NEAR EAST UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

DEPARTMENT OF ARTIFICIAL INTELLIGENCE ENGINEERING

FRAMEWORK FOR DIABETES PREDICTION
WITH MACHINE LEARNING
CLASSIFICATION ALGORITHM.

M.Sc. THESIS

Gershon Oghenetega OMORAKA

Nicosia June

2023

NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF ARTIFICIAL INTELLIGENCE ENGINEERING

FRAMEWORK FOR DIABETES PREDICTION
WITH MACHINE LEARNING
CLASSIFICATION ALGORITHM.

M.Sc. THESIS

Gershon Oghenetega OMORAKA

Supervisor
Prof. Dr. Fadi AL-TURJMAN

Nicosia June

2023

Approval

We certify that we have read the thesis submitted by Gershon Oghenetega Omoraka titled **"Framework for diabetes prediction with machine learning classification algorithm."** and that in our combined opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Educational Sciences.

Examining Committee	Name-Surname	Signature
Head of the Committee:	Assoc. Prof. Dr. Sertan Serte	
Committee Member:	Assis. Prof. Ibrahim Adeshola	
Supervisor:	Prof. Dr. Fadi Al-turjman	

Approved by the Head of the Department

13/07/2023

Prof. Dr. Fadi Al-turjman
Title, Name-Surname
Head of Department

Approved by the Institute of Graduate Studies

…../…../20…

Prof. Dr. Kemal Hüsnü Can Başer
Head of the Institute

Declaration

I hereby declare that all information, documents, analysis, and results in this thesis have been collected and presented according to the academic rules and ethical guidelines of the Institute of Graduate Studies, Near East University. I also declare that as required by these rules and conduct, I have fully cited and referenced information and data that are not original to this study.

Gershon Oghenetaga Omoraka

10/07/2023

First and foremost, I want to express my gratitude to God Almighty for His mercies, direction, and fortitude along this trip. My perseverance and resolve are the result of his unfailing support.

My profound gratitude goes out to my parents for their unfailing support, steadfast love, and unfailing belief in me. Their sacrifices, encouragement, and support have been crucial to my academic endeavors.

My supervisor, Prof. Fadi Al-Turjman, deserves all the credit for his outstanding advice, know-how, and unwavering support during this research project. His astute criticism, perseverance, and encouragement had a crucial role in determining the scope and caliber of this work.

I want to thank the distinguished jury members for their time, knowledge, and insightful comments as they evaluated my thesis. Their kind criticism and recommendations have tremendously helped me to improve my research.

Finally, I'd like to thank everyone who has helped me in different ways, both visible and invisible. They have been instrumental in helping me overcome obstacles and attain this goal because of their support, compassion, and faith in my talents. Please accept my sincere gratitude for your contributions to my thesis work on behalf of everyone listed and possibly not expressly named. Your unshakable confidence in me and support have served as a constant source of motivation.

Abstract

Framework for diabetes prediction with machine learning classification algorithm

Omoraka Gershon Oghenetega

MSc, Department of Artificial Intelligence Engineering

June 2023, 98 pages

Diabetes is a chronic metabolic condition that faces millions of people around the globe. Like prediabetes; needs to be detected on time and to reduce the risk, it is essential for the development of efficient preventive and management strategies. The goal of this study is to improve the classification of a diabetes dataset by adding the HbA1c property and establishing two new classes that correspond to hypoglycemia stages 1 and 2. The objective is to enhance risk assessment and make individualized therapies for diabetics possible. After preprocessing the dataset, a Random Forest Classifier with tuned hyperparameters is used. The effectiveness of the model is assessed via cross-fold validation. The findings show impressive accuracy and highlight the importance of the HbA1c characteristic in differentiating between stages of hypoglycemia. The enlarged classification system helps with individualized treatment strategies and enables a more thorough understanding of the disease spectrum. The study aids in the management of diabetes by facilitating tailored interventions and better risk stratification. Future work might focus on developing classification models and investigating new features. The findings have implications for future developments in diabetes care as well as decision support systems. A web application was created to enable the trained model's deployment, practical usage, and accessibility. The web application offers the patient a simple interface via which they can input the necessary data and get tailored assessments of their risk for diabetes.

The framework has the potential to facilitate effective preventative actions, individualized management regimens, and early identification of those who are at risk for diabetes.

Keywords: diabetes prediction, hypoglycemia detection, classification model, Random Forest, web application, feature importance.

Soyut

Makine öğrenimi sınıflandırma algoritması ile diyabet tahmini için çerçeve

Omoraka Gershon Oghenetega

Yüksek Lisans, Yapay Zeka Mühendisliği Bölümü

Haziran, 2023, 98 sayfa

Diyabet, dünya çapında milyonlarca insanın karşılaştığı kronik bir metabolik durumdur. Prediyabetikler gibi; zamanında tespit edilmesi ve riski azaltması gereklidir, etkili önleme ve yönetim stratejilerinin geliştirilmesi için gereklidir Bu çalışmanın amacı, HbA1c özelliğini ekleyerek ve iki yeni sınıf oluşturarak bir diyabet veri setinin sınıflandırmasını iyileştirmektir. Bu, hipoglisemi evre 1 ve 2'ye karşılık gelir. Amaç, risk değerlendirmesini geliştirmek ve diyabet hastaları için bireyselleştirilmiş tedavileri mümkün kılmaktır. Veri kümesini önceden işledikten sonra, ayarlanmış hiperparametrelere sahip bir Rastgele Orman Sınıflandırıcısı kullanılır. Modelin etkinliği, çapraz katlama doğrulaması yoluyla değerlendirilir. Bulgular etkileyici bir doğruluk gösteriyor ve HbA1c özelliğinin hipogliseminin evrelerini ayırt etmedeki önemini vurguluyor. Genişletilmiş sınıflandırma sistemi, bireyselleştirilmiş tedavi stratejilerine yardımcı olur ve hastalık spektrumunun daha kapsamlı bir şekilde anlaşılmasını sağlar. Çalışma, özel müdahaleleri ve daha iyi risk sınıflandırmasını kolaylaştırarak diyabet yönetimine yardımcı olur. Gelecekteki çalışmalar, sınıflandırma modelleri geliştirmeye ve yeni özellikleri araştırmaya odaklanabilir. Bulgular, diyabet bakımı ve karar destek sistemlerinde gelecekteki gelişmeler için çıkarımlara sahiptir. Eğitilen modelin devreye alınması, pratik kullanımı ve erişilebilirliği için web uygulaması oluşturulmuştur. Web uygulaması, hastaya gerekli verileri girebilecekleri ve diyabet risklerine ilişkin özel değerlendirmeler alabilecekleri basit bir arayüz sunar.

Çerçeve, etkili önleyici eylemleri, bireyselleştirilmiş yönetim rejimlerini ve diyabet riski taşıyanların erken tespitini kolaylaştırma potansiyeline sahiptir.

Anahtar Kelimeler: diyabet tahmini, hipoglisemi tespiti, sınıflandırma modeli,

Random Forest, web uygulaması, özellik önemi.

Table of Content

List of Tables

Page

List of figures

List of Abbreviations

AI: Artificial Intelligence

Cr: Creatinine ratio

BMI: Body Mass Index

HbA1c: Glycated hemoglobin

FPG: Fasting Plasma Glucose

MLP: Multilayer Perceptron

OGTT: Oral Glucose Tolerance Test

PAAS: Platform as a Service

RF: Random Forest

SVM: Support Vector Machine

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background of the study

High blood glucose levels and improper protein and fat metabolism are symptoms of diabetes, a chronic illness. The cells' inability to effectively use the insulin that is being produced, or the absence of insulin synthesis by the pancreas, causes blood glucose levels to increase (Felizardo et al., 2021) . Diabetes comes in four main varieties Type 1, in which the pancreas does not generate insulin; Type 2, in which the body cells are resistant to the effects of the insulin being produced and the production of insulin gradually diminishes over time; as well as gestational diabetes, which appears during pregnancy and raises the risk of certain difficulties both during labor and delivery, and prediabetics, whose blood sugar levels are elevated but not yet high enough to be categorized as type 2 diabetes (S. Roy et al., 2022) . The majority of people with type 1 diabetes are children, teenagers, and young adults. The reason or causes are unknown. It is believed that a combination of environmental factors and genetic predisposition causes type 1 diabetes. Due to type 2 diabetes's delayed onset and usual presentation without the rapid metabolic disturbance seen in type 1 diabetes, the precise time of commencement is difficult to determine. As a result, it is more difficult to tell apart aberrant from normal behavior, there is a long pre-detection period, and up to 50% of cases in the general population may receive a wrong diagnosis. Information on the prevalence of type 2 diabetes that has been clinically diagnosed can be useful in planning for health services, but this information is meaningless without knowledge on the prevalence of diabetes that has not yet been clinically identified. The OGTT has been utilized in specialized studies for epidemiological research to assess the presence and absence of disease (Forouhi

& Wareham, 2010). Prediabetes can be recognized using elevated glycated hemoglobin A1c (HbA1c) values, impaired fasting glucose (IFG) levels, or impaired glucose tolerance. FPG measurements above normal are used to identify cases with impaired fasting glucose levels. Impairment of post-meal insulin secretion and muscle insulin resistance are features of impaired glucose tolerance (Afsaneh et al., 2022).

Obesity, a family history of diabetes, prior signs of impaired blood glucose tolerance, a history of giving birth to infants weighing more than 4000 grams, fetal defects in the past, a history of stillbirths or multiple miscarriages, maternal age over 35, and polycystic ovary syndrome are just a few of the factors that increase the risk of gestational diabetes. These elements increase the risk of getting gestational diabetes during pregnant. Between 24 and 28 weeks of pregnancy is when gestational diabetes is frequently discovered. Blood is drawn to measure the blood sugar level one hour after 50 g of glucose were ingested in the form of water. If the blood sugar level is greater than 140 mg/dl, a second 3-hour oral glucose tolerance test is required to confirm the diagnosis. (Bhuiyan et al., 2021). The 3-hour oral glucose tolerance test requires an 8-hour fast, a blood sugar reading, the injection of 100 g of glucose into the water, and further blood tests one, two, and three hours later. To be diagnosed with gestational diabetes, two or more of the four blood glucose readings must be higher than the reference range. (Chou et al., 2023).

Even though diabetes cannot be treated completely, it can be controlled and avoided if a trustworthy early prediction can be made. Therefore, many approaches have been put forth in recent study employing statistical analysis and machine learning techniques for the prediction of diabetes. Presenting a classification algorithm for the prediction of diabetics takes a machine learning classifier into consideration and training the algorithm with some historical diabetics' data (Shah & Patel, 2019) . Machine learning (ML) is the use of numerous computer techniques that are capable of being improvised spontaneously through experimentation and the usage of data. To forecast or make decisions, algorithms build a model that depends on training data, which are samples of real data. In the realm of medicine, artificial intelligence has proved extremely helpful in the diagnosis and prognosis of several diseases.

