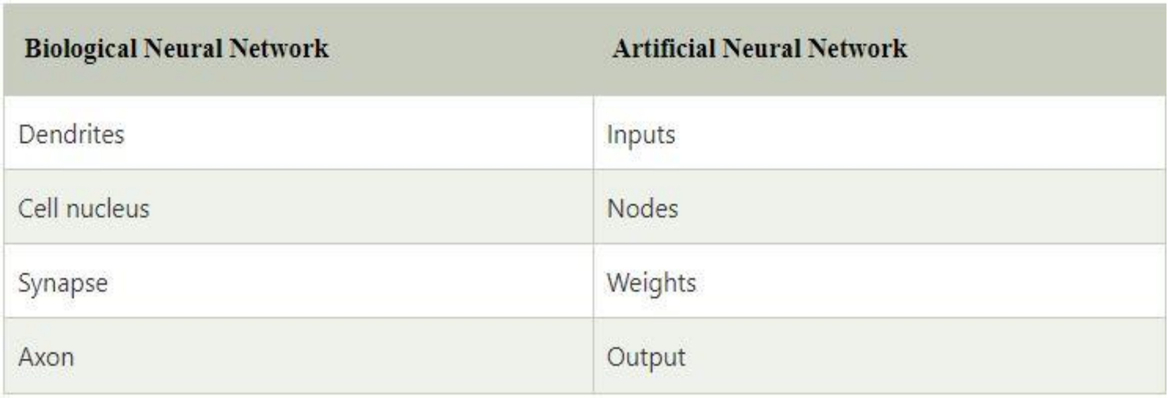
6.1.2 XG Boost

Gradient boosting is used by XGBoost to combine decision trees. Boosting approaches construct classifiers in a sequential manner, so that errors from one classifier are transmitted on to the next. XGBoost is capable of delivering prediction accuracies that equal several state-of-the-art supervised learning approaches, including neural networks, by developing decision trees on the gradient of the loss created by the preceding tree. Hyperparameters in machine learning configure many features of an algorithm; they must be specified prior to training and can have a significant impact on performance.

Cross-validation separates the training set at random to produce a tiny validation set that is used for assessing performance. The effect of altering a hyperparameter may therefore be assessed, and the best value can be chosen. It is typical to perform multiple rounds of cross-validation with distinct partitions.

6.1.3 Artificial neural network

Computer systems called Artificial Neural Networks, which are inspired by neural networks in organisms, are designed to mimic how people learn. It is a subfield of artificial intelligence that makes use of various optimisation methods to gather data, learn from prior experiences, and then apply that learning to categorise fresh data, find novel patterns, or forecast events.



Biological Neural Network	Artificial Neural Network
Dendrites	Inputs
Cell nucleus	Nodes
Synapse	Weights
Axon	Output

Fig.5. Biological vs. Artificial neural networks

The input and output layers, in addition to one or more hidden layers made up of nodes referred to as artificial neurons, are all components of neural networks. Fig showing the association between Biological and Artificial neural networks. Using a nonlinear function of inputs, artificial neurons calculate their outputs, which are then transformed by nodes into something that can be used by the output layer[60]. Using neural networks is limited by the fact that they are "black boxes," where data is input and results are generated. A black box refers to a situation in which solutions can be improved, but it is impossible to determine the exact process by which decisions are made. One of the major problems is the time required to train networks, which can consume a lot of computational power.

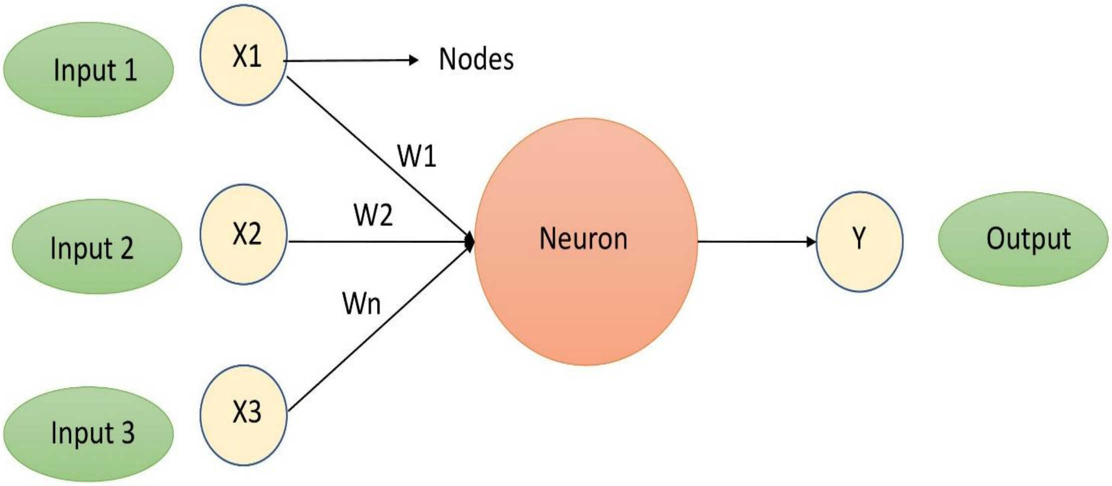


Fig.6. Workflow of Artificial Neural Network

In the pharmaceutical field, associating networks are frequently employed as a substitute to traditional response surface techniques, feature-extracting network as an alternate to principal component analysis, and nonadaptive networks for picture recognition. Based on these features, the ANN approach's potential application fields in the pharmaceutical fields are wide-ranging, encompassing everything from medical pharmacy to biopharmacy, drug and dose design, and data analysis.

Our dataset may have been somewhat tiny for a neural network, but it nevertheless performed well, leaving room for future growth with a larger dataset.

6.1.4 Logistic Regression

Among the most popular Machine Learning techniques, logistic regression is most commonly used in Supervised Learning. Forecasting a dependent variable with a categorical classification is done with a collection of uncorrelated variables. By using logistic regression, we can predict the output of a dependent category variable. Below fig showing the outcome of a dependent variable with a categorical attribute is predicted via logistic regression analysis[61]. Accordingly, the output needs to be a categorized or distinct value. Rather than defining 0 and 1, it provides probability values that fall somewhere between 0 and 1. It can be True or False, Yes or No, 0 or 1, etc.

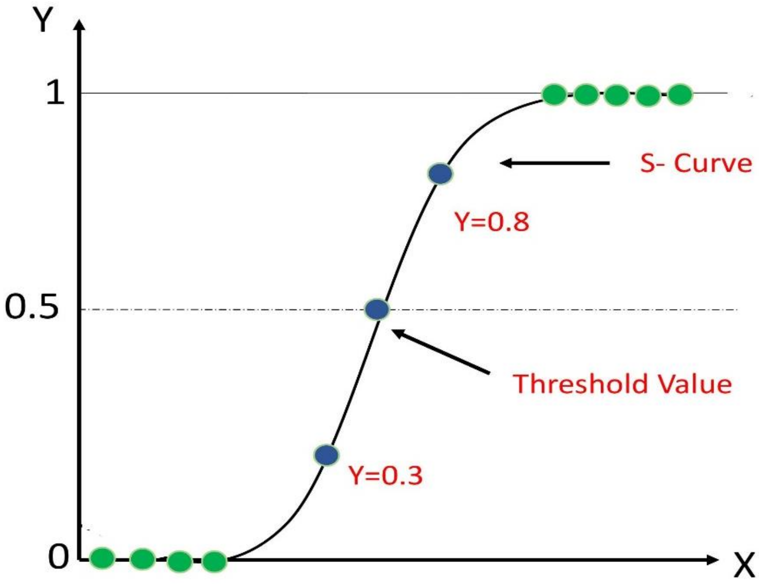


Fig.7. A categorized dependent variable's outcome is predicted via a logistic regression model in which As an alternative to a regression line, we construct a logistic function that forecasts the two highest numbers (0 or 1).

Logical regression models use the following sigmoid equation:

$$f(x) = \frac{1}{(1 + e^{-x})}$$

Where,

f(x): sigmoid function of x

e: epsilon (2.7182)

x: input value

The logistic regression technique is crucial for machine learning because it can categorize and compute probability using continuous and discrete datasets. With logistic regression, we can categorize observations using a variety of types of data and pinpoint which aspects are most advantageous for categorizing.

The basic formula for logistic regression is:

$$log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$

Here, X represents an independent variable, p(X) represents a dependent variable, 0 represents the intercept, and 1 represents the slope co-efficient. In order to produce an output (Y), input values (X) are linearly blended using coefficient values. A binary value is being modeled instead of a numeric number, which makes it distinct from linear regression. Despite having to fall within [0, 1], the score can have either a positive or negative value. Hence, logistic functions, also called sigmoid functions, are used to convert numbers. By using an S-shaped curve, any real number can be converted into a value between 0 and 1.

6.1.5 Naive Bayes

Statistical classification methods such as Nave Bayes have been successfully applied to bioinformatics. Using Bayes' theorem, one can make predictions regarding current events based on prior knowledge. A forecast evolves as new data becomes available. The investigators focus on the forecast, which is the probability that will come after. Prior probability, also referred to as prior knowledge, is a very reliable way

of predicting the outcome in the absence of further data[62]. Using current evidence, one can estimate the likelihood that a prediction will be correct given a certain result. The figure below demonstrates how the Naive Bayes method may be used to quickly solve the predictive classification issue

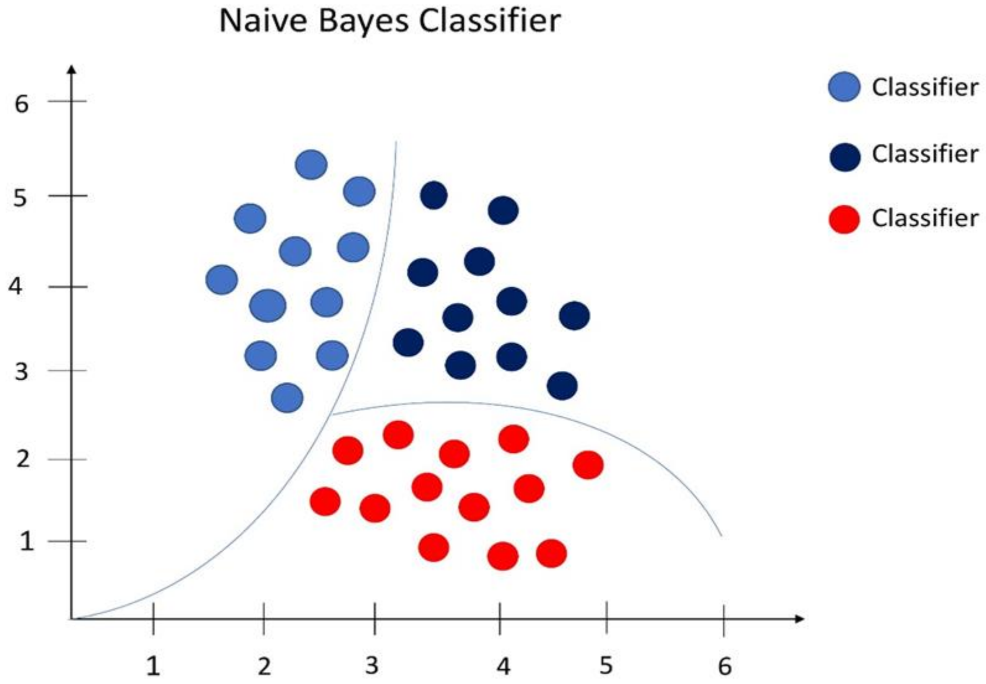


Fig.8. Figure depicting Naïve bayes classifier

The Bayesian network approach is crucial to the categorization or estimation of diabetes in machine learning. The Naive Bayesian network, which has the maximum accuracy of a score of up to 99.51%, is the kind of Bayesian network that is most frequently employed for classifications. The Bayesian Network employs the Naive Bayes theorem, which makes it much more favourable, effective, and independent since it firmly asserts that the existence of any specific characteristic in a class is not connected to the availability of any other attribute.

One of the significant popular methods for classifying diabetes is the the Bayes Network, which has an accuracy range of 71% to 99.51%. This Naive Bayesian method relies on conditional probability (the likelihood that a particular result would occur given a set of features):

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

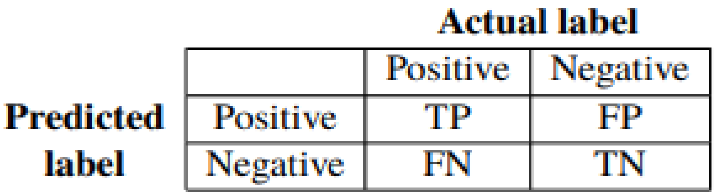
In the naive Bayes classification method, A stands for categorical result events, and B for a collection of predictors. If the result value is the same, the word "naive" suggests that the predictors are not related to

one another. As a consequence, P(b1, b2, b3|A) may be written as P(b1|A) P(b2|A) P(b3|A), greatly simplifying the computing process.

6.2 Comparison measurements

Confusion matrix

A confusion matrix is a performance measurement for machine learning classification problems. There are predicted values, which are the ones anticipated by the machine learning algorithm. The actual values are the true classification for that instance.



		Actual label	
		Positive	Negative
Predicted label	Positive	TP	FP
	Negative	FN	TN

Fig.9. Confusion matrix

When comparing predicted and actual values, four different columns are produced: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). For example, if an instance were predicted to have diabetes and did not have it, it is set as a false positive. Values from the confusion matrix is used to calculate precision, recall and accuracy.

Accuracy

The accuracy of a machine learning algorithm's classification of data is a measure of its effectiveness. It is calculated by dividing the whole number of guesses by the amount of predictions that were correct. The model is strong if the value is high.

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

Precision

Precision is the percentage of events that were accurately predicted to be positive. In this instance, the question is how many people have breast cancer despite receiving the designation.