Diabetes mellitus, a progressive illness that can worsen if left untreated and which affects 463 million people worldwide, is swiftly taking the role of a potential pandemic. Improved methods for detecting prediabetes are required to reduce the chance that it may develop into diabetes (Jahanur Rahman et al., 2020) . Diabetes biomarkers are biological molecules that can be discovered in blood, other bodily fluids, or tissues that can indicate whether the development of diabetes is normal or pathological. To determine how effectively the body responds to a disease or condition therapy, a biomarker may be utilized. also known as a signature molecule and a molecular marker. a biological molecule that may be discovered in tissues, bodily fluids, or blood and is a symptom of a condition, illness, or a normal or pathological process. To determine how effectively the body responds to a disease or condition therapy, a biomarker may be utilized. also known as a signature molecule and a molecular marker (Rjoob et al., 2020). The phrase "biomarker," also known as a "molecular marker" or a "signature molecule," must be defined clearly and separated from the term "risk factor." A biomarker is a biological molecule that may be discovered in blood, other bodily fluids, or tissue and is used to indicate if a process, condition, or sickness is healthy or harmful. To determine how effectively the body responds to a disease or condition therapy, a biomarker may be utilized. (Lyman et al., 2021). Biomarkers make it possible to identify people with subclinical disease before overt clinical disease manifests. They enable the monitoring of patient responses to therapeutic or preventive interventions as well as the application of preventive measures at the subclinical stage. They support the investigation of illness causes and the assessment of innovative preventive and therapeutic treatments by providing alternative end points for intervention studies. To put it another way, biomarkers enable us to monitor the effects of both clinical and subclinical disease. Below is the one important biomarker for the diagnosis of diabetes known as the glycated hemoglobin.

Glycated Hemoglobin A1C (HbA1c): The hemoglobin is a protein found in red blood cells. It is in charge of distributing oxygen throughout your body and giving blood its red hue. (Das et al., 2020).

Test for Glycated Hemoglobin A1C (HbA1c): The average blood sugar (glucose) levels over the past three months can be calculated using the marker HbA1c (1). It can thus be used to identify pre-diabetes, diabetes, and to rate the effectiveness of your diabetic therapy. Hemoglobin A1c, glycated hemoglobin, and simply A1c are other names for HbA1c (Tao et al., 2022). Glucose, a kind of sugar that is derived from the food you eat, is present in the blood. Your cells receive energy from glucose. Your cells can more easily absorb glucose with the help of the hormone insulin. If you have diabetes, either your body doesn't create enough insulin or your cells don't use it well. Your blood sugar levels increase as a result of the difficulty of glucose to enter your cells. Red blood cells contain a protein called hemoglobin, which binds to blood glucose. As your blood glucose levels increase, your hemoglobin will become more and more glucose-coated. An A1C test can estimate the proportion of your red blood cells that have hemoglobin coated with glucose.

1.2 Artificial Intelligence in disease diagnosis

In the past forty years, the discipline of computer science and many of its application fields have seen tremendous advancements in the field of artificial intelligence (AI). Although AI hasn't yet met predictions from the 1970s and 1980s, it has made great advances in planning, learning, modeling, automated reasoning, and knowledge representation. (Dagliati et al., 2017). The 'quadruple aim' for healthcare—improving the wellness of the population, the satisfaction of patients with treatment, caregiver experience, and decreasing the continually rising cost of care—presents significant challenges for healthcare systems everywhere. (Mastoli et al., 2022). The healthcare sector may be able to address some of these supply-and-demand problems by utilizing technology and artificial intelligence (AI). The growing availability of multi-modal data (genomics, economic, demographic, clinical, and phenotypic) signals a time when medical technology will fundamentally transform models of healthcare delivery through AI-augmented healthcare systems when combined with technological advancements in mobile, the internet of things (IoT), processing power, and data security (Sampath et al., 2020). Inevitably, AI offers a plethora of prospects in the field of health care, where it can be used to improve a number of standard medical procedures, from disease diagnosis to determining the best course of action for patients with serious illnesses like cancer. Artificial intelligence (AI)-enhanced robotic surgical equipment can improve surgical outcomes by reducing physical fluctuations and delivering real-time information to the physician (Goel & Satish, 2023) . Traditional diagnostic procedures are expensive, time-consuming, and frequently need for human involvement. Traditional diagnosis techniques are limited by the capacity of the individual, however ML-based systems are limitless and do not feel human tiredness. As a result, it may be possible to establish a method for illness diagnosis when unexpectedly large patient populations enter the healthcare system. X-ray and MRI pictures as well as tabular data on the diseases, age, and gender of patients are used to construct health smart systems (Ogwo et al., 2019).

1.3 Statement of the problem

Persistently high blood glucose levels can lead to serious problems affecting the heart and blood vessels, eyes, kidneys, nerves, and teeth. Additionally, people with diabetes are more susceptible to infections. In almost every industrialized country, diabetes is the leading cause of cardiovascular disease, blindness, renal failure, and lower limb amputation. (Kandhasamy & Balamurali, 2015). This problem has been on the desk of researchers in different fields to find the best and more relievable way to understand the diabetic status of patients. Most solutions have been brought out using artificial intelligence techniques but are only useful for binary classification, which mean tells if the patient is diabetic or not but has not really focused on the fact that there are other stages of diabetics in a patient. Implementing a model for this current problem will be able to create a more defined prediction and distinction of the output result in the prediction.

1.2 Research Aims and Objective

Based on a vast quantity of data, artificial intelligence (AI) can draw sophisticated conclusions. The two pillars of the AI explosion in 2021 will be machine learning (ML) and deep learning. Due to the increase in processing resources and the accompanying rapid improvement in computer performance, these technologies have improved tremendously (Schepart et al., 2023) . We introduce diabetes prediction models and AI/ML-based medical devices in this work. In order to enhance the healthcare system, it is also important to research the most up-to-date machine learning techniques for diagnosing diabetics and to use these techniques to classify and forecast the many types of diabetes.

The aging global population and contemporary lifestyles are significantly increasing the need for high-quality healthcare and well-being services since data is the foundation of an AI system. Wearable biosensing networks' remarkable high-dimensional data processing capabilities have recently drawn significant academic and technical interest to the integration of wearables with artificial intelligence (AI). (Zhang et al., 2023) , there by generating enough data for the training of AI system that can be integrated in health care system. This work also aims to create an end-to-end application for patients and health practitioners to easily diagnose diabetics using an ML model integrated within the backend of the web application. Making the process handy and assessable by any user. This work will cover not just the model training and evaluation but also the deployment of the model in a cloud-based server for more scalability.

1.3 General Approach

This work proposes a framework using the random forest (RF) classifier as the base classification model for the diagnosis and prediction. The classifier is trained with the diabetic dataset which will be processed and cleaned in other to avoid unwanted prediction results. Under the same experimental settings and dataset, the model is subjected to extensive experiments and hyperparameter adjustment to get the highest prediction accuracy (Sulistyawati & Murtadho, 2020). There are several classes in the categorization. In the sense that there are more than 2 classes to be predicted by the model. The proposed system shows a more complicated modeling of the data to the machine learning algorithm.

The dataset consists of medical information laboratory analysis with the input features: Number of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides (TG) and HDL Cholesterol, HBA1C and an output class or outcome. More information will be discussed in detail in the work and how this data can be used to create the desired model for the problem statement. The model for classification will be examined using the various model assessment metrics, including accuracy, sensitivity, AUC, and F1 score. This assessment indicator demonstrates how effectively the algorithm used for machine learning picks up new information from the data and how well it can perform on test data or new data. As well as in the real practice where the model will be deployed as a web API. This covers the design of a web app that can be easily used by the individual to predict their diabetics status. This web app will be built with the flask python framework to run the backend or the server side and the front end which serves as the user's interface.

1.4 Significance of the research

The study has a number of important ramifications for categorizing diabetes and managing patients, including:

- Enhanced Classification: A more thorough and in-depth comprehension of the disease spectrum is provided by the addition of hypoglycemic phases to the initial three classes of the three-class classification. This enables medical practitioners to recognize and classify individuals with varying levels of risk, resulting in individualized treatment plans and actions.

- Personalized Interventions: The research provides individualized therapies for people with diabetes by taking the HbA1c attribute and considering hypoglycemic stages. Depending on the precise hypoglycemia stage, healthcare providers can execute focused treatments that can improve patient outcomes overall.
- Improved Risk Stratification: The research improves risk stratification in diabetes management. By incorporating additional classes based on hypoglycemia stages, healthcare professionals can identify individuals at different levels of risk for developing hypoglycemia and adjust treatment plans accordingly. This proactive approach helps prevent complications and improves patient safety.
- Decision Support Systems: Decision support systems for managing diabetes can incorporate the enlarged classification scheme. These technologies can give healthcare workers real-time direction and recommendations based on the HbA1c characteristic and the recognized hypoglycemic stages, supporting them in making defensible judgments regarding patient care.

1.5 Structure of the research

The research work entails different sections starting with the introduction with the Background and context of the research topic. The literature review section presenting the epidemiology of diabetics and other related works in regards to the prediction of diabetes using machine learning. Methodology section, Describing the dataset used in the research,

explanation of the data preprocessing steps, including data cleaning and categorical feature encoding and details of the feature engineering process, also an overview of the Random Forest Classifier and its hyperparameter tuning, as well as Cross-fold validation methodology and performance evaluation metrics

CHAPTER 2

2.0 LITERATURE REVIEW

2.1 Epidemiology of Diabetes

Due to the global growth in obesity rates, the incidence of diabetes has increased over the past several decades (Safaei et al., 2021) . Given the early morbidity, mortality, decreased life expectancy, and huge financial burden it causes on patients, caregivers, and healthcare systems, this poses a serious threat to the public's health. Diabetes diagnosis and categorization have received substantial scrutiny and debate throughout time. Various diagnostic standards for diabetes have been produced by experts from the World Health Organization and American Diabetes Association, generally based on fasting or 2-hour post-load glucose levels. The use of glycated hemoglobin (HbA1c) for diagnosis is a topic of continuous discussion, despite some convergence and divergence in the opinions of the many parties involved. Type 1 and type 2 diabetes are the two primary forms, with type 2 accounting for more than 85% of all occurrences. The classification of diabetes' etiology is now universally accepted. (Forouhi & Wareham, 2010)

All the chronic metabolic diseases that make up diabetes mellitus have increased blood glucose levels because of the body's inability to manufacture insulin, resistance to its effect, or both (Ying et al., 2021). There are four clinically different kinds that this group of disorders may be split into.

- **Type 1 diabetes:** This category of diabetes is formerly known as juvenile diabetes (Sreenivasu et al., 2022). Type 1 diabetes is an autoimmune condition that develops when the immune system attacks insulin-producing beta cells. Insulin is a hormone that aids in controlling blood glucose levels and is necessary for the cells to utilize blood sugar for energy (S. Roy et al., 2022). This causes elevated blood sugar levels in the body prior to therapy. The beta cells in the pancreas, a large organ located behind the stomach, which are in charge of manufacturing insulin, are damaged by the immune system of the body when type 1 diabetes develops. As a result, the body stops producing insulin, which prevents glucose, a form of sugar, from entering the cells, where it is typically turned into energy. (Del Giorno et al., 2023) . The main therapy for type 1 diabetes is insulin. In this situation, individuals must regularly administer insulin injections or use an insulin pump to adequately control their illness. In order to reduce the risk of serious short- or long-term health problems, usually referred to as diabetic complications, blood sugar levels must be closely monitored and controlled. The patient's health might suffer greatly if insulin is not managed properly.
- **Prediabetes:** A person who has increased blood sugar levels but does not satisfy the diagnostic criteria for diabetes is said to have prediabetes, which is a disease that precedes diabetes. Prediabetic individuals may have both reduced glucose tolerance and impaired fasting glucose (Thirunavukkarasu & Umapathy, 2020).
- **Type 2 diabetes:** Type 2 diabetes that has persisted prevents the body from properly using insulin. Insulin sensitivity may be seen in type 2 diabetics. Young and middle-aged people tend to be more vulnerable to this kind of diabetes (Tiwari et al., 2021). Insufficient insulin production or inefficient insulin

utilization are the hallmarks of type 2 diabetes. Because of this, there is an overabundance of glucose in the bloodstream and insufficient glucose is supplied to the body's cells. If not treated appropriately, this illness causes increased blood sugar levels, which can harm general health.

2.2 Machine learning in predicting of diabetes

The application of machine learning and artificial intelligence in predicting and diagnosing diseases represents a promising area of research, offering significant potential for advancements in the health and medical field. Researchers are actively exploring these technologies to develop innovative solutions that can revolutionize healthcare practices. In this line of study, some previous work has been done to make sure the diagnosis of diabetics is done in an easy matter as artificial intelligence in brought into the picture (Patra et al., 2022). Given that AI and machine learning have become prevalent in the sectors related to healthcare and not transmittable chronic illnesses, their real medical application rate is quite low owing to the lack of explanation of these intricate algorithms or models.

(Kumari & Singh, 2013) employed neural networks in their work. The characteristics are gender, age, height, weight loss, and Increased thirst, hunger, and appetite, as well as nausea, fatigue, skin infections, and blurred vision. The 100 elements in the dataset were chosen at random, and 70 of them were used to train the model and the rest to test it. A low-cost, effective method of diabetes detection at home is the goal of the investigation. There are 13 characteristics in it that don't need to be clinically examined, saving money. 28 nodes make up the neural network, of which 13 nodes are input nodes, 14 nodes are hidden nodes, and 1 node is an output node. The patient is not impacted if the output is 0, but he is regarded to be affected otherwise.

However, there were just 20 test instances considered. Some of the variables, such as increased hunger and thirst, are vague and cannot unquestionably be caused by diabetes because the dataset is not specified. The accuracy attained using a back-propagation technique and an artificial neural network was 92.8%. There is no need to go to any clinics, and it is inexpensive.

A machine learning-based e-diagnosis system was proposed by (Chang et al., 2022). In this paper, machine learning techniques including Naive Bayes, J48, and Random Forest were used. The filling in of missing data and the preparation steps for the 3 factor and 5 factor attribute selection are finished. The J48, Naive Bayes, and Random Forest algorithms, with scores of 75.65%, 79.13%, and 79.57%, respectively, produced the best accuracy ratings. The accuracy, precision, sensitivity, and specificity of each algorithm are evaluated based on its performance. To enhance the model, the decision-making process is evaluated and for binary classification, a Naive Bayes model performs best with a smaller number of features, whereas random forest performs best with a larger number of features.

Algorithms based on artificial intelligence are being used to identify and analyze iris scans in order to diagnose diabetes. By utilizing several ML algorithms, (Samant & Agarwal, 2018) investigate the diagnosis of diabetes by changes in pigmentation in specific regions of the iris. They get the iris and crop out specific regions using pre-image processing techniques. They then employ statistical, wavelet, and textural aspects to observe the variations in tissue pigmentation. Finally, the presence of diabetes in the patient is determined using five classifiers. According to their findings, random forests perform better than other classifiers.

(Hasan et al., 2020) Different ML classifiers (k-nearest Neighbors, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron

(MLP) were employed in their suggested robust framework for diabetes prediction. This includes choosing features, normalization of data, feature rejection, K-fold cross-validation, and filling in missing values. This is an ensemble technique which involves the combination of several classification models and later checking their average accuracy and performance.

Researchers used a variety of machine learning algorithms to evaluate the potential for creating predictive models for calculating the risk of type 2 diabetes (T2D) based on electronic medical information. To determine the likelihood that these people will acquire T2D during the next six to twelve months, they examined demographic, clinical, and laboratory data from over two thousand individuals (Mani et al., 2012). The researchers accurately predicted the onset of type 2 diabetes 365 days and 180 days before the actual diagnosis of diabetes, achieving an Area Under the Curve (AUC) of more than 0.8.

2.3 Machine learning classification

Machine learning, a subfield of computer science and artificial intelligence (AI), focuses on using data and algorithms to imitate human learning processes. With this method, the system's accuracy increases over time as it absorbs and modifies the available data. (Maheshwari et al., 2020).

Classification is regarded as a subset of supervised machine learning in the field of artificial intelligence. Making a simple model that maps the distribution of class labels based on predictor characteristics is the main goal of supervised learning. (Cujilan et al., 2022) . Classification is a flexible method for grouping data into different categories. It has a wide range of uses in areas including voice recognition, picture classification, fraud detection, and email spam detection. The resultant classifier is used to give class labels to the testing instances in circumstances where the value of the class label is unknown, but the values of the predictor attributes are known (Kotsiantis et al., 2006). In a lighter note, supervised learning is done when the output or the target variables are known. Classification works with output variables that are categories or non-continuous values as opposed to regression, which uses an output variable with a continuous value. For instance, the outcome of categorization can be a binary category like "Diabetics" or "Non-diabetics". The method of classification includes input and output information since it employs labeled input data. (Al-Turjman et al., 2022) . Next section presents the most used type of classification techniques.

2.3.1 Binary classification

In machine learning and statistics, binary classification is a fundamental activity where the goal is to categorize or classify an input into one of two potential categories or classes. It entails giving each input a binary label, commonly denoted as 0 or 1, depending on its properties or qualities (Rabiha et al., 2021). The input data for binary classification is frequently represented as a series of feature vectors, where each vector holds a set of characteristics or measurements that characterize the input. The objective is to create a model or algorithm that can learn from labeled training data, where each input is connected to its corresponding class label, and then predict inputs that haven't yet happened or that will happen in the future (Fahim et al., 2022).

2.3.2 Multi - Class classification

To categorize or classify an input into one of many potential classes or categories is the goal of the machine learning problem known as multi-class classification. Multi-class classification involves placing an input into one of three or more classes, in contrast to binary classification, which has just two classes (Mary et al., 2023) .

Similar to binary classification, feature vectors are used in multi-class classification to represent the input data. Each vector includes measurements or properties that characterize the input. The objective is to create a model or algorithm that can learn from labeled training data, where each input is connected to its corresponding class label, and then predict inputs that haven't yet happened or that will happen in the future (Bidari et al., 2021).

2.3.3 Multi – Label classification

The goal of the machine learning job known as "multi-label classification" is to categorize or assign numerous labels to an input instance. Multi-label classification enables the simultaneous assignment of many class labels, in contrast to binary or multi-class classification, which assign instances to a single class (Sangkatip & Phuboon-Ob, 2020) . Similar to other classification tasks, multi-label classification involves feature vectors as the representation of the input data. Each vector includes measurements or properties that characterize the input. Each instance, however, may have numerous labels attached to it rather than just one (Usman et al., 2023) . Creating a model or algorithm that can learn from labeled training data, where each input is connected to a set of class labels, is the aim of multi-label classification. The model's job is to forecast the appropriate labels for unknown or upcoming inputs based on the relationships and patterns discovered from the training data.

2.3.4 Imbalanced classification

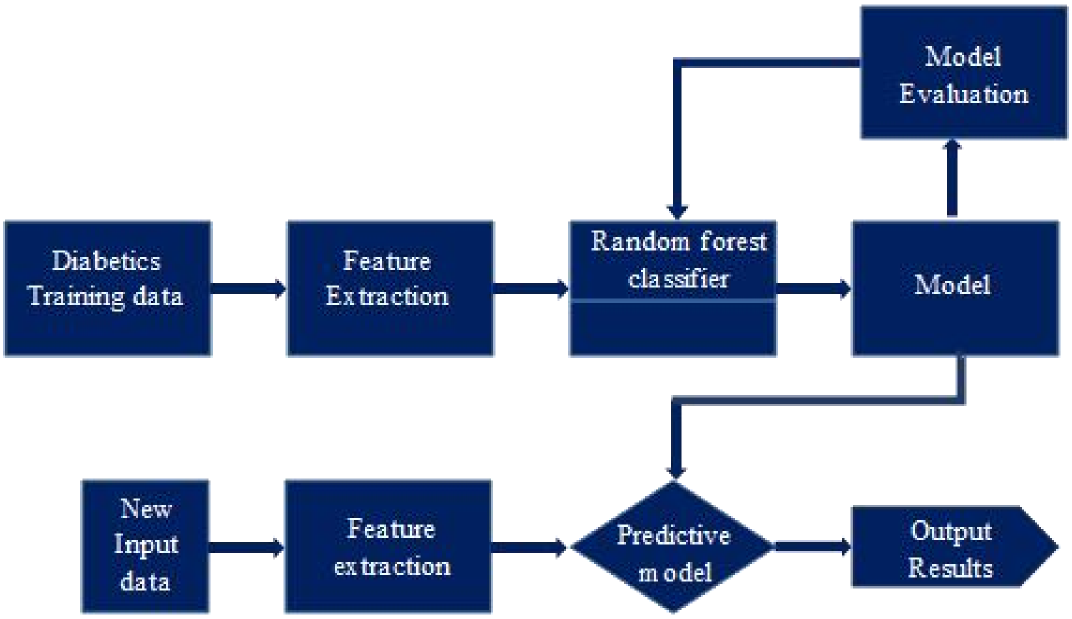
A situation in machine learning when the distribution of class labels in the training data is noticeably skewed or lopsided is known as imbalanced classification, also known as imbalanced learning or class imbalance. It denotes an uneven representation of classes because one class has a significantly higher number of instances than another class or classes (Huang et al., 2020) . In an unbalanced classification issue, the dataset is dominated by the majority class (commonly called the negative class), while the minority class (positive class) includes comparatively fewer examples (Hassan & Amiri, 2019a). Traditional classification algorithms may have difficulties as a result of this class imbalance since they are prone to favor the majority class and may find it difficult to anticipate the minority class.