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall counts determining which of the affirmative examples had accurate labels applied. How many people were given a diabetes diagnosis. A better value denotes a successful machine learning algorithm.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

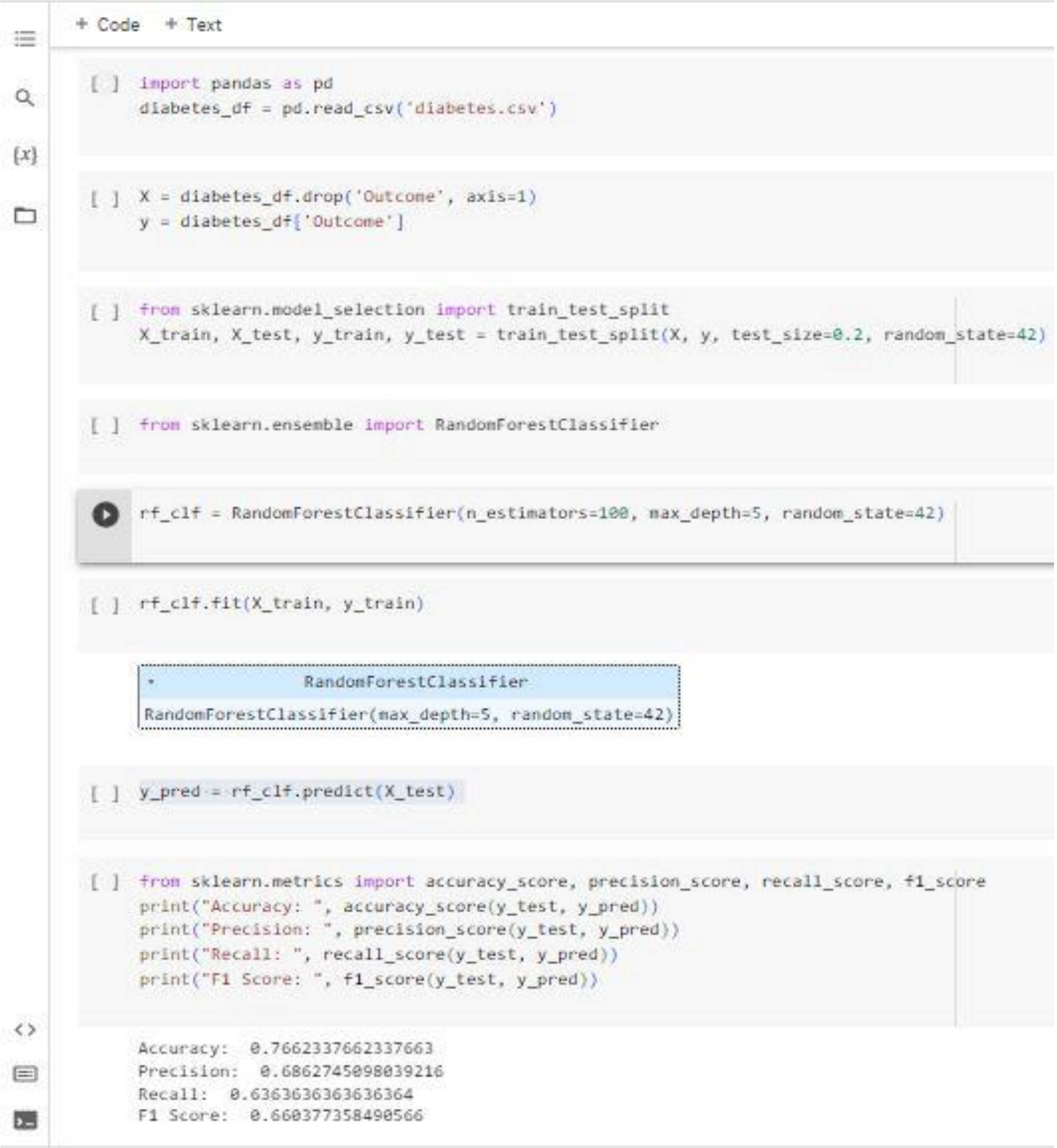
When evaluating a model's performance in binary classification tasks, the F1 score takes into account both precision and recall. It serves as a single parameter to assess the effectiveness of a classification system and is the harmonic average of accuracy and recall.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

CHAPTER 7

7.1 Algorithms and their Results

7.1.1 Random Forest Algorithm



```
import pandas as pd
diabetes_df = pd.read_csv('diabetes.csv')
```

```
X = diabetes_df.drop('Outcome', axis=1)
y = diabetes_df['Outcome']
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
rf_clf = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=42)
```

```
rf_clf.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(max_depth=5, random_state=42)
```

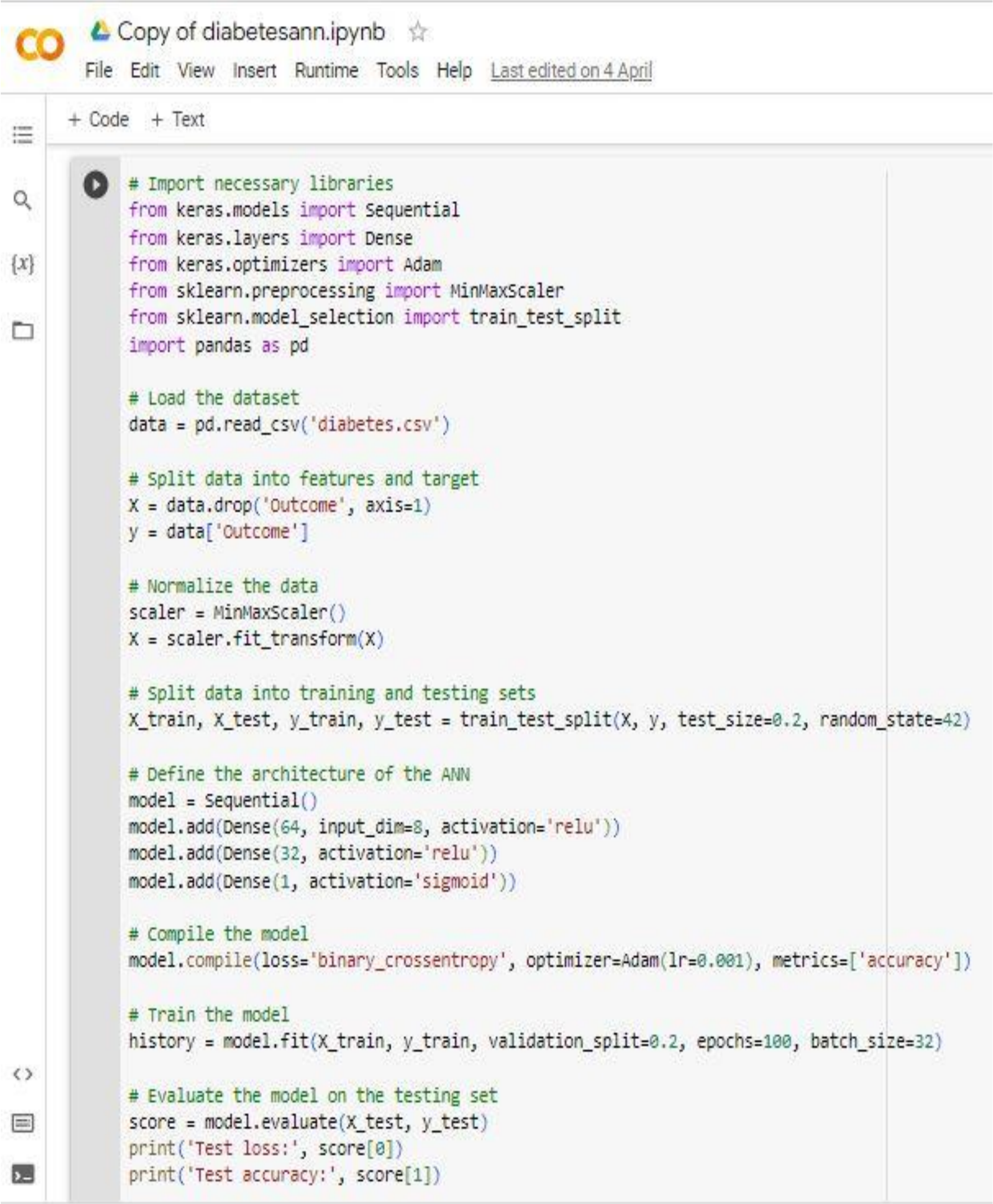
```
y_pred = rf_clf.predict(X_test)
```

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
print("Accuracy: ", accuracy_score(y_test, y_pred))
print("Precision: ", precision_score(y_test, y_pred))
print("Recall: ", recall_score(y_test, y_pred))
print("F1 Score: ", f1_score(y_test, y_pred))
```

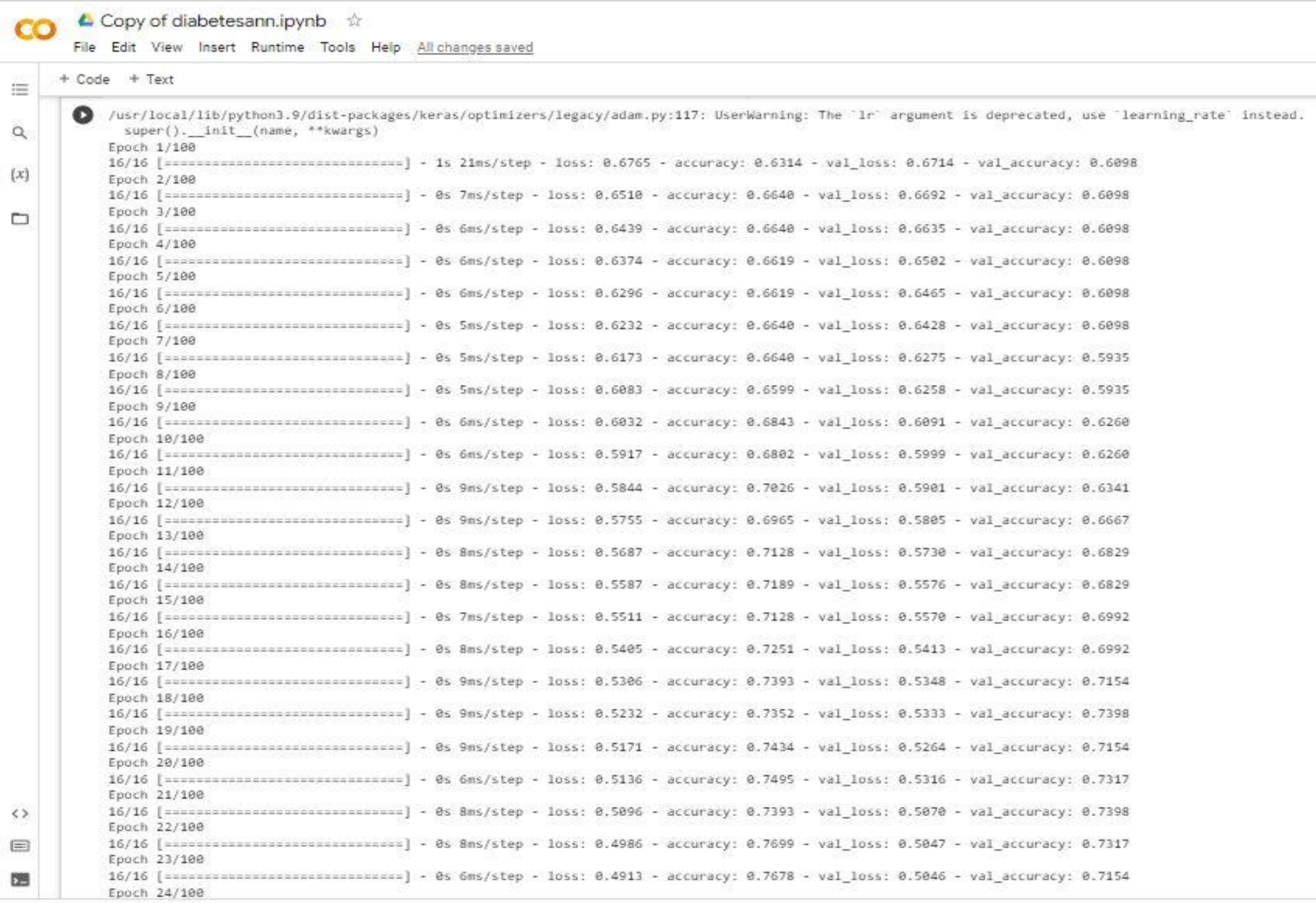
```
Accuracy:  0.7662337662337663
Precision:  0.6862745098039216
Recall:  0.6363636363636364
F1 Score:  0.660377358490566
```

Fig.10. Random Forest Result

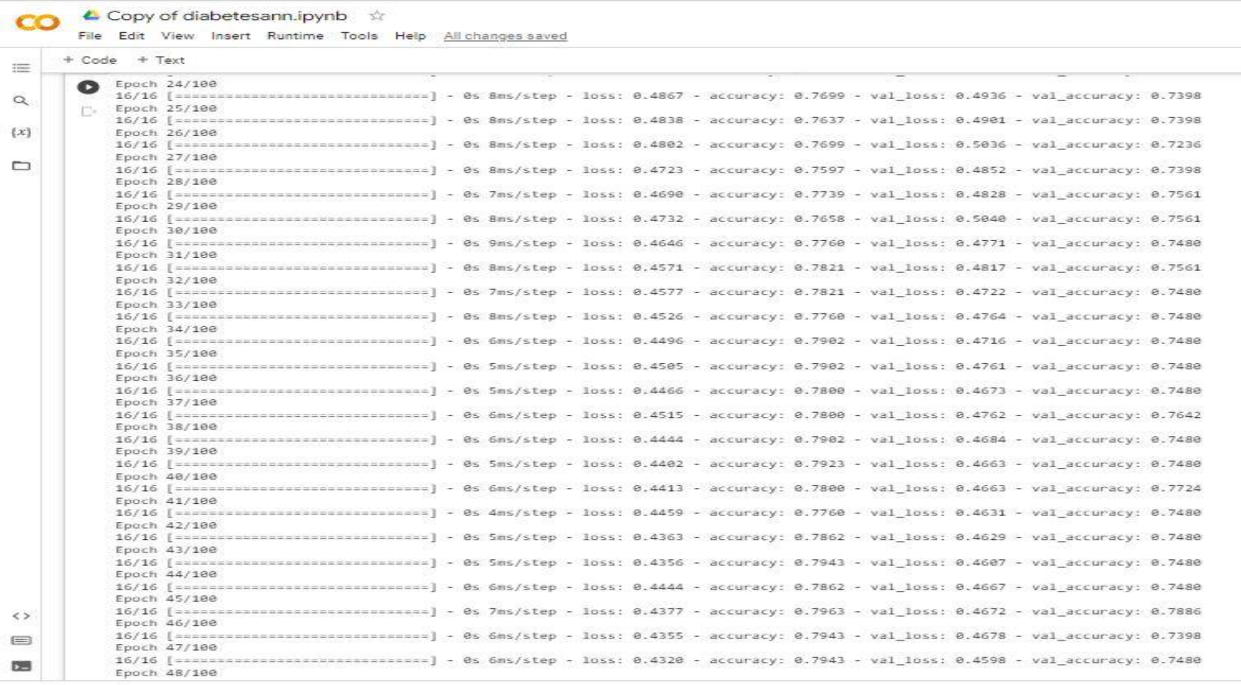
7.1.2 Artificial Neural Network Algorithm



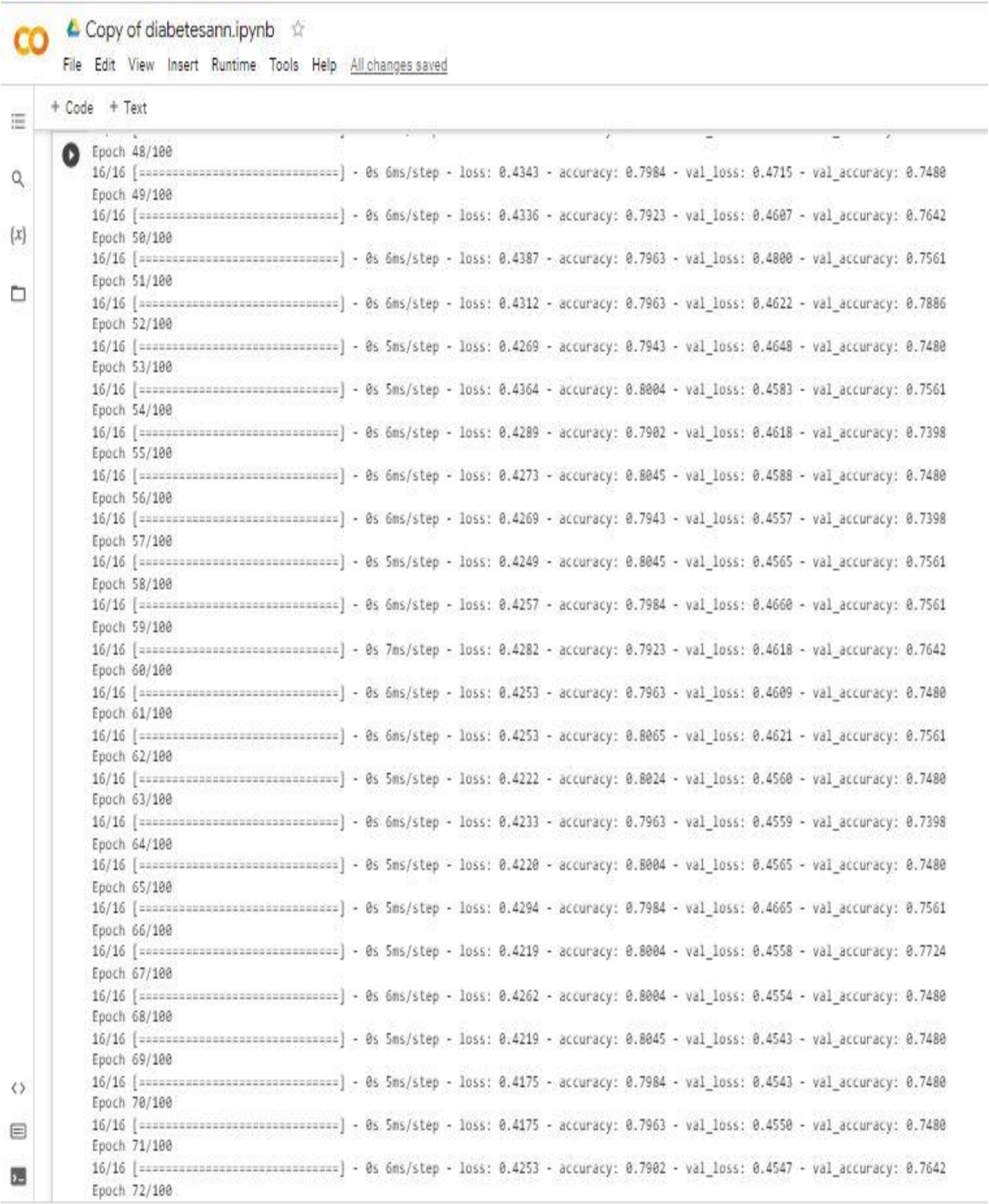
```
# Import necessary libraries
from keras.models import Sequential
from keras.layers import Dense
from keras.optimizers import Adam
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
import pandas as pd