CHAPTER 3

3.0 MATERIALS AND METHODS

The materials and procedures employed for the study are the main topics of this section. includes complete disclosure of the dataset, its statistical analysis, and feature selection through data cleaning and correlation. Next the classification model is built using the scikit-learn python library and with the various evaluation techniques, the model is been examined for better performance in prediction. This section also covers the development of the simple web application for deploying the model. In this case the model can practically use for prediction diabetics. Where the user of the patients enters the data from his or her test result, with data is then passed to the model for prediction.

Figure 1: General framework of the proposed technique



3.1 Data Collection and Statistical Description of Dataset

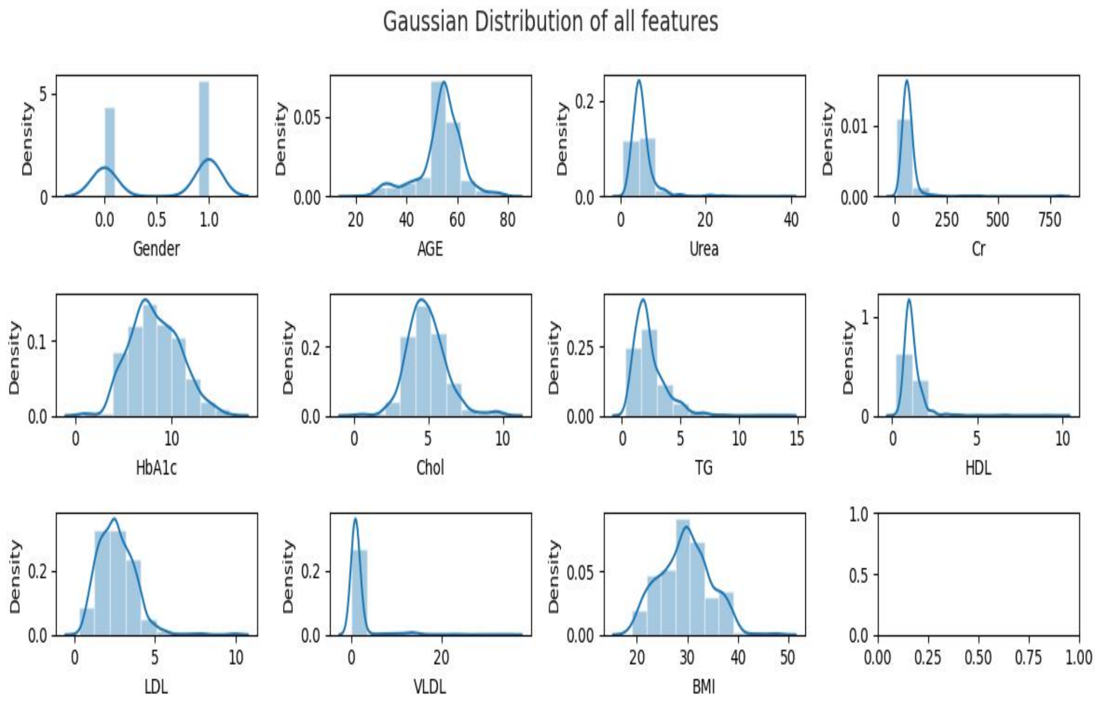
The Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital, Medical City Hospital labs, and the Iraqi society were the sources of the dataset used to train the machine learning model. To develop the dataset exclusively for diabetes-related cases, the data were taken from patient files that contained medical information and laboratory analyses. It consists of 1000 patient health information whereby 565 records were for male, and 435 records were for female. From the data analysis staged it was observed that the dataset entails 844 diabetics cases, 53 pre-diabetics and 103 non-diabetic, with eleven important features. The Gender is known to be the only categorical feature in the dataset, where it has to do with either male or female. The two main categories of the 11 health-related variables are demographic information and laboratory test results. However, assessments of a number of markers, including total cholesterol, triglycerides, glucose (AC), hemoglobin A1c, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol, were included in the laboratory data. These findings showed that people with diabetes mellitus (DM) have impaired metabolic activity. (Guasch-Ferré et al., 2016)

Table 1: Functional description of features.

Features	Description	Measurement	Mean	Std
Gender	Gender (male or female)	Discrete	0.56	0.49
Age	Age of the patient (years)	Continuous	53.60	8.73
Urea	Urea level of the patient	Continuous	5.12	2.94
Cr	Creatinine ratio	Continuous	68.93	60.13
HbA1c	Hemoglobin A1C test to measure blood sugar(glucose) level	Continuous	8.24	2.53

Chol	Cholesterol	Continuous	4.86	1.30
TG	Triglycerides	Continuous	2.35	1.40
HDL	high-density lipoprotein	Continuous	1.21	0.66
LDL	Low-density lipoprotein	Continuous	2.61	1.11
VLDL	Very-low-density lipoprotein	Continuous	1.82	3.62
BMI	Body mass index	Continuous	29.56	4.95

Figure 2: Gaussian distribution for all diabetics feature in the dataset



3.1.1 Data Processing and Cleaning

Finding and fixing faults, inconsistencies, and errors within the dataset is data cleaning, sometimes referred to as data cleansing or data pretreatment. In order to make dirty and unreliable data appropriate for analysis or modeling, this method seeks to turn information into a clean and correct format. (Parajuli et al., 2022). Data cleaning helps to enhance data quality and provides accurate and useful findings, making it an important step in the data processing process. The process of cleaning data is iterative, and it may take several rounds to reach the required degree of cleanliness and quality(Gopal et al., 2021). To preserve openness and repeatability, it is crucial to record the actions followed during data cleansing. The accuracy and efficacy of following analysis or modeling operations are improved by thoroughly cleaning and preprocessing the data. In this work, the data processing implemented are explained as follows:

Handing missing values: Determine the dataset's missing values and choose the best course of action to manage them. Rows or columns with missing values can be removed, or missing values can be imputed using the mean, median, or other statistical measures, or using more sophisticated imputation methods like regression or multiple imputation. Locating the missing values and cleaning them was done using the pandas isnull() function. This function scans through the dataset and give a summary of the number of null values present in the dataset. For easy processing, the rows with these missing values were deleted.

Categorical features encoding: In order for machine learning methods to be used successfully for analysis or modeling, categorical variables or features are encoded into numerical representations. The intervals themselves are categorical, but the data utilized represents integers. Comparable discrete data is still numerical. The total of two dice throws is an illustration of discrete data. There is a known and finite set of outcomes, but these outcomes are numerically represented (Pan et al., 2022). In our dataset the variable that indicate qualitative traits, like the gender categories are known as categorical features.

The gender it shows that the gender of the only discrete value in the dataset. Because the categorical data gender is either male or female and was converted to binary values 1 and 0. The processing also involved detecting and handling missing values.

3.1.2 Feature Correlation

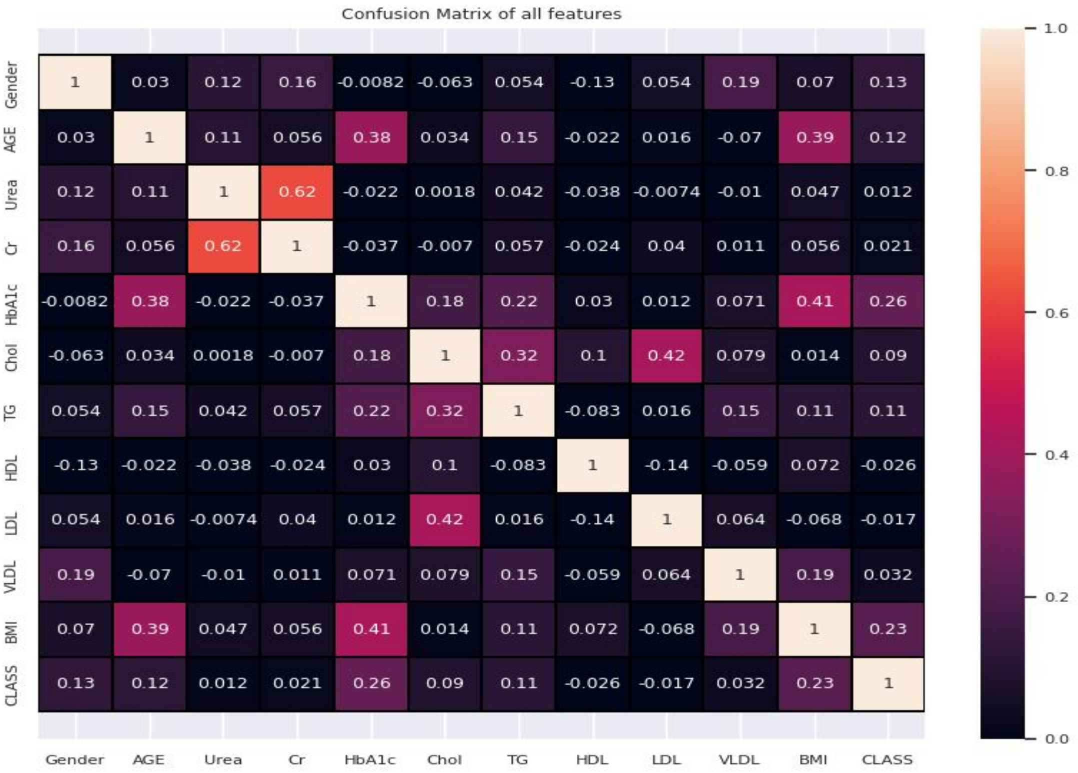
The statistical link or relationship between two or more features (variables) in a dataset is referred to as feature correlation. It assesses how well changes in one attribute translate into changes in another. For feature selection, dimensionality reduction, and comprehending the underlying patterns in the data, feature correlation analysis can offer insights into the dependencies and interactions between variables (Buyrukoğlu & Akbaş, 2022). The correlation coefficients for several variables are simply listed in a table as a correlation matrix. The matrix demonstrates the relationships between every conceivable pair of values in a table. It is an effective tool for detecting and displaying data trends as well as summarizing the huge data collection.

This stage involves the cleaning and processing of the data, as well as feature selection. The data was observed to have some missing values and some irrelevant data. During the analysis of the data, it was discovered that some features were not relevant and had no impact on the problem or had no direct correlation with the output variable. Such as the patient number and the Id.

In other to get the feature importance and to understand which variable or diabetics test has more influence on the target variable, a correlation test is done using a heatmap to plot the confusion matrix. Confusion matrices show when a model regularly misclassifies two classes, which makes it easy to assess how trustworthy a model's output is likely to be. When a classification model is effective, it gives understanding and helps to determine which data it might not be able to classify correctly. This data is then known to have a very low correlation and sometime times having a negative correlation with the target class.