# Load the dataset
data = pd.read_csv('diabetes.csv')

# Split data into features and target
X = data.drop('Outcome', axis=1)
y = data['Outcome']

# Normalize the data
scaler = MinMaxScaler()
X = scaler.fit_transform(X)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Define the architecture of the ANN
model = Sequential()
model.add(Dense(64, input_dim=8, activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

# Compile the model
model.compile(loss='binary_crossentropy', optimizer=Adam(lr=0.001), metrics=['accuracy'])

# Train the model
history = model.fit(X_train, y_train, validation_split=0.2, epochs=100, batch_size=32)

# Evaluate the model on the testing set
score = model.evaluate(X_test, y_test)
print('Test loss:', score[0])
print('Test accuracy:', score[1])
```


Copy of diabetesann.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

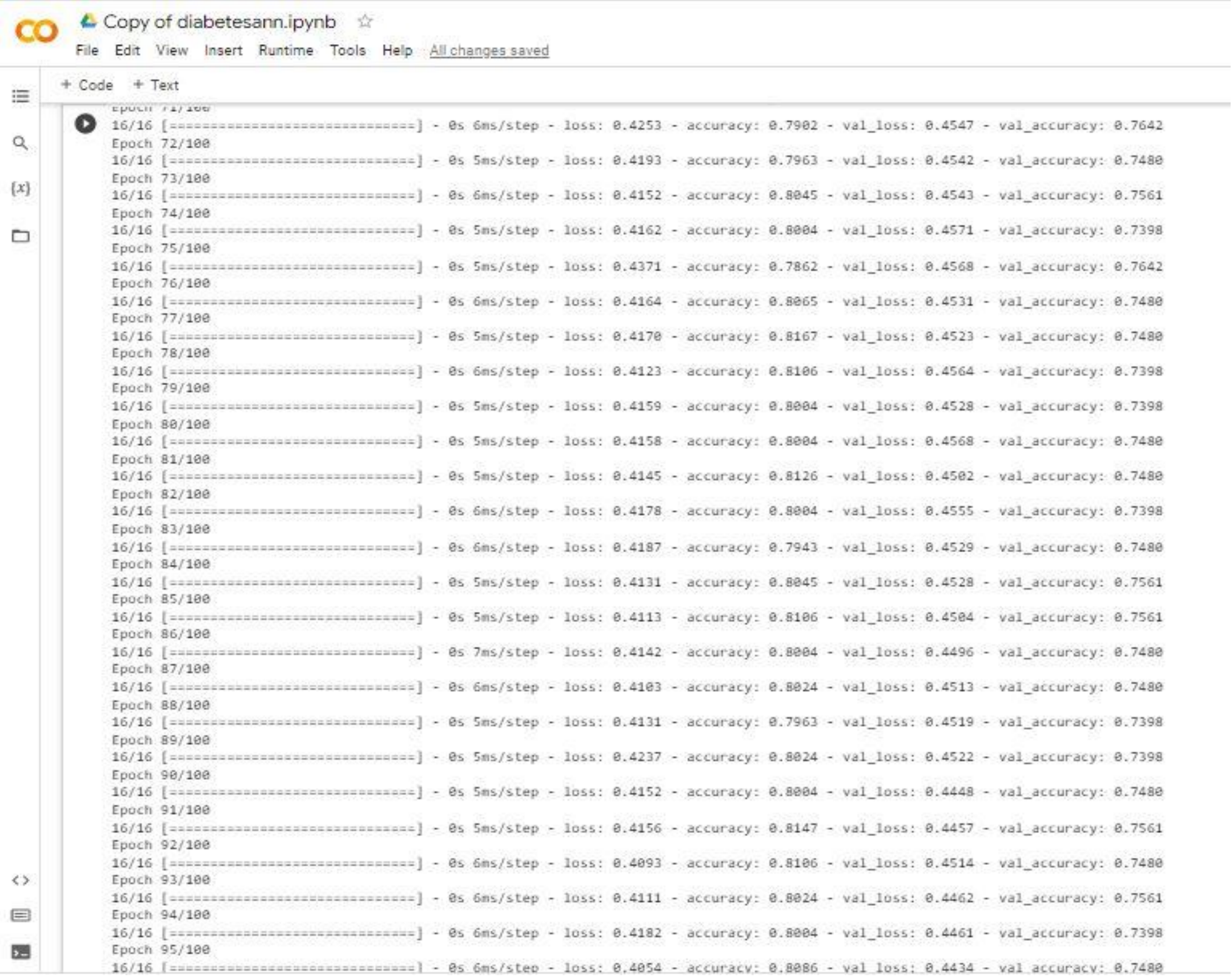
```
/usr/local/lib/python3.9/dist-packages/keras/optimizers/legacy/adam.py:117: UserWarning: The `lr` argument is deprecated, use `learning_rate` instead.
  super().__init__(name, **kwargs)
Epoch 1/100
16/16 [==============================] - 1s 21ms/step - loss: 0.6765 - accuracy: 0.6314 - val_loss: 0.6714 - val_accuracy: 0.6098
Epoch 2/100
16/16 [==============================] - 0s 7ms/step - loss: 0.6510 - accuracy: 0.6640 - val_loss: 0.6692 - val_accuracy: 0.6098
Epoch 3/100
16/16 [==============================] - 0s 6ms/step - loss: 0.6439 - accuracy: 0.6640 - val_loss: 0.6635 - val_accuracy: 0.6098
Epoch 4/100
16/16 [==============================] - 0s 6ms/step - loss: 0.6374 - accuracy: 0.6619 - val_loss: 0.6502 - val_accuracy: 0.6098
Epoch 5/100
16/16 [==============================] - 0s 6ms/step - loss: 0.6296 - accuracy: 0.6619 - val_loss: 0.6465 - val_accuracy: 0.6098
Epoch 6/100
16/16 [==============================] - 0s 5ms/step - loss: 0.6232 - accuracy: 0.6640 - val_loss: 0.6428 - val_accuracy: 0.6098
Epoch 7/100
16/16 [==============================] - 0s 5ms/step - loss: 0.6173 - accuracy: 0.6640 - val_loss: 0.6275 - val_accuracy: 0.5935
Epoch 8/100
16/16 [==============================] - 0s 5ms/step - loss: 0.6083 - accuracy: 0.6599 - val_loss: 0.6258 - val_accuracy: 0.5935
Epoch 9/100
16/16 [==============================] - 0s 6ms/step - loss: 0.6032 - accuracy: 0.6843 - val_loss: 0.6091 - val_accuracy: 0.6260
Epoch 10/100
16/16 [==============================] - 0s 6ms/step - loss: 0.5917 - accuracy: 0.6802 - val_loss: 0.5999 - val_accuracy: 0.6260
Epoch 11/100
16/16 [==============================] - 0s 9ms/step - loss: 0.5844 - accuracy: 0.7026 - val_loss: 0.5901 - val_accuracy: 0.6341
Epoch 12/100
16/16 [==============================] - 0s 9ms/step - loss: 0.5755 - accuracy: 0.6965 - val_loss: 0.5805 - val_accuracy: 0.6667
Epoch 13/100
16/16 [==============================] - 0s 8ms/step - loss: 0.5687 - accuracy: 0.7128 - val_loss: 0.5730 - val_accuracy: 0.6829
Epoch 14/100
16/16 [==============================] - 0s 8ms/step - loss: 0.5587 - accuracy: 0.7189 - val_loss: 0.5576 - val_accuracy: 0.6829
Epoch 15/100
16/16 [==============================] - 0s 7ms/step - loss: 0.5511 - accuracy: 0.7128 - val_loss: 0.5570 - val_accuracy: 0.6992
Epoch 16/100
16/16 [==============================] - 0s 8ms/step - loss: 0.5405 - accuracy: 0.7251 - val_loss: 0.5413 - val_accuracy: 0.6992
Epoch 17/100
16/16 [==============================] - 0s 9ms/step - loss: 0.5306 - accuracy: 0.7393 - val_loss: 0.5348 - val_accuracy: 0.7154
Epoch 18/100
16/16 [==============================] - 0s 9ms/step - loss: 0.5232 - accuracy: 0.7352 - val_loss: 0.5333 - val_accuracy: 0.7398
Epoch 19/100
16/16 [==============================] - 0s 9ms/step - loss: 0.5171 - accuracy: 0.7434 - val_loss: 0.5264 - val_accuracy: 0.7154
Epoch 20/100
16/16 [==============================] - 0s 6ms/step - loss: 0.5136 - accuracy: 0.7495 - val_loss: 0.5316 - val_accuracy: 0.7317
Epoch 21/100
16/16 [==============================] - 0s 8ms/step - loss: 0.5096 - accuracy: 0.7393 - val_loss: 0.5070 - val_accuracy: 0.7398
Epoch 22/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4986 - accuracy: 0.7699 - val_loss: 0.5047 - val_accuracy: 0.7317
Epoch 23/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4913 - accuracy: 0.7678 - val_loss: 0.5046 - val_accuracy: 0.7154
Epoch 24/100
```


Copy of diabetesann.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

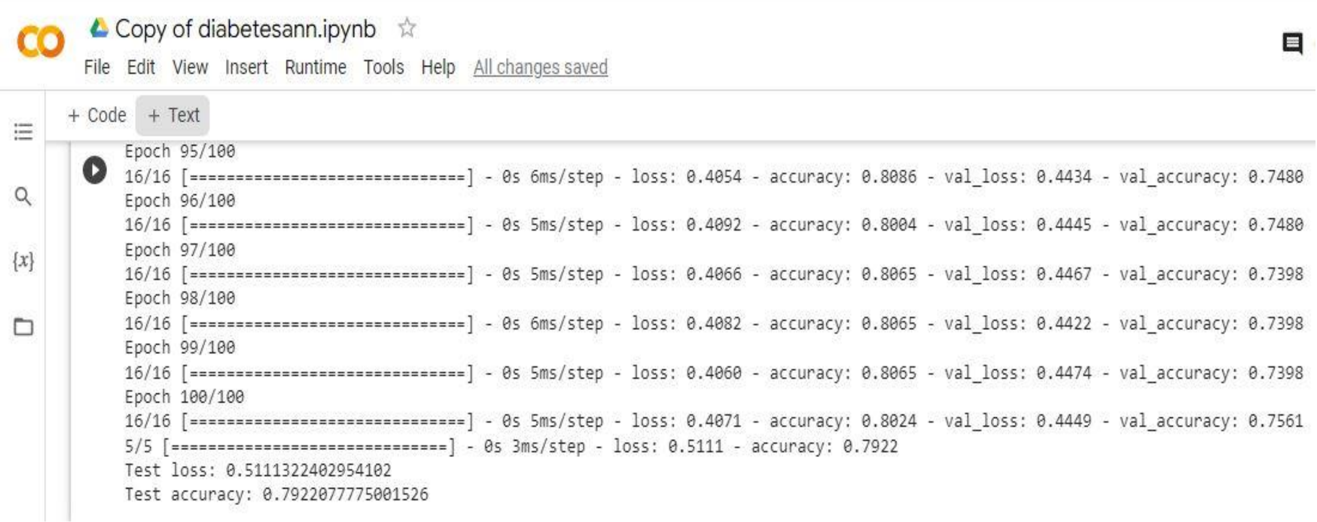
```
Epoch 24/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4867 - accuracy: 0.7699 - val_loss: 0.4936 - val_accuracy: 0.7398
Epoch 25/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4838 - accuracy: 0.7637 - val_loss: 0.4901 - val_accuracy: 0.7398
Epoch 26/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4802 - accuracy: 0.7699 - val_loss: 0.5036 - val_accuracy: 0.7236
Epoch 27/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4723 - accuracy: 0.7597 - val_loss: 0.4852 - val_accuracy: 0.7398
Epoch 28/100
16/16 [==============================] - 0s 7ms/step - loss: 0.4690 - accuracy: 0.7739 - val_loss: 0.4828 - val_accuracy: 0.7561
Epoch 29/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4732 - accuracy: 0.7658 - val_loss: 0.5040 - val_accuracy: 0.7561
Epoch 30/100
16/16 [==============================] - 0s 9ms/step - loss: 0.4646 - accuracy: 0.7760 - val_loss: 0.4771 - val_accuracy: 0.7480
Epoch 31/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4571 - accuracy: 0.7821 - val_loss: 0.4817 - val_accuracy: 0.7561
Epoch 32/100
16/16 [==============================] - 0s 7ms/step - loss: 0.4577 - accuracy: 0.7821 - val_loss: 0.4722 - val_accuracy: 0.7480
Epoch 33/100
16/16 [==============================] - 0s 8ms/step - loss: 0.4526 - accuracy: 0.7760 - val_loss: 0.4764 - val_accuracy: 0.7480
Epoch 34/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4496 - accuracy: 0.7902 - val_loss: 0.4716 - val_accuracy: 0.7480
Epoch 35/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4505 - accuracy: 0.7902 - val_loss: 0.4761 - val_accuracy: 0.7480
Epoch 36/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4466 - accuracy: 0.7800 - val_loss: 0.4673 - val_accuracy: 0.7480
Epoch 37/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4515 - accuracy: 0.7800 - val_loss: 0.4762 - val_accuracy: 0.7642
Epoch 38/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4444 - accuracy: 0.7902 - val_loss: 0.4684 - val_accuracy: 0.7480
Epoch 39/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4402 - accuracy: 0.7923 - val_loss: 0.4663 - val_accuracy: 0.7480
Epoch 40/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4413 - accuracy: 0.7800 - val_loss: 0.4663 - val_accuracy: 0.7724
Epoch 41/100
16/16 [==============================] - 0s 4ms/step - loss: 0.4459 - accuracy: 0.7760 - val_loss: 0.4631 - val_accuracy: 0.7480
Epoch 42/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4363 - accuracy: 0.7862 - val_loss: 0.4629 - val_accuracy: 0.7480
Epoch 43/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4356 - accuracy: 0.7943 - val_loss: 0.4607 - val_accuracy: 0.7480
Epoch 44/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4444 - accuracy: 0.7862 - val_loss: 0.4667 - val_accuracy: 0.7480
Epoch 45/100
16/16 [==============================] - 0s 7ms/step - loss: 0.4377 - accuracy: 0.7963 - val_loss: 0.4672 - val_accuracy: 0.7886
Epoch 46/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4355 - accuracy: 0.7943 - val_loss: 0.4678 - val_accuracy: 0.7398
Epoch 47/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4320 - accuracy: 0.7943 - val_loss: 0.4598 - val_accuracy: 0.7480
Epoch 48/100
```


Copy of diabetesann.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

```
Epoch 48/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4343 - accuracy: 0.7984 - val_loss: 0.4715 - val_accuracy: 0.7480
Epoch 49/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4336 - accuracy: 0.7923 - val_loss: 0.4607 - val_accuracy: 0.7642
Epoch 50/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4387 - accuracy: 0.7963 - val_loss: 0.4800 - val_accuracy: 0.7561
Epoch 51/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4312 - accuracy: 0.7963 - val_loss: 0.4622 - val_accuracy: 0.7886
Epoch 52/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4269 - accuracy: 0.7943 - val_loss: 0.4648 - val_accuracy: 0.7480
Epoch 53/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4364 - accuracy: 0.8004 - val_loss: 0.4583 - val_accuracy: 0.7561
Epoch 54/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4289 - accuracy: 0.7902 - val_loss: 0.4618 - val_accuracy: 0.7398
Epoch 55/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4273 - accuracy: 0.8045 - val_loss: 0.4588 - val_accuracy: 0.7480
Epoch 56/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4269 - accuracy: 0.7943 - val_loss: 0.4557 - val_accuracy: 0.7398
Epoch 57/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4249 - accuracy: 0.8045 - val_loss: 0.4565 - val_accuracy: 0.7561
Epoch 58/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4257 - accuracy: 0.7984 - val_loss: 0.4660 - val_accuracy: 0.7561
Epoch 59/100
16/16 [==============================] - 0s 7ms/step - loss: 0.4282 - accuracy: 0.7923 - val_loss: 0.4618 - val_accuracy: 0.7642
Epoch 60/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4253 - accuracy: 0.7963 - val_loss: 0.4609 - val_accuracy: 0.7480
Epoch 61/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4253 - accuracy: 0.8065 - val_loss: 0.4621 - val_accuracy: 0.7561
Epoch 62/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4222 - accuracy: 0.8024 - val_loss: 0.4560 - val_accuracy: 0.7480
Epoch 63/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4233 - accuracy: 0.7963 - val_loss: 0.4559 - val_accuracy: 0.7398
Epoch 64/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4220 - accuracy: 0.8004 - val_loss: 0.4565 - val_accuracy: 0.7480
Epoch 65/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4294 - accuracy: 0.7984 - val_loss: 0.4665 - val_accuracy: 0.7561
Epoch 66/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4219 - accuracy: 0.8004 - val_loss: 0.4558 - val_accuracy: 0.7724
Epoch 67/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4262 - accuracy: 0.8004 - val_loss: 0.4554 - val_accuracy: 0.7480
Epoch 68/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4219 - accuracy: 0.8045 - val_loss: 0.4543 - val_accuracy: 0.7480
Epoch 69/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4175 - accuracy: 0.7984 - val_loss: 0.4543 - val_accuracy: 0.7480
Epoch 70/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4175 - accuracy: 0.7963 - val_loss: 0.4550 - val_accuracy: 0.7480
Epoch 71/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4253 - accuracy: 0.7902 - val_loss: 0.4547 - val_accuracy: 0.7642
Epoch 72/100
```



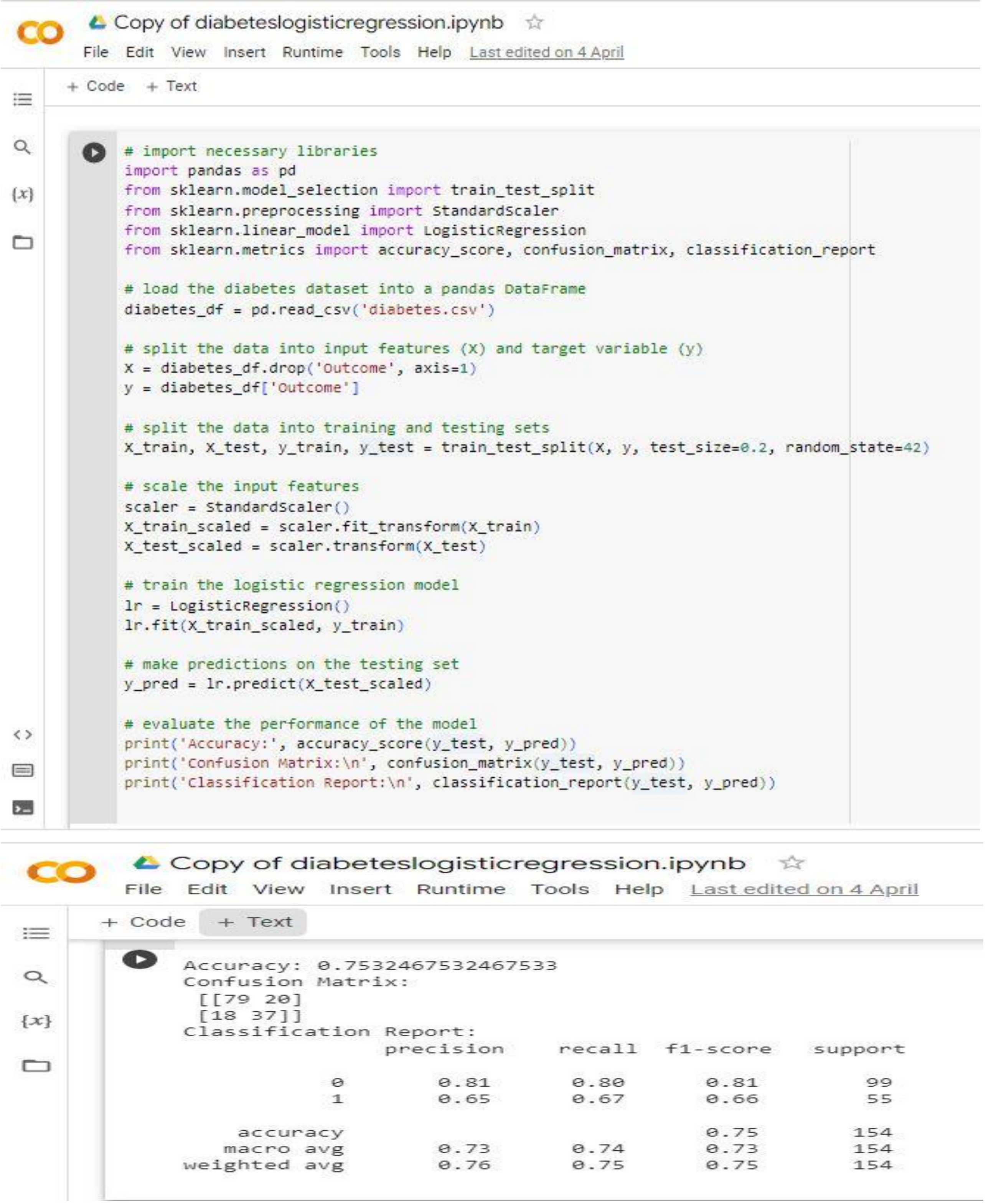
```
Epoch 71/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4253 - accuracy: 0.7902 - val_loss: 0.4547 - val_accuracy: 0.7642
Epoch 72/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4193 - accuracy: 0.7963 - val_loss: 0.4542 - val_accuracy: 0.7480
Epoch 73/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4152 - accuracy: 0.8045 - val_loss: 0.4543 - val_accuracy: 0.7561
Epoch 74/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4162 - accuracy: 0.8004 - val_loss: 0.4571 - val_accuracy: 0.7398
Epoch 75/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4371 - accuracy: 0.7862 - val_loss: 0.4568 - val_accuracy: 0.7642
Epoch 76/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4164 - accuracy: 0.8065 - val_loss: 0.4531 - val_accuracy: 0.7480
Epoch 77/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4170 - accuracy: 0.8167 - val_loss: 0.4523 - val_accuracy: 0.7480
Epoch 78/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4123 - accuracy: 0.8106 - val_loss: 0.4564 - val_accuracy: 0.7398
Epoch 79/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4159 - accuracy: 0.8004 - val_loss: 0.4528 - val_accuracy: 0.7398
Epoch 80/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4158 - accuracy: 0.8004 - val_loss: 0.4568 - val_accuracy: 0.7480
Epoch 81/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4145 - accuracy: 0.8126 - val_loss: 0.4502 - val_accuracy: 0.7480
Epoch 82/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4178 - accuracy: 0.8004 - val_loss: 0.4555 - val_accuracy: 0.7398
Epoch 83/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4187 - accuracy: 0.7943 - val_loss: 0.4529 - val_accuracy: 0.7480
Epoch 84/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4131 - accuracy: 0.8045 - val_loss: 0.4528 - val_accuracy: 0.7561
Epoch 85/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4113 - accuracy: 0.8106 - val_loss: 0.4504 - val_accuracy: 0.7561
Epoch 86/100
16/16 [==============================] - 0s 7ms/step - loss: 0.4142 - accuracy: 0.8004 - val_loss: 0.4496 - val_accuracy: 0.7480
Epoch 87/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4103 - accuracy: 0.8024 - val_loss: 0.4513 - val_accuracy: 0.7480
Epoch 88/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4131 - accuracy: 0.7963 - val_loss: 0.4519 - val_accuracy: 0.7398
Epoch 89/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4237 - accuracy: 0.8024 - val_loss: 0.4522 - val_accuracy: 0.7398
Epoch 90/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4152 - accuracy: 0.8004 - val_loss: 0.4448 - val_accuracy: 0.7480
Epoch 91/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4156 - accuracy: 0.8147 - val_loss: 0.4457 - val_accuracy: 0.7561
Epoch 92/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4093 - accuracy: 0.8106 - val_loss: 0.4514 - val_accuracy: 0.7480
Epoch 93/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4111 - accuracy: 0.8024 - val_loss: 0.4462 - val_accuracy: 0.7561
Epoch 94/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4182 - accuracy: 0.8004 - val_loss: 0.4461 - val_accuracy: 0.7398
Epoch 95/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4054 - accuracy: 0.8086 - val_loss: 0.4434 - val_accuracy: 0.7480
```