From the heatmap below, the variable with the highest correlation with the target class is the HbA1c, which represents the Hemoglobin A1C test and measures blood sugar(glucose) level. In diabetics, the blood sugar content is a very important attribute and a strong measure to determine if the patient is diabetics type 2 positive. The HbA1c has a correlation to the target index of 0.26 and to the body mass index (BMI) 0.41.

Figure 3:heatmap representing the correlation matrix



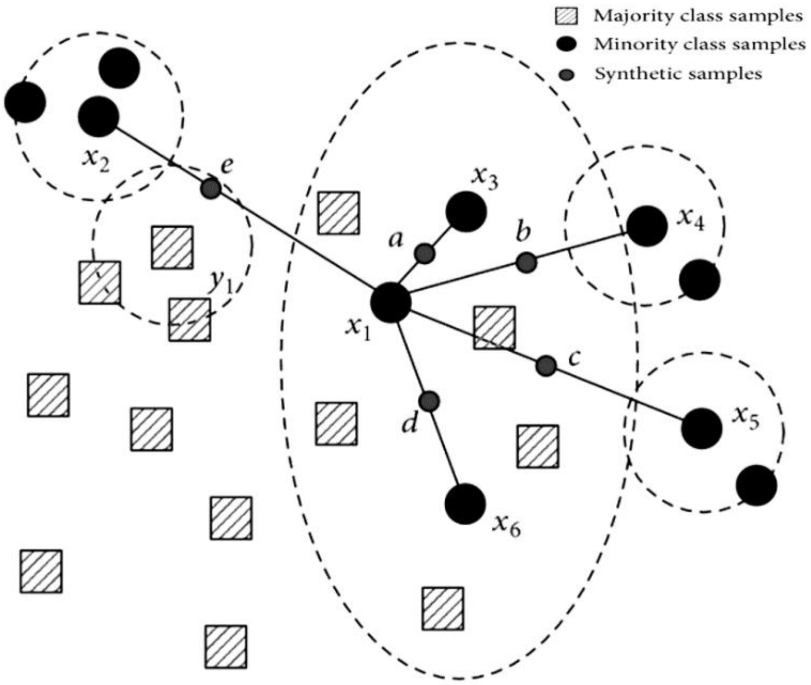
These features are said to be the most important features in the dataset, because they have a lot impact on the output.

3.1.3 Dealing with the Imbalanced dataset

The distribution of classes or categories in a dataset is referred to as an "imbalanced dataset" if it is not equal or balanced(Yadav et al., 2021). In such datasets, one class predominates over the others considerably or has a greater number of instances. In many different fields, including fraud detection, disease diagnosis, text categorization, and anomaly detection, unbalanced datasets are typical. From the diabetes dataset present in this work, the dataset is observed to be imbalanced, having over 80% of the dataset. This situation can lead to the over fitting of the model and coursing the whole data not to fit well during the modeling. When it comes to the learning task, the minority class is typically more engaging. The classification of satellite images, risk management, and medical diagnosis are just a few examples of circumstances where there is an imbalance between classes. Using the classifiers created by traditional machine learning algorithms without modifying the output threshold may be a crucial error while researching difficulties with imbalanced data (Hu & Li, 2013). Due to their propensity to be skewed in favor of the majority class, imbalanced datasets provide difficulties for machine learning algorithms since they do a poor job of recognizing and forecasting the minority class. Due to this imbalance, models may perform well for the majority class but badly for the minority class, which is frequently the class of interest (Hassan & Amiri, 2019b). There are different ways to handle this problem of imbalanced dataset but this work engraved the Synthetic Minority Over-sampling Technique (SMOTE).

SMOTE: is a popular approach for dealing with class imbalance in machine learning. By interpolating between instances of the minority class that already exist, For the minority class, it is an oversampling strategy that produces fake samples (Azad et al., 2022) . SMOTE makes use of generated samples to enhance the dataset's representation of the minority diabetic class. This enables the machine learning algorithm to gain knowledge from a more diverse dataset and perhaps enhance its capability of foretelling the minority class (Dimililer et al., 2022).

Figure 4: The smote oversampling algorithm



A k-nearest neighbor method is used by SMOTE to generate artificial data points. The SMOTE algorithm is displayed as follows:

1. *Begin by identifying the vector representing the minority class in the dataset.*
2. *Choose the k nearest data points to the minority class for further evaluation.*
3. *Create a line and introduce a synthetic point between each minority data point and any of its k nearest neighbors.*
4. *Repeat steps 3 and 4 until a balanced distribution is achieved between the minority data points and their k neighbors.*

The minority class samples x1 and x6 from figure 4 were chosen using the k closest neighbors (KNN) technique and then interpolated to create a new sample d.

3.1.4 Feature Engineering and data binning

The feature engineering pipeline is the process of transforming raw data into meaningful features that may be used in machine learning methods, such as predictive models. Predictor variables are created and chosen to be utilized in the predictive model throughout this step. The outcome variable and the chosen predictor variables are the two components of these prediction models (Connie et al., 2022). This work presents a situation for bring out new classes to the dataset class by understanding the importance of the various features and how they can affect the target class. The mean blood glucose (sugar) concentrations during the previous two to three months, expressed as HbA1c and expressed in mmol/mol, are a highly significant factor to consider.

Data Binning for the HbA1c feature: The binning method has been used to smooth data. As demonstrated in Table 2, it has proved advantageous to categorize age into five groups. It is shown that the blood glucose level has a tangible effect on the diabetic results. And, with the help of these result gotten from the bins, a new class is created. Patients without diabetes have a different blood glucose concentration than patients with the disease. So, this situation imposes a problem when the blood glucose level is lower than the normal state or not the state for diabetes. In the section of the work, we introduce 2 new classes called the **Hypoglycemia (HG).** When the blood glucose level is less than 80 mg/mol, these phases appear (Wang et al., 2020) . Hemoglobin is a protein found in red blood cells. It is in charge of distributing oxygen throughout your body and giving blood its red hue.

Table 2: Binning for glycated hemoglobin (HbA1c) to Blood glucose

Glycated hemoglobin test (%)	**Blood Glucose (mg/mol)**	**Blood Glucose bin for the new class**
Less than 3.9	Less than 60	Hypoglycemia stage 2(HG 2)
3.9 – 4.5	61 - 70	Hypoglycemia stage 1(HG 1)
4.6 – 5.6	71 - 140	Normal
5.7 – 6.5	141 - 180	Prediabetes
6.6 above	181 above	Diabetes

From the table 2, the dataset is relabeled based on the value gotten from the glycated hemoglobin test. These bins are said to be the new classes for our diabetes prediction. So, the dataset was relabeled to 5 classes for the prediction. In this case this work then to also determine the opposite case for diabetes which was said to be the case when the blood sugar level is below 71 mg/mol.

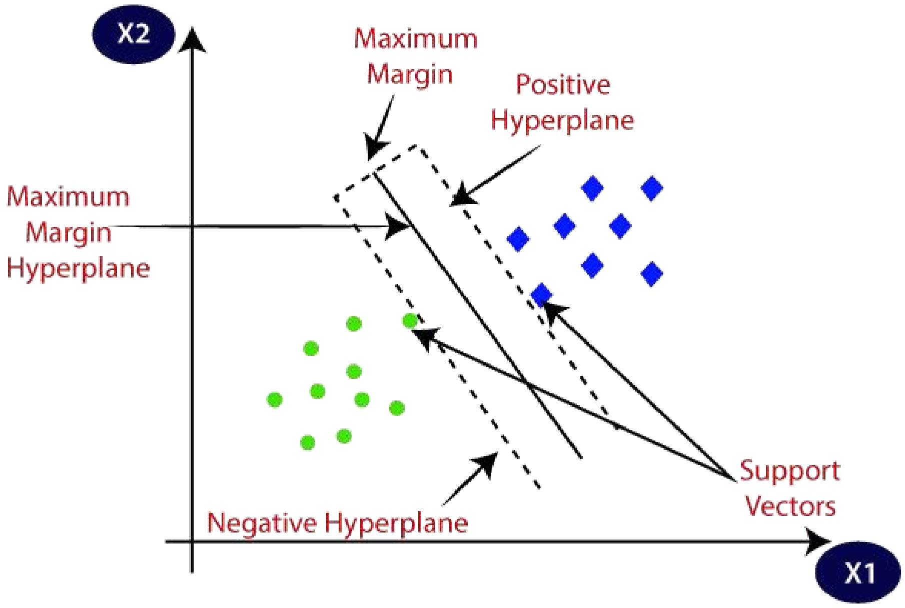
3.2 Modeling

Details on the application of the diabetes categorization model are provided in this section. After the data processing, the major step is to select a classification algorithm suitable for the diabetic dataset. And giving the type of classification, which is a multi-class and an imbalanced classification. The data is fitted to the classification method after it has been chosen to produce a model. The model is then evaluated and checked with different classification evaluation techniques. The evaluation tells if the model is suitable for the prediction or not, it gives an over on how the model performs either on the training set or on the testing set.

3.2.1 Support vector machine

A very well-liked supervised learning approach called Support Vector Machine (SVM) is used to resolve Classification and Regression issues. However, its main use in the field of machine learning is to handle Classification jobs (Irkham et al., 2022). Finding the best line or decision boundary in an n-dimensional space that effectively divides various classes is the basic objective of the SVM method. The SVM permits quick categorization of fresh data points in the future by building this hyperplane. The name of this optimal decision boundary is the hyperplane (A. Roy & Chakraborty, 2023). Support vectors, which are extreme data points that are critical for creating the hyperplane in SVM, are identified and chosen by the algorithm. The SVM methodology is built on these support vectors. Figure 5 shows the choice boundary, or hyperplane, which essentially divides the world into two groups.

Figure 5: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes



The two most well-known SVM kinds are linear and non-linear. The linear SVM can handle data that can be separated into two classes by a single straight line, also known as linearly separable data. The best linear decision boundary to divide the data into their appropriate classes is discovered in this instance using the linear SVM technique (Kurani et al., 2023). The classification technique used is a Linear SVM classifier for datasets that may be divided by a straight line. However, a non-linear SVM classifier is employed when working with datasets that cannot be divided into segments by a straight line. By identifying the most suitable non-linear decision limits to accurately classify the data, the non-linear SVM classifier can handle larger and more complicated datasets (Zhou et al., 2023).