```
Epoch 95/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4054 - accuracy: 0.8086 - val_loss: 0.4434 - val_accuracy: 0.7480
Epoch 96/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4092 - accuracy: 0.8004 - val_loss: 0.4445 - val_accuracy: 0.7480
Epoch 97/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4066 - accuracy: 0.8065 - val_loss: 0.4467 - val_accuracy: 0.7398
Epoch 98/100
16/16 [==============================] - 0s 6ms/step - loss: 0.4082 - accuracy: 0.8065 - val_loss: 0.4422 - val_accuracy: 0.7398
Epoch 99/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4060 - accuracy: 0.8065 - val_loss: 0.4474 - val_accuracy: 0.7398
Epoch 100/100
16/16 [==============================] - 0s 5ms/step - loss: 0.4071 - accuracy: 0.8024 - val_loss: 0.4449 - val_accuracy: 0.7561
5/5 [==============================] - 0s 3ms/step - loss: 0.5111 - accuracy: 0.7922
Test loss: 0.5111322402954102
Test accuracy: 0.7922077775001526
```

Fig.11. Artificial Neural Network Result

7.1.3 Logistic Regression Algorithm

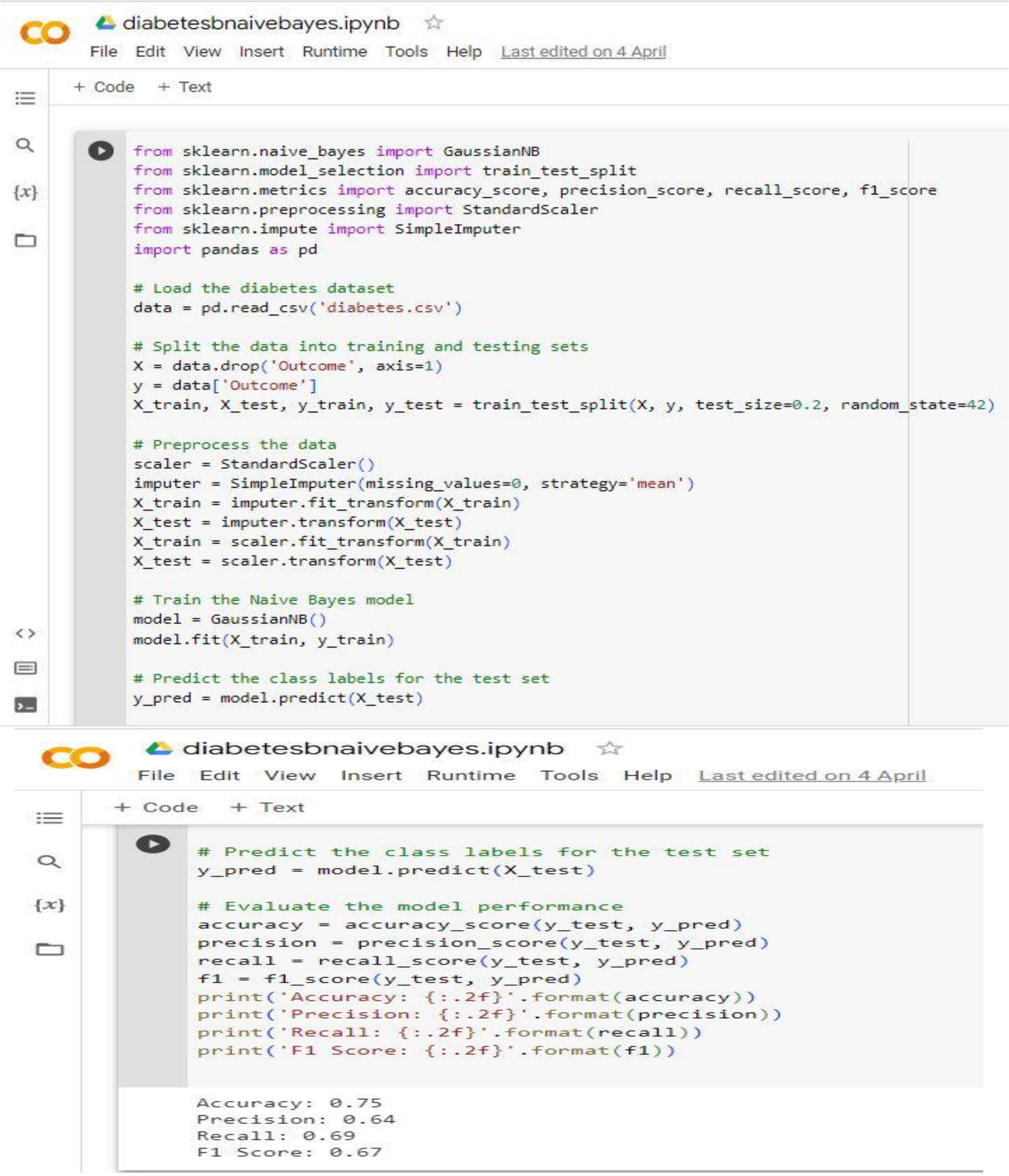


```
# import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# load the diabetes dataset into a pandas DataFrame
diabetes_df = pd.read_csv('diabetes.csv')

# split the data into input features (X) and target variable (y)
X = diabetes_df.drop('Outcome', axis=1)
y = diabetes_df['Outcome']

# split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# scale the input features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# train the logistic regression model
lr = LogisticRegression()
lr.fit(X_train_scaled, y_train)

# make predictions on the testing set
y_pred = lr.predict(X_test_scaled)

# evaluate the performance of the model
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Confusion Matrix:\n', confusion_matrix(y_test, y_pred))
print('Classification Report:\n', classification_report(y_test, y_pred))
```

```
Accuracy: 0.7532467532467533
Confusion Matrix:
 [[79 20]
 [18 37]]
Classification Report:
               precision    recall  f1-score   support

           0       0.81      0.80      0.81        99
           1       0.65      0.67      0.66        55

    accuracy                           0.75       154
   macro avg       0.73      0.74      0.73       154
weighted avg       0.76      0.75      0.75       154
```

Fig.12. Logistic Regression Result

7.1.4 Naive Bayes Algorithm


diabetesbnaivebayes.ipynb
File Edit View Insert Runtime Tools Help Last edited on 4 April
+ Code + Text

```
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
import pandas as pd

# Load the diabetes dataset
data = pd.read_csv('diabetes.csv')

# Split the data into training and testing sets
X = data.drop('Outcome', axis=1)
y = data['Outcome']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Preprocess the data
scaler = StandardScaler()
imputer = SimpleImputer(missing_values=0, strategy='mean')
X_train = imputer.fit_transform(X_train)
X_test = imputer.transform(X_test)
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Train the Naive Bayes model
model = GaussianNB()
model.fit(X_train, y_train)

# Predict the class labels for the test set
y_pred = model.predict(X_test)
```

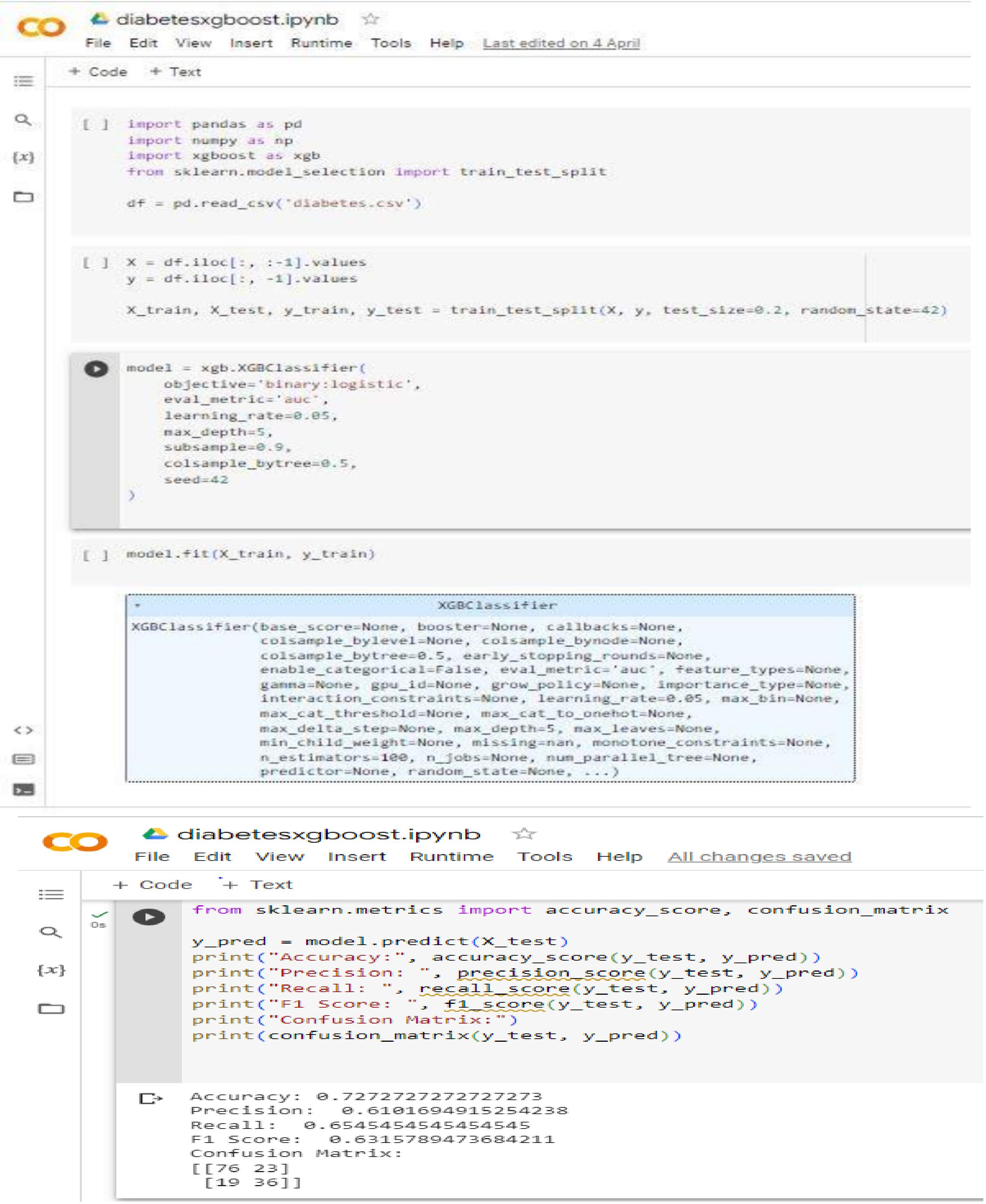
diabetesbnaivebayes.ipynb
File Edit View Insert Runtime Tools Help Last edited on 4 April
+ Code + Text

```
# Predict the class labels for the test set
y_pred = model.predict(X_test)

# Evaluate the model performance
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
print('Accuracy: {:.2f}'.format(accuracy))
print('Precision: {:.2f}'.format(precision))
print('Recall: {:.2f}'.format(recall))
print('F1 Score: {:.2f}'.format(f1))
```

```
Accuracy: 0.75
Precision: 0.64
Recall: 0.69
F1 Score: 0.67
```

Fig.13. Naive Bayes Result

7.1.5 XG Boost Algorithm


diabetesxgboost.ipynb
File Edit View Insert Runtime Tools Help Last edited on 4 April
+ Code + Text

```
import pandas as pd
import numpy as np
import xgboost as xgb
from sklearn.model_selection import train_test_split

df = pd.read_csv('diabetes.csv')
```

```
X = df.iloc[:, :-1].values
y = df.iloc[:, -1].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
model = xgb.XGBClassifier(
    objective='binary:logistic',
    eval_metric='auc',
    learning_rate=0.05,
    max_depth=5,
    subsample=0.9,
    colsample_bytree=0.5,
    seed=42
)
```

```
model.fit(X_train, y_train)
```

```
XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.5, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='auc', feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.05, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=5, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, ...)
```

```
from sklearn.metrics import accuracy_score, confusion_matrix

y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision: ", precision_score(y_test, y_pred))
print("Recall: ", recall_score(y_test, y_pred))
print("F1 Score: ", f1_score(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

```
Accuracy: 0.7272727272727273
Precision:  0.6101694915254238
Recall:  0.6545454545454545
F1 Score:  0.6315789473684211
Confusion Matrix:
[[76 23]
 [19 36]]
```

Fig.14. XG Boost Result

CHAPTER 8

8.1 Conclusion

On Pima Indian heritage dataset that is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, we applied five primary algorithms: Logistic Regression, Ann, Random Forest, XG Boost and Naïve Byes to calculate, compare, and evaluate various results according to confusion matrix, accuracy, sensitivity, precision, and AUC in order to determine the best machine learning algorithm that is accurate, dependable, and finds the higher accuracy. All algorithms were created using the Python sci-kit-learn library and the Anaconda environment. Following a precise comparison of our models, we discovered that ANN performed better than all other algorithms, achieving higher efficiency, of 79.22%. Finally ANN has proven to be effective in predicting and diagnosing diabetes and achieves the greatest efficiency in terms of precision and accuracy. The fact that all of the results are restricted to the PIMA database should be noted as a limitation of our work. As a result, it is important to consider applying the same algorithms and techniques to other databases in future work to validate the results obtained using this database. In addition, we intend to use our and additional machine learning algorithms with additional parameters on bigger sets of data with numerous disease classes to obtain.

Model	Dataset	Algorithms compared	Accuracy	Best Accuracy
Model proposed	Pima Indian Diabetes dataset	Random Forest	76.62%	**ANN- 79.22%**
		ANN	**79.22%**	
		Logistic Regression	75.32%	
		Naiive Bayes	75%	
		Xg Boost	72.27%	

TABLE 5. Comparison between various ML algorithm for Best Accuracy

In conclusion, by utilising vast and varied datasets to extract relevant patterns and characteristics, machine learning techniques have shown tremendous promise in identifying diabetes. This method has benefits including combining numerous data sources and offering a thorough evaluation of a person's risk. Machine learning models are capable of ongoing learning and adaptation, keeping up with new information and research. But there are drawbacks, such as the poor quality of training data as well as the difficulty of interpretation. Despite these drawbacks, with more study and improvement, machine learning algorithms might improve diabetes care in terms of early identification, individualized treatment programs, and general treatment results.

8.2 Limitation

- Limited data accessibility: Training and evaluation of machine learning models need vast volumes of labelled data. It might be difficult to locate thorough datasets that are expressly designed for diabetes diagnosis. The efficacy and generalisation of machine learning models for precise diagnosis might be hampered by a lack of accessibility to relevant and wide-ranging information.

- Disease that is multifaceted, complicated, and constantly changing: Diabetes is a disease that is multifaceted, complex, and ever-evolving. Given the disease's development and interactions with other medical diseases, machine learning algorithms may find it challenging to fully capture all of the complexities and dynamics of the illness. The state of a patient may change over time, and models that only take into account static variables may find it difficult to adjust.

- As was previously established, machine learning models like deep neural networks frequently lack interpretation and accountability. Patients and healthcare professionals must comprehend the thinking underlying the model's predictions in the instance of diabetes diagnosis in order to have confidence in the outcomes. It may be difficult to validate the model's judgements and defend them to beneficiaries if it is not interpretable.

- Considerations for ethics and privacy: Using machine learning to diagnose diabetes necessitates managing private patient information. Applications used in healthcare must maintain privacy, provide data security, and solve ethical issues. Appropriate implementation requires observing patient privacy rights, preventing data breaches, and creating strong data governance systems.

8.3 Future work

On a bigger and more varied dataset, further research may be done to predict diabetes utilising sophisticated machine learning methods like ensemble learning and deep learning. The prediabetic database can be used to predict diabetes mellitus earlier on. More models created in earlier studies can be used to compare the performance of machine learning models. Additionally, a comparison of the model's effectiveness with and without the data set's outliers may be performed. This will increase the predictability of diabetes mellitus, as well as determine whether outliers affecting performance are appreciably different when the model is trained with or without them.

REFERENCES

[1] Care, D. (2022). Addendum. 11. Chronic Kidney Disease and Risk Management: Standards of Medical Care in Diabetes—2022: Diabetes Care 2022; 45 (Suppl. 1): S175–S184. Diabetes care.

[2] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.

[3] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *Journal of medical systems*, *42*, 1-17.

[4] Kim, E., Caraballo, P. J., Castro, M. R., Pieczkiewicz, D. S., & Simon, G. J. (2019). Towards more accessible precision medicine: building a more transferable machine learning model to support prognostic decisions for micro-and macrovascular complications of type 2 diabetes mellitus. *Journal of medical systems*, *43*, 1-12.

[5] SK, S. (2017). A machine learning ensemble classifier for early prediction of diabetic retinopathy. *Journal of Medical Systems*, *41*, 1-12.

[6] Boutilier, J. J., Chan, T. C., Ranjan, M., & Deo, S. (2021). Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis. *Journal of medical Internet research*, *23*(1), e20123.

[7] Gopalan, A., Mishra, P., Alexeeff, S. E., Blatchins, M. A., Kim, E., Man, A. H., & Grant, R. W. (2018). Prevalence and predictors of delayed clinical diagnosis of type 2 diabetes: A longitudinal cohort study. *Diabetic Medicine*, *35*(12), 1655-1662.

[8] Goyal, R., & Jialal, I. (2022). Diabetes Mellitus Type 2.[Updated 2021 Sep 28]. *StatPearls [Internet]*. *Treasure Island (FL): StatPearls Publishing*.

[9] Indoria, P., & Rathore, Y. K. (2018). A survey: detection and prediction of diabetes using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, *7*(3), 287-291.

[10] Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ open*, *3*(5), e002457.

[11] Rahman, M. S., Hossain, K. S., Das, S., Kundu, S., Adegoke, E. O., Rahman, M. A., ... & Pang, M. G. (2021). Role of insulin in health and disease: an update. *International journal of molecular sciences*, *22*(12), 6403.

[12] Accili, D. (2018). Insulin action research and the future of diabetes treatment: the 2017 banting medal for scientific achievement lecture. *Diabetes*, *67*(9), 1701-1709.

[13] Okolo, C. T. (2022). *Diabetes Prediction Using Machine Learning Algorithm* (Doctoral dissertation, University of Louisiana at Lafayette).

[14] Misra, S., & Shukla, A. K. (2023). Teplizumab: type 1 diabetes mellitus preventable?. *European Journal of Clinical Pharmacology*, 1-8.

[15] Centers for Disease Control and Prevention (CDC. (2004). Prevalence of overweight and obesity among adults with diagnosed diabetes--United States, 1988-1994 and 1999-2002. *MMWR. Morbidity and mortality weekly report*, *53*(45), 1066-1068.

[16] Okolo, C. T. (2022). *Diabetes Prediction Using Machine Learning Algorithm* (Doctoral dissertation, University of Louisiana at Lafayette).

[17] Olokoba, A. B., Obateru, O. A., & Olokoba, L. B. (2012). Type 2 diabetes mellitus: a review of current trends, Oman Med. J. 27 (2012) 269–273.

[18] Lang, I. A., Galloway, T. S., Scarlett, A., Henley, W. E., Depledge, M., Wallace, R. B., & Melzer, D. (2008). Association of urinary bisphenol A concentration with medical disorders and laboratory abnormalities in adults. *Jama*, *300*(11), 1303-1310.

[19] Alberti, G., Zimmet, P., Shaw, J., & IDF Epidemiology task force consensus group. Metabolic syndrome: a world-wide definition. *Lancet*, *366*, 1059.

[20] Soriguer, F., Goday, A., Bosch-Comas, A., Bordiú, E., Calle-Pascual, A., Carmena, R., ... & Vendrell, J. (2012). Prevalence of diabetes mellitus and impaired glucose regulation in Spain: the Di@ bet. es Study. *Diabetologia*, *55*, 88-93.

[21] Porta, M., Curletto, G., Cipullo, D., Rigault de la Longrais, R., Trento, M., Passera, P., ... & Cavallo, F. (2014). Estimating the delay between onset and diagnosis of type 2 diabetes from the time course of retinopathy prevalence. *Diabetes care*, *37*(6), 1668-1674.

[22] Edelman, D. (2002). Outpatient diagnostic errors: unrecognized hyperglycemia. *Effective Clinical Practice*, *5*(1).

[23] Kiefer, M. M., Silverman, J. B., Young, B. A., & Nelson, K. M. (2015). National patterns in diabetes screening: data from the National Health and Nutrition Examination Survey (NHANES) 2005–2012. *Journal of general internal medicine*, *30*, 612-618.

[24] Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, 1-28.

[25] Goyal, R., & Jialal, I. (2022). Diabetes Mellitus Type 2.[Updated (2022) Jun 19]. *Stat Pearls [Internet]*. *Treasure Island (FL): Stat Pearls Publishing*.

[26] Plows, J. F., Stanley, J. L., Baker, P. N., Reynolds, C. M., & Vickers, M. H. (2018). The pathophysiology of gestational diabetes mellitus. *International journal of molecular sciences*, *19*(11), 3342.

[27] Tripathi, G., & Kumar, R. (2020, June). Early prediction of diabetes mellitus using machine learning. In *2020 8th international conference on reliability, Infocom technologies and optimization (trends and future directions)(ICRITO)* (pp. 1009-1014). IEEE.

[28] Holst, J. J., & Madsbad, S. (2016). Mechanisms of surgical control of type 2 diabetes: GLP-1 is key factor. *Surgery for Obesity and Related Diseases*, *12*(6), 1236-1242.

[29] Kramer, C. K., Araneta, M. R. G., & Barrett-Connor, E. (2010). A1C and diabetes diagnosis: the Rancho Bernardo Study. *Diabetes care*, *33*(1), 101-103.

[30] Sacks, D. A., Chen, W., Wolde-Tsadik, G., & Buchanan, T. A. (2003). Fasting plasma glucose test at the first prenatal visit as a screen for gestational diabetes. *Obstetrics & Gynecology*, *101*(6), 1197-1203.

[31] Bartoli, E., Fra, G. P., & Schianca, G. C. (2011). The oral glucose tolerance test (OGTT) revisited. *European journal of internal medicine*, *22*(1), 8-12.

[32] American Diabetes Association. (2009). Diagnosis and classification of diabetes mellitus. *Diabetes care*, *32*(Suppl 1), S62.

[33] Berne, C., & Agardh, D. (1997). Diabetes mellitus-nya svenska nationella riktlinjer. *Nordisk medicin*, *112*(5), 151-153.

[34] Tinajero, M. G., & Malik, V. S. (2021). An update on the epidemiology of type 2 diabetes: a global perspective. *Endocrinology and Metabolism Clinics*, *50*(3), 337-355.

[35] Peyrot, M., Rubin, R. R., & Khunti, K. (2010). Addressing barriers to initiation of insulin in patients with type 2 diabetes. *Primary Care Diabetes*, *4*, S11-S18.

[36] Schmitt, A., Reimer, A., Kulzer, B., Haak, T., Gahr, A., & Hermanns, N. (2014). Assessment of diabetes acceptance can help identify patients with ineffective diabetes self-care and poor diabetes control. *Diabetic medicine*, *31*(11), 1446-1451.

[37] Kirk, J. K., Passmore, L. V., Bell, R. A., Narayan, K. V., D'Agostino Jr, R. B., Arcury, T. A., & Quandt, S. A. (2008). Disparities in A1C levels between Hispanic and non-Hispanic white adults with diabetes: a meta-analysis. *Diabetes care*, *31*(2), 240-246.

[38] American Diabetes Association. (2010). Standards of medical care in diabetes—

2010. *Diabetes care*, *33*(Suppl 1), S11.

[39] Barr, R. G., Nathan, D. M., Meigs, J. B., & Singer, D. E. (2002). Tests of glycemia for the diagnosis of type 2 diabetes mellitus. *Annals of Internal Medicine*, *137*(4), 263-272.

[40] Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). Prediabetes: a high-risk state for diabetes development. *The Lancet*, *379*(9833), 2279-2290.

[41] Care, D. (2014). In the same tables, the sentence “The diagnosis of GDM is made when the plasma glucose level measured 3 h after the test is $140 mg/dL (7.8 mmol/L)” is incorrect. The corrected sentence is as follows:“The diagnosis of GDM is made when at least two of the following four plasma glucose levels (measured fasting, 1 h, 2 h, 3 h after the OGTT) are met or exceeded. *Diabetes Care*, *37*, 887.

[42] Inzucchi, S. E., Bergenstal, R. M., Buse, J. B., Diamant, M., Ferrannini, E., Nauck, M., ... & Matthews, D. R. (2015). Management of hyperglycemia in type 2 diabetes, 2015: a patient-centered approach: update to a position statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes care*, *38*(1), 140-149.

[43] Montvida, O., Shaw, J., Atherton, J. J., Stringer, F., & Paul, S. K. (2018). Long-term trends in antidiabetes drug usage in the US: real-world evidence in patients newly diagnosed with type 2 diabetes. *Diabetes care*, *41*(1), 69-78.

[44] Tuomilehto, J., Lindström, J., Eriksson, J. G., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., ... & Uusitupa, M. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *New england journal of medicine*, *344*(18), 1343-1350.

[45] Mitchell, T. M. (2007). *Machine learning* (Vol. 1). New York: McGraw-hill.

[46] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[47] Alić, B., Gurbeta, L., & Badnjević, A. (2017, June). Machine learning techniques for classification of diabetes and cardiovascular diseases. In *2017 6th mediterranean conference on embedded computing (MECO)* (pp. 1-4). IEEE.

[48] Magoulas, G. D., & Prentza, A. (2001). Machine learning in medical applications. *Machine Learning and Its Applications: advanced lectures*, 300-307.

[49] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542*(7639), 115-118.

[50] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... &

Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, *316*(22), 2402-2410.

[51] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.

[52] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.

[53] Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

[54] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[55] Charpentier, G., Benhamou, P. Y., Dardari, D., Clergeot, A., Franc, S., Schaepelynck-Belicar, P., ... & TeleDiab Study Group. (2011). The Diabeo software enabling individualized insulin dose adjustments combined with telemedicine support improves HbA1c in poorly controlled type 1 diabetic patients: a 6-month, randomized, open-label, parallel-group, multicenter trial (TeleDiab 1 Study). *Diabetes care*, *34*(3), 533-539.

[56] Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., ... & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174-311ra174.

[57] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395-405.

[58] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

[59] Vermeulen, I., Weets, I., Asanghanwa, M., Ruige, J., Van Gaal, L., Mathieu, C., ... & Belgian Diabetes Registry. (2011). Contribution of antibodies against IA-2β and zinc transporter 8 to classification of diabetes diagnosed under 40 years of age. *Diabetes care*, *34*(8), 1760-1765.

[60] van Gerven, M. A. J., & Bohte, S. M. (2017). Artificial neural networks as models of neural information processing.

[61] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, 100179.

[62] Priya, K. L., Kypa, M. S. C. R., Reddy, M. M. S., & Reddy, G. R. M. (2020, June). A novel approach to predict diabetes by using Naive Bayes classifier. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)* (pp. 603-607). IEEE.

Behaviour Analysis Using Machine Learning Algorithms In Health Care Sector

Anukriti Yadav

Department of Biotechnology Engineering,Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

anukritiyadav454@gmail.com

Deepak Kumar

Department of Biotechnology Engineering, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

deeepak545@gmail.com

Yasha Hasija*

Department of Biotechnology Engineering, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

yashahasija06@gmail.com

*Corresponding Author

***Abstract* - A behavioral analytics approach uses big data analytics in combination with machine learning (ML) to identify patterns, trends, aberrations, and other useful insights. The behavior of an individual can be analyzed by expressions, postures, and activity levels. Using ML algorithms could revolutionize the way clinicians make decisions in health care sector. Studies of human behavior have been conducted in a range of scientific disciplines (e.g sociology, psychology, computer science). ML algorithms have the potential to transform the way doctors and instructors make choices. This methodology has been slow to be adopted by behavior analysis experts to maximize its application to practical issues and to aid them in learning more about human behavior. ML algorithms are dominating the healthcare industry. Recent researches have indicated that these techniques can be used to anticipate disease based on health data. Our study examines several machine learning algorithms used in early disease detection and identifies key trends in their performance. The analysis suggests that human behavior may play a role in a variety of conditions, including diabetes, cancer, heart disease, autism, mental illness, Alzheimer's, and others. A number of daily habits are associated with this behavior, including food, respiration rate, blood pressure, voice output, social abnormalities, insomnia, and so on. A few examples of ML applications integrated into healthcare services are naive bayes (NB), support vector machines (SVM), random forest (RF), and convolutional neural networks (CNN). In a variety of cancer classification applications, these models are proved to be highly efficient in diagnosing various cancer types. This review includes a number of research investigations that employ ML to analyze behavioral data. As we gain further insights into the factors influencing organisms' behavior, we are able to create computational models which allow disease prediction and management to become more accurate.**

Keywords— Behavioral analytics, Machine learning`, Algorithm, accuracy, healthcare services

I. INTRODUCTION

Patients with neurological diseases, head traumas, and mental illnesses benefit greatly from behavioral analysis in the health sector It is helpful to determine the root cause of a disease by analyzing the patient's behavior. There are many challenges associated with patient behavioral analysis in traditional healthcare [1]. Through the development of smart healthcare, it is possible to analyze patient behavior more easily. Using ML for human behavior recognition has become a new topic of analysis due to the issues relating to potency and accuracy of conventional artificial feature-extraction behavior identification. **Figure 1** shows different type of machine learning algorithms that are used behavioral analysis

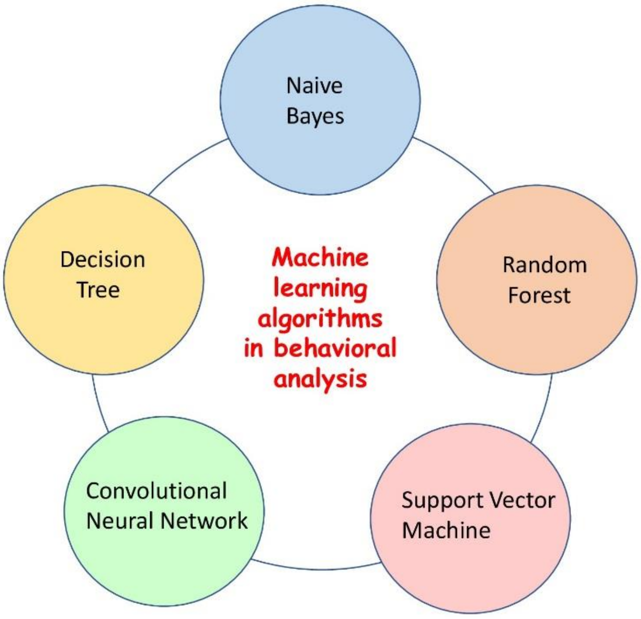

Naive Bayes
Decision Tree
Machine learning algorithms in behavioral analysis
Random Forest
Convolutional Neural Network
Support Vector Machine

Fig. 1. Different type of machine learning algorithms in behavioral analysis

Research investigators used ML tools to detect behavioral patterns for different patient groups based on the experimental case study they conducted such as pattern analysis for anxiety and depression level [2], assessing patients who have been diagnosed with autism and those who have not [3].