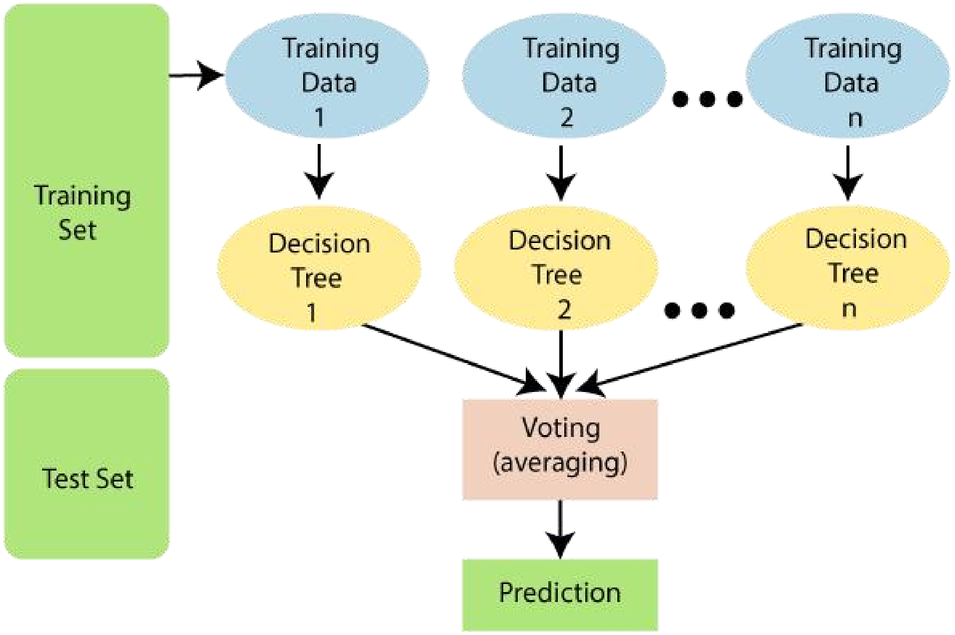
Multiple lines or decision boundaries may be used to possibly divide the classes of the data in an n-dimensional space. Finding the right decision boundary to classify the data points, however, is crucial. In the context of SVM, this ideal boundary is referred to as the hyperplane. In order to accurately classify data points, the hyperplane serves as the optimum decision boundary that optimizes the margin between the classes (Junaid et al., 2022).

The characteristics included in the dataset dictate the SVM hyperplane's dimensions. For instance, the hyperplane will be represented as a straight line if the dataset only has two characteristics. On the other hand, the hyperplane will have two dimensions if the dataset has three characteristics. Building a hyperplane with the highest margin, or the biggest separation between the data points of distinct classes, is the main goal of SVM. This guarantees the most effective and precise categorization of the data points, optimizing the SVM algorithm's performance (Umar Ibrahim et al., 2023) . Support vectors are certain data points or vectors that are in close proximity to the hyperplane and have a big influence on where it is. Because they actively assist and direct the positioning of the hyperplane, these vectors are known as "support vectors". Support vectors are important in the decision-making process of the SVM algorithm since they are the main determinants of the placement of the hyperplane. For best performance, SVMs need to have several hyperparameters tweaked. The regularization parameter (C), the kernel-specific parameters (such as the degree for polynomial kernels or gamma for radial basis function kernels), and the type of the kernel (e.g., linear, polynomial, or radial basis function) are all important hyperparameters. To get the best classification results, the ideal hyperparameter combination must be found. SVMs can manage datasets with an uneven distribution of cases across different classes (Yang et al., 2023). SVMs can be used successfully to solve imbalanced classification problems by modifying the class weights or incorporating methods like oversampling or under sampling (Zhang, 2012) . SVM classifiers, in general, are flexible and powerful tools for both linear and nonlinear classification applications. They are frequently used in a variety of fields, such as image recognition, text classification, and medical diagnosis, due to their capacity for handling complex decision boundaries, interpretability through support vectors, and robustness to outliers. To achieve the best results using SVMs, proper hyperparameter selection and treatment of imbalanced datasets are essential (Stephan et al., 2022).

3.2.2 Random Forest Classifier

Talking about the various classification algorithms, the random forest classifier is simply known to be an ensemble of several Decision Trees (DTs) also known as the rain of forest algorithm. A class of forecasting is represented by each tree in the random forest. The one that receives the most votes among them is regarded as the model's forecast (Afsaneh et al., 2022). Considering the foregoing, optimizing the method for multi-class categorization. A non-parametric approach called the Decision Tree can be used to solve issues involving classification and regression. Two ideas, decision nodes and leaves, can be used to define the tree. The leaves are identified to be the various decision or practical events to be made by the algorithms and can be defined as the place where data split (Woldaregay et al., 2019).

Figure 6: The working of the Random Forest algorithm



There are many trees in a forest, and the more trees there are, the more active the forest is. Different decision trees are created by Random Forest using randomly chosen data, and the trees collective votes are then used to determine the test object's class. The final output is decided using the majority vote process. The outcome that the majority of the decision trees in this scenario have chosen is the rain forest system's ultimate output. (Daghistani & Alshammari, 2020) . The design in figure 6 shows a simple random forest classifier. This classification algorithm is the selected model for this work based on its capability to handle multi class dataset. The diabetics dataset which consists of the pre-diabetic, diabetic and the non-diabetics' class is passed to the ensembled decision trees in the forest, and their output is then counter

3.2.3 Random Forest hyperparameter tuning.

A number of hyperparameters in the random forest can be adjusted to enhance performance. The number of decision trees in the ensemble, the maximum depth of each tree, and the number of characteristics taken into account at each split are all significant parameters employed in this study.

Firstly, in training the model, the random state parameter was set to zero and the tree depth was set to 3. The random state helps the model give the same output for different execution. The output performance changes when these parameters are changed.

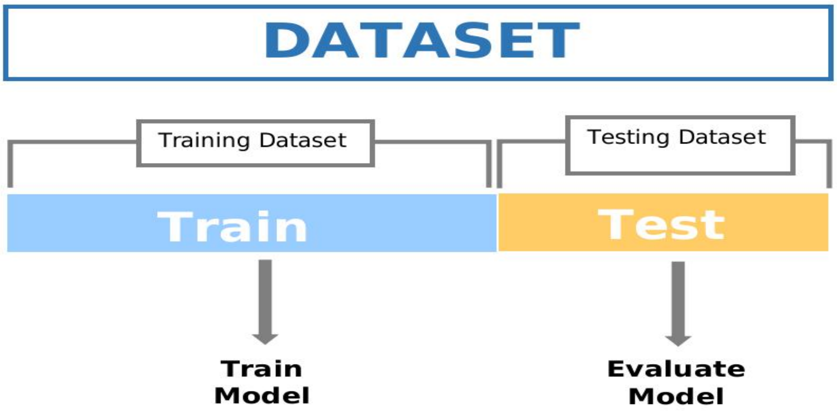
- *n_estimators: It represents the number of trees present in the forest.*
- *max_features: This parameter specifies the maximum number of features considered for splitting a node during the tree construction.*
- *max_depth: It denotes the maximum number of levels that each decision tree can have.*
- *min_samples_split: This parameter sets the minimum number of data points required in a node before the node can be split further.*
- *min_samples_leaf: It determines the minimum number of data points allowed in a leaf node.*

These parameters play crucial roles in shaping the random forest model and influence its performance in handling different types of data and decision-making processes.

3.2.4 Train Test Split

Before training the model or fitting the data to the RF classifier, then divide the dataset into a train set and test set. The training data had a percentage of 70 and testing data was 30 percent. It is observed that we just have a little amount of data and the dataset wasn't balanced as shown in figure 7. The class with diabetics has more data distribution to compare the prediabetics and the non-diabetics, this method results in a large bias because the model misses some information about the data that wasn't used for training coursing the model to be overfitted. The scikit-learn python library was used to split the data using the **train test split** method.

Figure 7: The train test split for the dataset

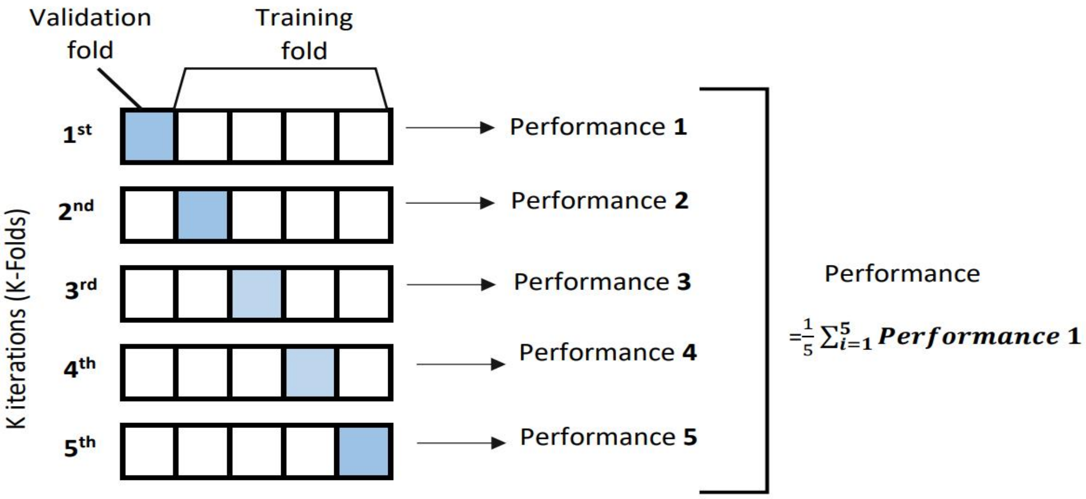


3.2.5 K – fold Cross Validation

K-fold Cross-validation (KCV) is one of the most well-liked techniques for classifier model selection and error estimates (Arlot & Celisse, 2010). The practice of cross-validation (CV) is a well-liked algorithm selection method. To estimate the risk associated with each algorithm, CV splits data once or multiple times: Each algorithm is trained using a portion of the data (the training sample), and the remaining portion (the validation sample) is used to calculate the algorithm's risk. The algorithm with the lowest calculated risk is then chosen by CV (Munna et al., 2020) . Due to the independence of the training sample from the validation sample, CV prevents overfitting in comparison to the resubstituting error.

Since the dataset is not properly distributed, and this leads to an imbalanced dataset, using the normal train test split method is not suitable for the process and cannot attain proper accuracy. The stratified K-fold cross-validation object is a kind of K-Fold cross-validation that produces stratified folds. By tracking the sample percentages for each class, this method makes sure that the folds preserve the same class distribution as the original dataset (A. Ramezan et al., 2019) . Using the train/test indices supplied by this approach, the data is then split into train and test sets. This approach, which makes use of stratified folds, makes it possible to evaluate machine learning models more robustly, especially when working with datasets that are unbalanced and have different class proportions. When using the stratified k-fold cross validation, we want to maintain the desired class ratio and the number of trainings set a limited or the amount of training data are not that much for the estimator. Figure 7 depicts the visual representation of the data splitting (5-fold cross-validation) employed in this study.

Figure 8: Split fold for the cross-fold cross validation



This method splits our dataset in 5 different stratified cross-fold, allowing the classifier to be trained with the knowledge of all the data in each split. This method was used to avoid the overfitting of the model and avoid the model been trained with just the data with the highest count because of how imbalanced the dataset is. From the figure above, our dataset was divided based on ratio of the data classes for each fold. After the training for each fold, the accuracy is averaged as the general accuracy for the model.

3.3 MODEL EVALUATION

In machine learning, the type of evaluation done is based on if it is a classification or regression problem. In this section the model is evaluated after training based on accuracy. In our case we have a five-class classification problem, and confusion metrics explain more on the performance of the model. The Receiver Operating Characteristics (ROC) with Area Under the ROC Curve (AUC) is also offered to measure the accuracy of predictions rather than their absolute values (Severeyn et al., 2021). Evaluation of the models is one of the important process and task involved in the life cycle of a machine learning project. This section tells the better or the best model to be used. This means the model that performed well for the training data (Mishra et al., 2020).

ACCURACY is a typical parameter used to assess the performance of a machine learning model, notably in classification tasks (Park et al., 2023) . It computes the fraction of correctly identified examples in relation to the total number of instances in the dataset. While accuracy is a simple and obvious indicator, it may not always offer a whole view of a model's performance, especially in circumstances with unbalanced classes or where the cost of false positives and false negatives fluctuates (Jakka & Rani, 2019).

Here's how accuracy is determined and some things to keep in mind when utilizing it.

Calculation for accuracy:

$$Accuracy = \frac{numbers\ of\ correct\ predictions}{total\ number\ of\ predictions} \times 100$$

Some considerations in utilizing accuracy in model evaluation

3.3.1 Imbalanced classes

Imbalanced classes occur when the quantity of cases in distinct classes is notably uneven in classification tasks. One class, known as the majority class, has many more instances than one or more minority classes, which have far fewer instances. This class imbalance can make training and evaluating machine learning models difficult (Elseddawy et al., 2022).

3.3.2 Confusion matrix

Confusion matrix is a key machine learning technique for assessing the effectiveness of a classification model. It gives a detailed breakdown of the model's predictions and actual results, providing for a more in-depth knowledge of how well the model performs across different classes(Krstinic et al., 2023). The confusion matrix is most useful in binary classification (two classes) and multiclass classification (more than two classes) settings. It is made up of four major parts. True Positives (TP): Instances that the model accurately predicts as positive.

- True Negatives (TN): Instances that the model accurately predicts as negative.
- False Positives (FP): Instances that the model mistakenly predicts as positive when they are in fact negative (Type I error).
- False Negatives (FN): Instances that the model mistakenly predicts as negative when they are in fact positive (Type II error).

The confusion matrix and its related metrics aid in providing a more detailed insight of a model's performance, namely its strengths and shortcomings in predicting distinct classes. It is an essential tool for making educated judgments regarding model changes, feature engineering, and threshold tuning in order to achieve the ideal balance of precision and recall (Markoulidakis et al., 2021).

3.3.3 Receiver Operating Characteristic (ROC) curve

The ROC is a graphical tool for evaluating the performance of binary classification algorithms at various thresholds. It shows how the true positive rate (sensitivity) and false positive rate (fallout) fluctuate when the decision threshold for distinguishing positive and negative cases varies (de Hond et al., 2022).

The ROC curve shows:

The x-axis shows the false positive rate (FPR), which is the percentage of true negatives that are wrongly categorized as positives. The true positive rate (TPR), also known as sensitivity or recall, is the proportion of genuine positives that are accurately categorized as positives on the y-axis (Mandrekar, 2010).

The ROC curve is especially useful in situations when the balance of sensitivity and specificity is critical. A good model will have a ROC curve that is closer to the top-left corner, suggesting better sensitivity and lower fallout over a wide range of thresholds.

The AUC-ROC (Area Under the ROC Curve) is a single statistic that describes the model's overall performance. AUC of 0.5 suggests random guessing, whereas AUC of 1 shows a flawless model. The higher the AUC-ROC score, the better the model's capacity to differentiate between positive and negative cases (Cook, 2007). The ROC curve is a useful model comparison tool since it allows you to examine and evaluate the performance of many models on a single plot. It assists you in determining the best threshold for your unique problem and tolerance for false positives and false negatives (Gneiting & Vogel, 2022).

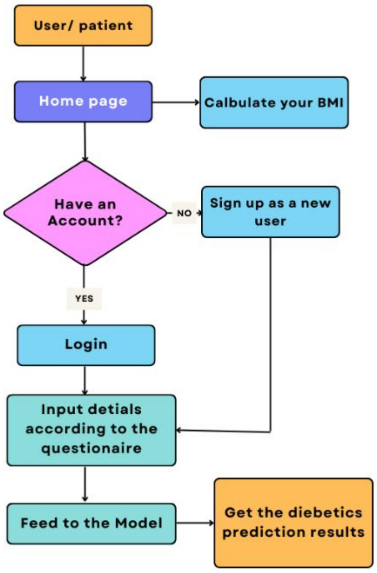
3.4 WEB APPLICATION INTERFACE

The web application was built to practically use the model for real life scenarios. It was designed to make the model usable. It was implemented using the flask python framework which serves the model as an API for the inference(Pankaj et al., 2021). The diabetics data is sent as a request from the client which is the patient to the server using a **post method**, and the request is passed through the flask RESTAPI to the model in the server. The prediction from the model is sent back through the API as a response(Hussain & Al-Turjman, 2022).

The action to carry out on the resource identified by the uniform resource identifier is known as the method.

The flow chart in figure 10 explains how the system works and the user interacts with the interface. The user login into the system or signup if he or she does not have an account. The login ensures that the patient's data is saved to the patient's account for future use and for other recommendations for the patient. The test data entered in the questionnaire page is saved to the database and passed to the model saved in the backend of the web app. The response from the model is a Json data containing the prediction and this response is then sent to the frontend of the web app.

Figure 9: flowchart showing the process and operation of the web application.



3.4.1 Architecture and technology stack

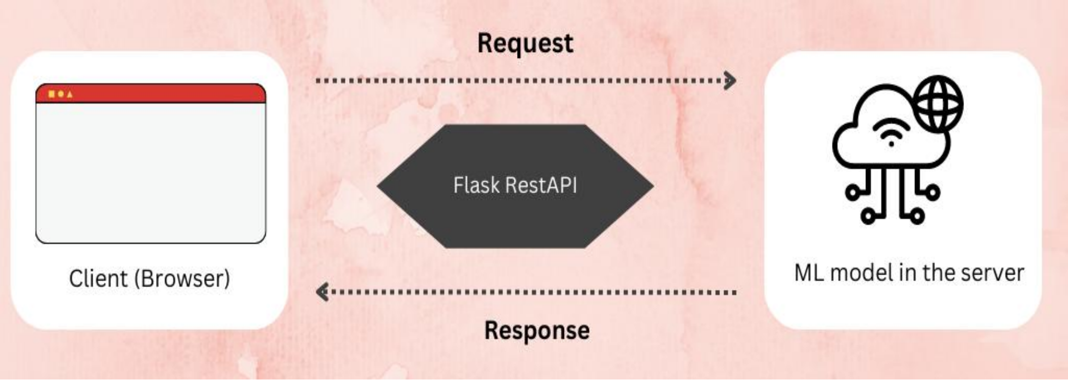
Flask framework for the back-end, HTML and CSS for the front-end, and responsive CSS for the best user experience across devices are all used in the web app's design, which is sturdy and effective. The API was developed using Python, providing smooth data exchange between the front-end and back-end components (Huckvale & Moseley, 2023) . The back-end architecture is based on Flask, a compact and adaptable web framework that makes routing and request processing simple.

Python's adaptability makes it possible for Flask to be seamlessly integrated with it, facilitating the construction of APIs and data processing. The modular design of Flask makes it simple to scale, allowing for future development of the app's functionalities. On the front end, an easy-to-use and aesthetically pleasing user interface was made using HTML and CSS. In order to create a seamless user experience across PCs, tablets, and mobile devices, responsive CSS ensures flexibility to multiple screen sizes. User engagement is increased by the smart front-end design, which emphasizes user accessibility, navigation, and readability (Umar Ibrahim et al., 2022). The use of Python-based API improves the functionality of the web app by enabling dynamic data retrieval and presentation. Real-time communication with the back end is made possible through the API, giving users access to the most recent data and enhancing their user experience. Overall, the web app's design is a seamless synthesis of Python's API building skills, HTML/CSS's front-end adaptability, and Flask's back-end functionality. The web app offers a remarkable user experience because to its user-centric design and responsive layout, giving users easy access to useful features and data.

3.4.2 Random Forest model to RESTApi

A Random Forest model that has been smoothly included as an API has been added to the web interface, boosting its usability and forecasting skills. The Random Forest model can be actively interacted with by users using the model API, which produces real-time predictions based on user inputs (Ahmad et al., 2021). Users may quickly enter data into the Random Forest model and receive immediate predictions without leaving the web app by displaying the model as an API.

Figure 9: working of the web app with the rest Api



The user experience is improved by this feature, which provides dynamic and customized insights. Data security and effective communication between the front-end and back-end components are given first priority in the painstakingly developed integration of the Random Forest model as an API. The flexibility of the Python programming language provides seamless integration with the Flask framework, enabling the model API to function without interruption within the framework of the web app.

Appendix B, in the appendix section shows the code for implemented to stream the model to an API for the prediction. The result from the prediction is also send to the frond-end page through another form of API. These helps the user interact well with the model. The web app's functionality is improved because to this integration, which also creates opportunities for potential future improvements. The model API's modular architecture makes it simple to update existing models and create new ones, providing chances for growth and scalability.

3.4.2 Deployment and Hosting

PythonAnywhere's cloud platform as a service (PaaS) was used to deploy and host the web app. Platform as a Service (PaaS) is a cloud computing architecture that provides developers with a platform and environment to build, deploy, and manage applications without the burden of infrastructure administration (Sania Febriani & Fitri Purwaningtias, 2022) . The cloud service provider provides a complete development and runtime environment in PaaS, allowing developers to concentrate entirely on building and operating their applications. Without the need for intensive server maintenance, PythonAnywhere offers a convenient and user-friendly environment for hosting and running web applications.

The following steps were part of the deployment process:

- **Creating a PythonAnywhere Account:** Initially, an account was created on the PythonAnywhere platform to access their services and tools.
- **Uploading the Web App:** The web app's files, including the front-end HTML/CSS, back-end Flask code, and machine learning model, were uploaded to the PythonAnywhere server.
- **Setting up Virtual Environment:** A virtual environment was created to isolate the web app's dependencies and ensure compatibility.
- **Installing Required Packages:** All necessary Python packages and dependencies were installed within the virtual environment to run the web app smoothly.
- **Configuring WSGI File:** The Web Server Gateway Interface (WSGI) file was configured to specify the Flask application and enable the web app's interaction with the PythonAnywhere server.
- **Running the Web App:** The web app was launched on the PythonAnywhere server, making it accessible through a unique URL provided by the platform.