In spite of extensive instrumental as well as scoring noise, ML is capable of detecting interactions that are complex, high-dimensional and, non-linear that may notify prognosis [4]. **Figure 2** depicts how behavioral analysis is carried out using machine learning algorithms. Many biology and behavioral research laboratories, however, find it strenuous to implement these advanced analyses, which may explain why they have not yet been widely adopted. In this article, we have reviewed about various machine learning techniques (like SVM, CNN, Decision tree, Random forest, Naive bayes) to study behavior analysis in healthcare sector.

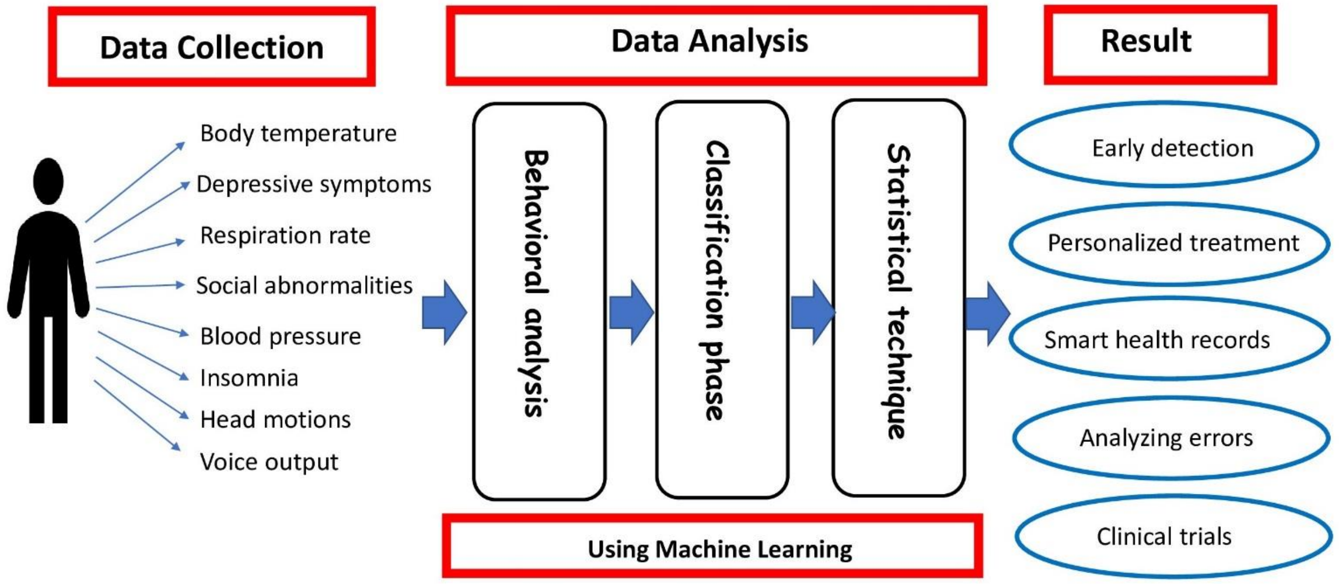


Fig. 2. Use of machine learning algorithm in behavioral analysis

II. Algorithms in detecting behavioural analysis

A. Naive bayes

The Naive Bayes algorithm is a statistical classifier which predicts the likelihood that a given tuple belongs to a given class in accordance with Bayesian analysis theorem [5]. A naive bayes method is used in healthcare to assess a patient's behavior, such as their mental health, using a multiclass classification approach and probilistic algorithm. Radwan Qasrawi et al used an empirical Bayesian approach to investigate the factors associated with depression and anxiety in school-aged children. A total of 3984 West Bank students from 5th to 9th grades, age ranging 10-15 years, were studied in refugee and community schools. During the academic year 2013-2014 data was assembled using the health behaviors of school children examination to identify risk variables related to student mental health symptoms. ML was then used to analyze the data. An analysis of 5 ML algorithms, including RF, neural networks, decision trees, SVMs, and NB were used and it was concluded that NB had the best accuracy in predicting depressive disorders (87.1) and anxiety (72.7) [6].

As a tool for predicting how a person's body will behave if she/he contracts Covid-19, Rabie et al. developed the Covid-19 Prudential Expectation Strategy (CPES). In this method there are three steps: Outlier Rejection Phase (ORP), Feature Selection Phase (FSP), and Classification Phase (CP). CPES makes use of a Statistical Naive Bayes (SNB) classifier, CP, to categorize people according to their body's reaction to Covid-19 infection. There were 2215 persons that filled out the form in total. Compared to current classification algorithms, Prudential Expectation Strategy performs better in accuracy, precision, error and recall, measuring 0.87, 0.84, 0.13 and 0.79, respectively [7].

B. SVM (Support Vector Machine)

A support vector network is the most prevailing supervised learning model which uses deep learning algorithms to map data into a high-dimensional feature space for classifying and predicting data from two groups. SVM increases effectiveness and makes healthcare more convenient and personalized for patients in a healthcare institution. The algorithm is utilized in several healthcare practices to anticipate if a patient has a particular heath issue. Its high classification accuracy, sensitivity, and specificity make it an excellent option for diagnosing diseases like heart disease, stress, and influenza.

Athira et al. [8] used SVM technique for development of multi-parameter for monitoring patients health. The researchers developed a multi parameter system based on IOT which had four parameters including heart rate, hotness and coldness, pulse rate, and oxygenation are observed using analogous sensors and in case of emergency an email is conveyed to patient's guardian. They achieved classification accuracy to 95 percent of the mpm system (Multi-Parameter Patient Monitor) by improving the algorithm of SVM.

Using behavioral risk as a predictor of cervical cancer, Degrimenci [9] investigated KNN, Random Forest, and SVM algorithms for their potential role in cervical cancer prediction. The information utilized in this study was collected from the UCI Machine Learning Repository (a library of data on cervical cancer behavior risk). A total of 72 indonesians were recruited to provide the samples. Result showed that 21 were in danger (positive) and the remaining

51 were not in danger (negative). The facts that were collected related to cervix cancer behavior which includes 19 characteristics, such as diet, hygiene practices, sexual risk, emotion, etc. Out of all the approaches that were tested, the SVM technique had the most appreciative accuracy (91.67% with sigmoid-SVM).

C. Random Forest

Random Forest (RF) algorithm is an ensemble classifier that uses a variety of decision tree models to improve presaging accuracy. Based on the set of training data, it generates several classification trees that are each trained using bootstrap sampling [10]. Random forests are effective at estimating variable significance with neural networks. Their ability to handle huge data sets with hundreds of variables makes them an excellent approach for dealing with missing data. Probabilities of contracting a disease can be forecast using a random forest model based on past diagnoses. The concept may be useful for managing risks, communicating customized health inforation, and assisting with healthcare decisions [11].

Khalilia et al. [12] trained random forest classifiers for predicting diseases using HCUP (Healthcare Cost and Utilization Project) data from National Inpatient Sample (NIS). To predict the threat of 8 chronic disease, they evaluated the effectiveness of SVM, bagging, boosting, and RF ensemble learning hinge on the Area Under the ROC Curve (AUC). Additionally, RF calculates the significance of each and every variable in the classification procedures, which helped them overcome the class imbalance issue. An average AUC of 88.79% was calculated for eight disease types using the HCUP data set and found to be promising in disease prediction.

Kazuya et al. [13] conducted a study in Japan to show a link between stroke-related search behavior and stroke-related mortality. The regression analysis used a number of characteristics as predictors, including sex, lifespan, hospitalizations, progress, strokes, etc. They identified 9476 abstracts from Japanese literature relating to stroke symptoms and signs and a score of 89.94% was achieved using age-adjusted mortality from stroke in the RF analysis. It revealed that the query with high relevance score was stroke, and that it was associated with Japan's age-adjusted mortality rate.

D. CNN (CONVOLUTED NEURAL NETWORK)

CNN is a kind of artificial neural network which uses mathematical operations that it is typically employed in image recognition, segmentation, classification and has one or more convolutional layers, and other correlated tasks. As CNN categorizes hundreds of pictures each minute, it could be useful because new photos could be categorized instantly. Photos are sent for grading to physicians when a patient comes in for screening, but they are not appropriately rated. A trained CNN is capable of making a rapid diagnosis and responding to a patient immediately. In similar fashion, the network produced these results using only one image per eye.

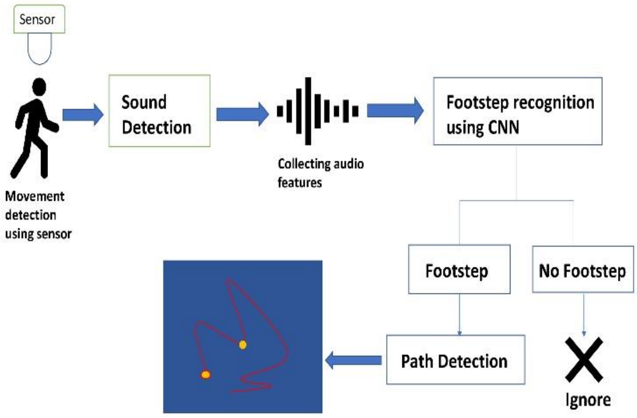

Sensor
Movement detection using sensor
Sound Detection
Collecting audio features
Footstep recognition using CNN
Footstep
No Footstep
Path Detection
Ignore

Fig. 3. An illustration of how data filtering works

Oliveira et al. [14] developed a CNN method to detect the wandering movements of Alzheimer's patients based on data gathered from non-intrusive sensors around the house (**Figure 3).** 220 paths were generated in the dataset. This data was identified by CNN using visual features (such as loops or random movements). The data was compared with 60 min and 30 min datasets and the 30 min datasets had a precision difference of 55.57%, a recall difference of 20% and a F1 score difference of 17.86%.

Pratt et al. [15] carried out a CNN based model for the classification of 5 classes of diabetic retinopathy disease. They took 80000 images from the Kaggle through which state-of-the-art DR stage classification technique was put together by utilizing complex DR characteristics such as transude on the retina, MAS, and HEMS, they employed a CNN architecture with expansion of data to designate 5 degrees of diabetic retinopathy severity.

D. Decision Tree

Decision tree is a type of supervised machine learning technique which predicts and processes data using classification and regression analogies based on real life [5]. Health systems can use decision trees to determine the initial course of treatment for behavior problems and to implement empirically validated treatment procedures [16].

In a study by Cohen et al. [17] behavioral profiles of children with Autism Spectrum Disorder are used to guide treatment decisions through CART decision trees. They compared the PDDBI with the ADOS-2 to determine its criterion-related validity. A total of 110 candidates were selected between 1.5 and 6.9 years age and grouped into 2 behavioral aspects: Receptive/Expressive Social Communication Abilities (REXSCAs) and Approach Withdrawal Problems (AWPs). Based on T-scores, various cut-off scores was evaluated in this study for domains such as sensory behaviors, traditions, repetition of language, offensive behaviour, and ability to express.

Using a decision tree algorithm, Batterham et al. analysed depression outcomes. This study showed that environmental factors are associated with various social factors that are related to depression. After 4 weeks of follow-up, patients were allowed to begin taking new antidepressants; 25% of placebo patients and 7% of zuranolone patients received new antidepressants in the phase 2 study. Decision trees were discovered to have higher susceptibility and selectivity than logistic regressions when analogous predictors were used [18].

TABLE I. RECENT STUDIES CONDUCTED ON BEHAVIOR ANALYSIS WITH THE HELP OF ML FOR THE TREATMENT OF VARIOUS DISEASE

S.No	Technique	Paper	Worked on	Accuracy	Year
1	Naïve Bayes	Radwan Qasrawi et al. [6]	Depression and anxiety	87.1%(depression) 72.7%(anxiety)	2022
		Rabie et al. [7]	Covid	87%	2022
2	SVM	Degirmenci [9]	Cervical cancer	91.67%	2022
		Athira et al. [8]	Heart rate, temperature, respiration rate, oxygen and saturation	95%	2020
3	Random Forest	Khalilia et al. [12]	8 chronic diseases	88.79%	2011
		Kazuya et al. [13]	Stroke	89.94%	2022
4	CNN	Pratt et al. [15]	Daibetic retinopathy disease	-	2016
		Oliveira et al. [14]	Alzheimer disease	82.65%	2022
5	Decision Tree	Cohen et al. [17]	Autisum spectrum disorder	-	2019
		Batterham et al. [18]	Depression	-	2009

III. CONCLUSION

As machine learning becomes more prevalent, it is being used for diagnosing diseases in multiple industries. A number of scientists have discussed the benefit of machine learning-based disease diagnostics (MLBDD) in terms of time and cost efficiency. Traditionally, diagnostic techniques are labor-intensive, expensive, and require human involvement frequently. According to WHO estimates, lifestyle changes are responsible for 30% of all deaths worldwide. These deaths can be avoided by correctly identifying the risk factors that go along with them and developing behavior modification strategies. Prevention of potentially fatal consequences requires changes in health-related behavior. Life expectancy will be increased if early diagnosis, prevention assistance, and appropriate treatment are provided as soon as possible. Machine learning algorithms will be implemented to further investigate methods like sensor based feature extraction, such as Electrocardiogram (ECG), Electroencephalogram (EEG) etc, for the diagnosis of early-stage diseases from human behavior on various platforms. The idea behind automated patient and disease monitoring activities are to conserve time and fill in when all doctors are busy, like during an emergency. The deployment of smart technology in this industry can help save lives during pandemics, such as the COVID-19 epidemic.

REFERENCES

[1] F. Yao, “Deep learning analysis of human behaviour recognition based on convolutional neural network analysis,” *Behav. Inf. Technol.*, vol. 0, no. 0, pp. 1–9, 2020, doi: 10.1080/0144929X.2020.1716390.

[2] T. Richter, B. Fishbain, A. Markus, G. Richter-Levin, and H. Okon-Singer, “Using machine learning-based analysis for behavioral differentiation between anxiety and depression,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-72289-9.

[3] D. P. Wall, R. Dally, R. Luyster, J. Y. Jung, and T. F. DeLuca, “Use of artificial intelligence to shorten the behavioral diagnosis of autism,” *PLoS One*, vol. 7, no. 8, 2012, doi: 10.1371/journal.pone.0043855.

[4] B. Kanchanatawan, S. Thika, S. Sirivichayakul, A. F. Carvalho, M. Geffard, and M. Maes, “In Schizophrenia, Depression, Anxiety, and Physiosomatic Symptoms Are Strongly Related to Psychotic Symptoms and Excitation, Impairments in Episodic Memory, and

Increased Production of Neurotoxic Tryptophan Catabolites: a Multivariate and Machine Learning," *Neurotox. Res.*, vol. 33, no. 3, pp. 641–655, 2018, doi: 10.1007/s12640-018-9868-4.

[5] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," *J. Med. Syst.*, vol. 42, no. 5, 2018, doi: 10.1007/s10916-018-0934-5.

[6] R. Qasrawi, S. P. V. Polo, D. A. Al-Halawa, S. Hallaq, and Z. Abdeen, "Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis," *JMIR Form. Res.*, vol. 6, no. 8, pp. 1–15, 2022, doi: 10.2196/32736.

[7] A. H. Rabie, N. A. Mansour, A. I. Saleh, and A. E. Takieldeen, "Expecting individuals' body reaction to Covid-19 based on statistical Naïve Bayes technique," *Pattern Recognit.*, vol. 128, no. April, 2022, doi: 10.1016/j.patcog.2022.108693.

[8] A. Athira, T. D. Devika, K. R. Varsha, and S. S. Bose, "Design and Development of IOT Based Multi-Parameter Patient Monitoring System," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020,* pp. 862–866, 2020, doi: 10.1109/ICACCS48705.2020.9074293.

[9] A. Degirmenci, "Performance Comparison of kNN , Random Forest and SVM in the Prediction of Cervical Cancer from Behavioral Risk," vol. 7, no. 10, pp. 71–79, 2022.

[10] W. Mao and F.-Y. Wang, "Cultural Modeling for Behavior Analysis and Prediction," *Adv. Intell. Secur. Informatics*, pp. 91–102, 2012, doi: 10.1016/b978-0-12-397200-2.00008-7.

[11] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[12] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, 2011, doi: 10.1186/1472-6947-11-51.

[13] K. Taira and S. Fujita, "Prediction of Age-Adjusted Mortality from Stroke in Japanese Prefectures: Ecological Study Using Search Engine Queries," *JMIR Form. Res.*, vol. 6, no. 1, 2022, doi: 10.2196/27805.

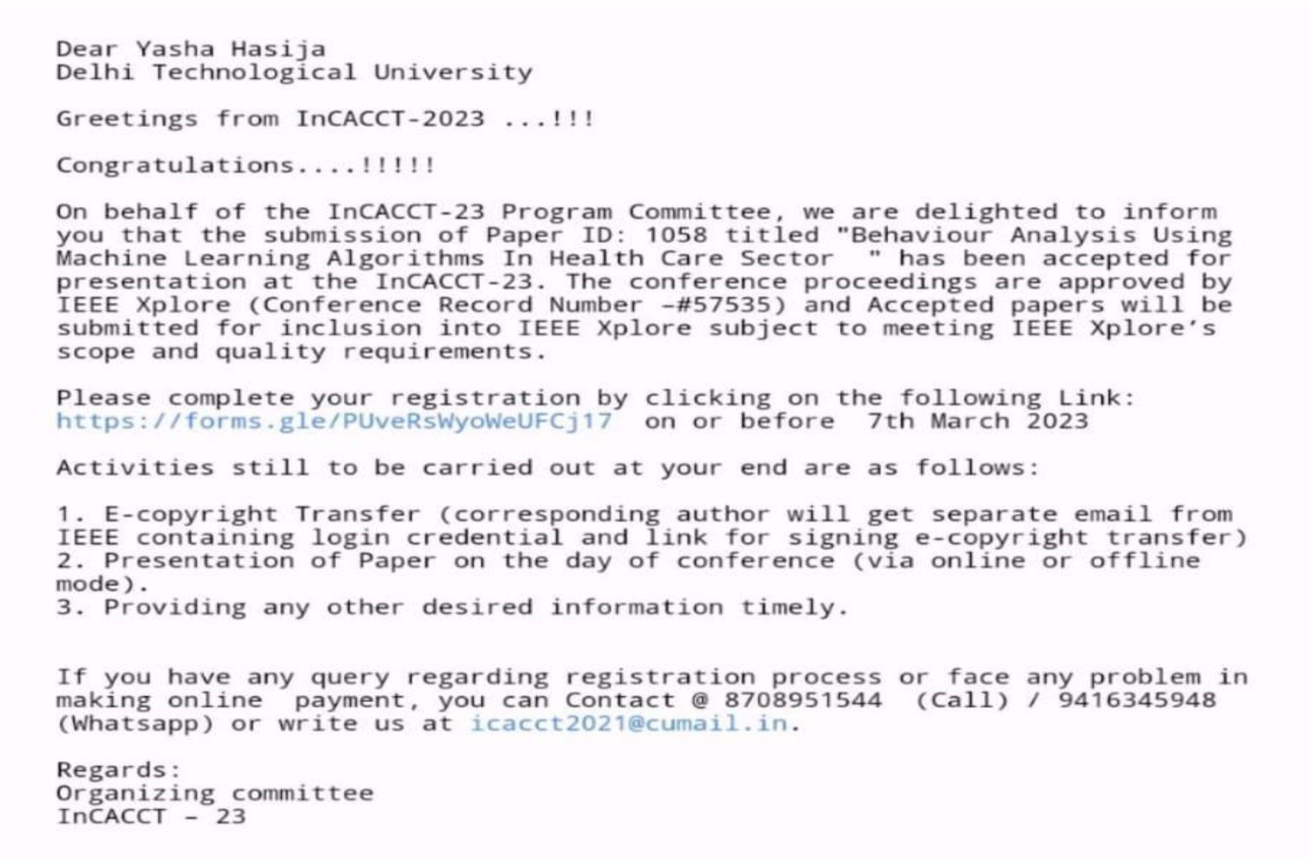
[14] R. Oliveira, R. Feres, F. Barreto, and R. Abreu, "CNN for Elderly Wandering Prediction in Indoor Scenarios," *SN Comput. Sci.*, vol. 3, no. 3, pp. 1–11, 2022, doi: 10.1007/s42979-022-01091-3.

[15] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," *Procedia Comput. Sci.*, vol. 90, no. July, pp. 200–205, 2016, doi: 10.1016/j.procs.2016.07.014.

[16] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002, doi: 10.1023/A:1016409317640.

[17] I. L. Cohen and M. J. Flory, "Autism Spectrum Disorder Decision Tree Subgroups Predict Adaptive Behavior and Autism Severity Trajectories in Children with ASD," *J. Autism Dev. Disord.*, vol. 49, no. 4, pp. 1423–1437, 2019, doi: 10.1007/s10803-018-3830-4.

[18] P. J. Batterham, H. Christensen, and A. J. Mackinnon, "Modifiable risk factors predicting major depressive disorder at four year follow-up: A decision tree approach," *BMC Psychiatry*, vol. 9, pp. 4–11, 2009, doi: 10.1186/1471-244X-9-75.



Dear Yasha Hasija
Delhi Technological University

Greetings from InCACCT-2023 ...!!!

Congratulations....!!!!!

On behalf of the InCACCT-23 Program Committee, we are delighted to inform you that the submission of Paper ID: 1058 titled "Behaviour Analysis Using Machine Learning Algorithms In Health Care Sector " has been accepted for presentation at the InCACCT-23. The conference proceedings are approved by IEEE Xplore (Conference Record Number -#57535) and Accepted papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

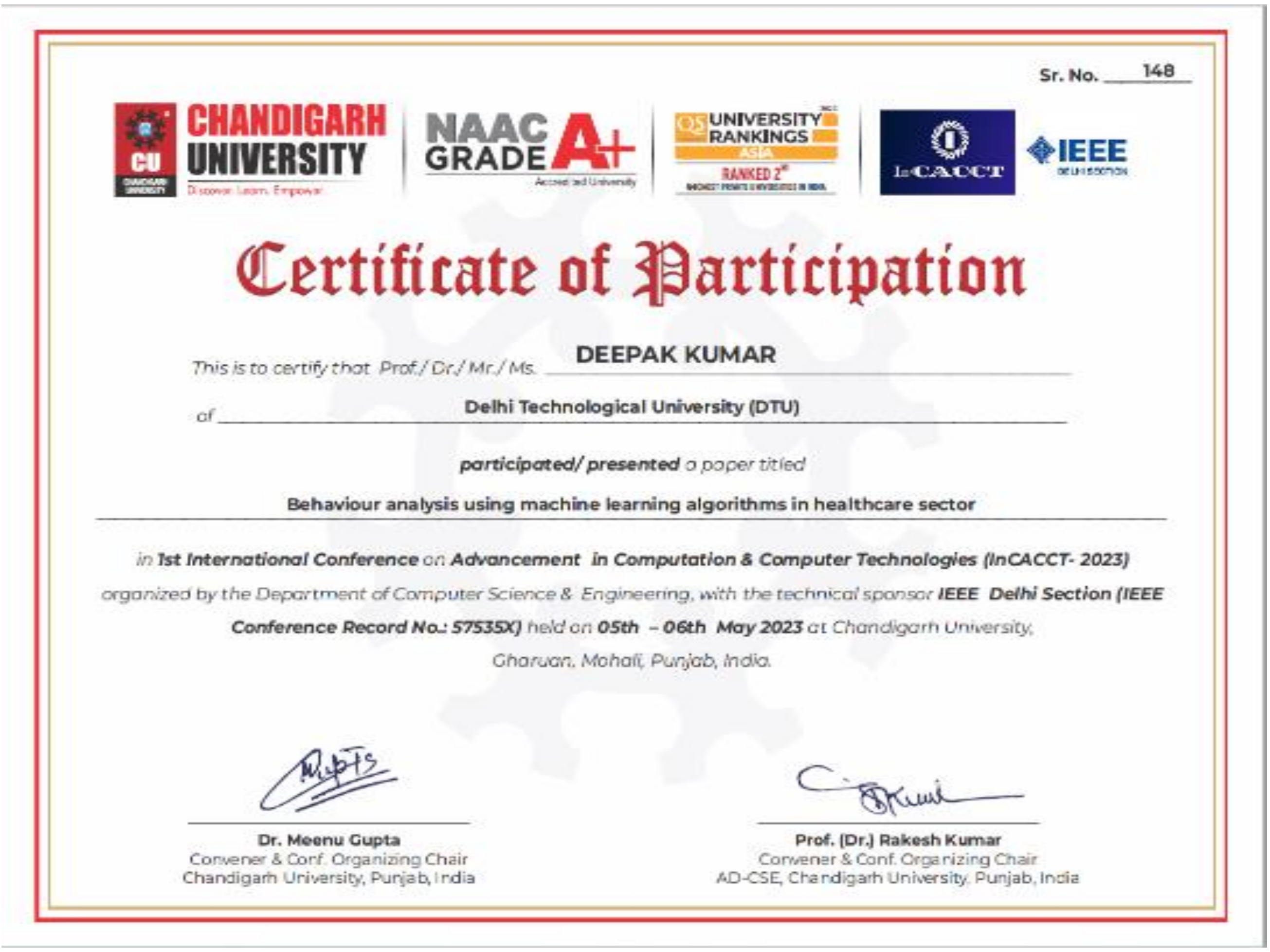
Please complete your registration by clicking on the following Link: https://forms.gle/PUveRsWyoWeUFCj17 on or before 7th March 2023

Activities still to be carried out at your end are as follows:

1. E-copyright Transfer (corresponding author will get separate email from IEEE containing login credential and link for signing e-copyright transfer)
2. Presentation of Paper on the day of conference (via online or offline mode).
3. Providing any other desired information timely.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8708951544 (Call) / 9416345948 (Whatsapp) or write us at icacct2021@cumail.in.

Regards:
Organizing committee
InCACCT - 23



Sr. No. 148

CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

NAAC GRADE A+

QS UNIVERSITY RANKINGS ASIA RANKED 2nd

InCACCT

IEEE

Certificate of Participation

This is to certify that Prof./Dr./Mr./Ms. **DEEPAK KUMAR**

of **Delhi Technological University (DTU)**

participated/ presented *a paper titled*

Behaviour analysis using machine learning algorithms in healthcare sector

in ***1st International Conference*** *on* ***Advancement in Computation & Computer Technologies (InCACCT- 2023)*** *organized by the Department of Computer Science & Engineering, with the technical sponsor* ***IEEE Delhi Section (IEEE Conference Record No.: 57535X)*** *held on* ***05th – 06th May 2023*** *at Chandigarh University, Gharuan, Mohali, Punjab, India.*

Dr. Meenu Gupta
Convener & Conf. Organizing Chair
Chandigarh University, Punjab, India

Prof. (Dr.) Rakesh Kumar
Convener & Conf. Organizing Chair
AD-CSE, Chandigarh University, Punjab, India

turnitin

Similarity Report ID: oid:27535:36255800

PAPER NAME

diabetes writeup.docx

WORD COUNT 8375 Words	CHARACTER COUNT 48139 Characters
PAGE COUNT 43 Pages	FILE SIZE 1.8MB
SUBMISSION DATE May 26, 2023 7:54 PM GMT+5:30	REPORT DATE May 26, 2023 7:55 PM GMT+5:30

● 11% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- Crossref database
- 7% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College Of Engineering)
Bhawana Road, Delhi-110042

Certificate

I hereby certify that the Project Dissertation titled **"Breaking Barriers in Diabetes Management: Unleashing the Potential of Machine Learning in Diagnoses and Treatment"** which is submitted by **Deepak Kumar (2K21/MSCBIO/10)**, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is recorded for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or anydiploma to this university or elsewhere.

Place: Delhi
Date : 29.05.23

29.05.23

Prof. Yasha Hasija

(Supervisor)

Professor

Department of Biotechnology

Delhi Technological University

30/05/2023

Prof. Pravir Kumar

Head of Department

Dean (International Affairs)

Department of Biotechnology

Delhi Technological University