The web app was made internet accessible by being hosted on PythonAnywhere, enabling users to access its features and forecasts using any web browser. The platform's managed hosting environment reduced the hassle of server management, allowing the emphasis to stay on the creation and functionality of the web app. The web application was easily deployable and made available to its target audience thanks to the use of PythonAnywhere as a PaaS solution.

CHAPTER 4

Results And Discussion

In this section, we looked at the outcomes from our model on a dataset for categorizing diabetes. There were three classes in the dataset at first: "normal," "prediabetes," and "diabetes." The HbA1c feature, however, allowed us to define two new groups, "HG 1" and "HG 2," which stand for stage 1 and stage 2 hypoglycemia, respectively. A more thorough grasp of the disease spectrum was made possible by this enlargement of the classification system. We measured the effectiveness of our model using a number of parameters, including accuracy and the Area under the curve AUC. The model's accuracy was commendable, demonstrating its aptitude for appropriately classifying events into the five categories. To comprehend the distribution of cases among the various classes, we also looked at the class frequencies.

We further tested the stability and generalizability of the model using cross-fold validation. We learned about the dataset's consistency and variation across various subsets of data by folding the dataset into many folds and training/testing the model on various combinations.

The outcomes also demonstrated the significance of feature engineering, particularly the inclusion of the HbA1c characteristic, which offered useful details for the detection and classification of hypoglycemic stages. Healthcare providers may now adjust treatment strategies and treatments to each patient's unique hypoglycemia stage because to this improved classification system.

Overall, our investigation produced encouraging findings, demonstrating our model's ability to correctly categorize occurrences throughout the broadened classification scheme. A thorough analysis and better comprehension of the dataset were made possible by the use of cross-fold validation and the inclusion of the HbA1c attribute. These results set the groundwork for further investigation and potential improvements in the classification and treatment of diabetes.

4.1 Data Transformation and Class Generation

Figure 10: Diabetic class from the original dataset

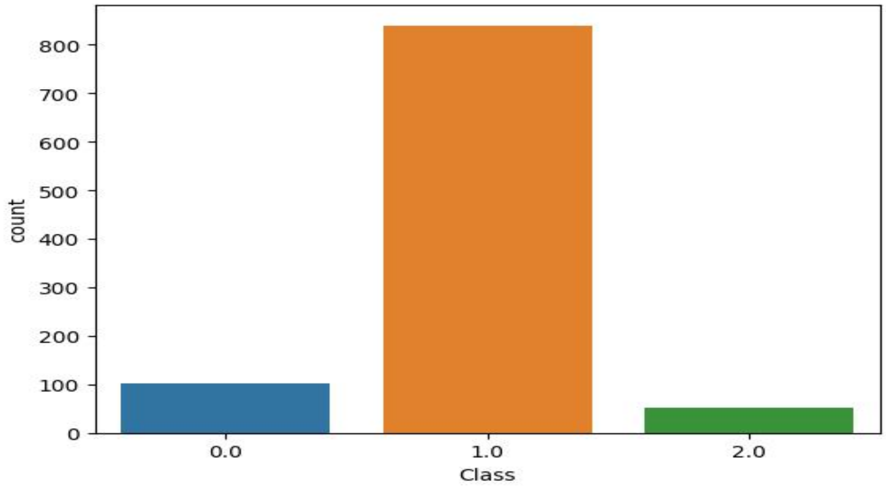
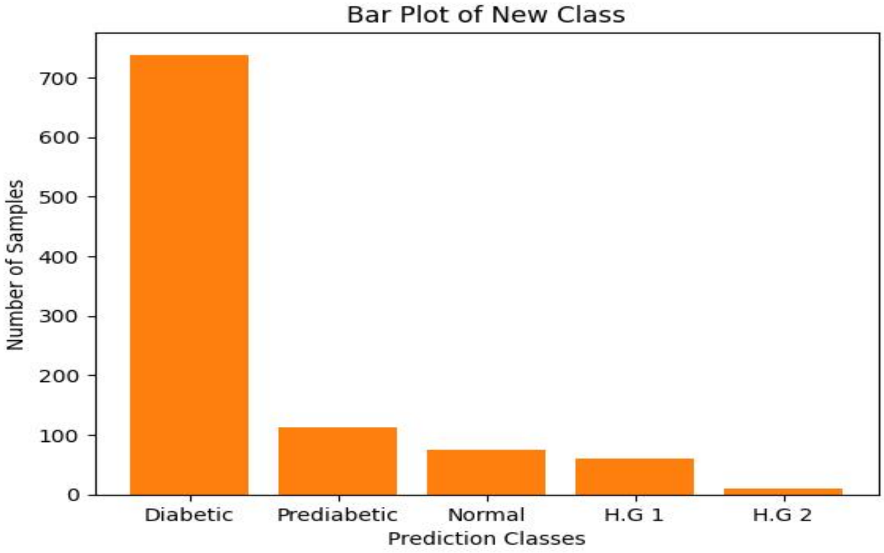


Figure 11: New diabetes class from the HbAc1 test binning



There were three classes in the original diabetes dataset: "normal," "prediabetes," and "diabetes." However, a new class was produced using the HbA1c characteristic, leading to an extended categorization with five classes. The two classes, "diabetes." Figure 10 shows the plot for the count and the number of classes in the initial dataset. However, a new class was produced using the HbA1c characteristic, leading to an extended categorization with five classes. The two classes, "HG 1" and "HG 2," stand for the various stages of hypoglycemia, namely "hypoglycemia stage 1" and "hypoglycemia stage 2."

The dataset was enhanced as shown in the next figure 11 by the addition of the HbA1c characteristic, a useful marker of long-term blood glucose control, which provides more details about the stages of hypoglycemia (Bartolome & Prioleau, 2022). This improvement enables more precise categorization and perhaps improved identification of people at various risk levels. Adding stages of hypoglycemia to the classification allows for a more thorough grasp of the disease spectrum. It enables the identification of those who are more or less likely to have hypoglycemia or who are already going through milder or more severe phases of the ailment. This more precise granularity can help in customizing treatment strategies, treatments, and monitoring tactics to the particular hypoglycemic stage.

The classification model can now better reflect the variety within the diabetes community with the addition of HG 1 and HG 2 as new classifications. It enables medical personnel to recognize patients who need closer observation or treatment to properly avoid or manage hypoglycemia.

As a result, the diabetes dataset is enriched by the addition of the HG 1 and HG 2 classes based on the HbA1c attribute, resulting in a more thorough classification scheme. This growth presents opportunities for enhanced risk classification, individualized treatment plans, and focused interventions for people experiencing various degrees of hypoglycemia.

4.2 Model selection

We investigated the application of machine learning methods for predictive modeling in this work. Support Vector Machines (SVMs) were one of the techniques that were examined for classification jobs. The SVM model did not, however, attain adequate accuracy in comparison to the Random Forest classifier, which finally emerged as the chosen base model for the study, despite numerous attempts at hyperparameter adjustment (Zhu et al., 2023) . SVMs are strong algorithms that search for the best hyperplane to divide several classes in a dataset. They can handle high-dimensional data and perform well with complex and nonlinear relationships. However, choosing the right hyperparameters, such as the kernel type, regularization parameter (C), and gamma value, is crucial for determining how well they function.

The SVM model could not display the necessary accuracy in predicting diabetes classes when compared to the Random Forest model, despite substantial work being put into tuning these hyperparameters and experimenting with other configurations. This outcome may have been influenced by a number of factors:

Dataset characteristics: SVMs work best when the dataset is linearly separable or when there is a distinct difference between classes. The SVM may have trouble determining the best decision boundary if the dataset is highly unbalanced or contains overlapping regions(Furey et al., 2000).

Complexity of the Data: SVMs demand careful kernel function selection and adjustment. The model's performance may deteriorate if the data displays intricate patterns or nonlinear correlations that the chosen kernel is unable to effectively capture.

Sensitivity to Hyperparameters: SVMs are sensitive to hyperparameter settings, making it difficult to choose the best ones. Sometimes the hyperparameters may not have been properly adjusted, producing less-than-ideal outcomes (Deebak & Al-Turjman, 2023).

The dataset's intricate linkages, however, were better captured by the Random Forest classifier, which synthesizes several decision trees. Random Forest models are more robust when dealing with skewed data and less susceptible to hyperparameters (Harimoorthy & Thangavelu, 2021) . Additionally, they offer feature importance rankings that can be used to better comprehend the fundamental principles behind the classification. Based on a thorough analysis of several performance criteria, including accuracy, confusion matrix, Area under the ROC curve, the Random Forest model was chosen above the SVM. For the specific diabetes dataset utilized in the study, the Random Forest model consistently beat the SVM in terms of accuracy and displayed better overall classification performance. It is important to acknowledge the iterative nature of the research process even though the SVM model did not achieve the target accuracy criterion. Machine learning studies sometimes include trial and error, investigating various configurations and techniques to find the best model for a given problem (Shrivastava et al., 2022). The Random Forest model was chosen as the basic model for the thesis work since it performed better in this instance.

4.3 Model evaluation result

The classification accuracy is the mostly used evaluation metric used to evaluate the performance of the random forest classifier. Several criteria were used to assess the model's performance, with an emphasis on accuracy in particular. The model successfully classified cases accurately, as evidenced by its accuracy as shown in the table 3. This indicates that the model is successfully forecasting the target variable using the provided features. And these accuracies vary based on the training parameters used for the random forest classifier.

By experimenting with various Random Forest Classifier hyperparameters, the model's performance was assessed. To determine their effect on the model's performance, various hyperparameters including the number of estimators, maximum depth, and minimum samples split were changed. It was discovered through this study that adding more estimators often enhanced performance up until a certain point, at which time the gains plateaued or even shrank as a result of overfitting. Similarly, changing the trees' maximum depth had an effect on how well the model worked. Underfitting might result from setting the maximum depth too low, while overfitting could result from setting it too high. The model's capacity to generalize data patterns was influenced by the minimum samples split hyperparameter. A smaller value encouraged splits that were more complicated, which might result in overfitting, whereas a greater value encouraged divides that were simpler, which could result in underfitting.

The model's performance was enhanced by carefully adjusting the hyperparameters. Grid search and randomized search are two methods that could be used to methodically explore the hyperparameter space and identify the ideal combination. It is significant to keep in mind that the effects of each hyperparameter can change based on the particular dataset and issue at hand. As a result, a thorough investigation should be carried out to determine the hyperparameters that produce the greatest performance for the specified task.

In conclusion, the choice of hyperparameters had an impact on the model's performance. Finding the ideal configuration that maximized the model's predictive power required careful analysis and trial with various settings. The generalizability and robustness of the model can be increased through further optimization and validation on different datasets.

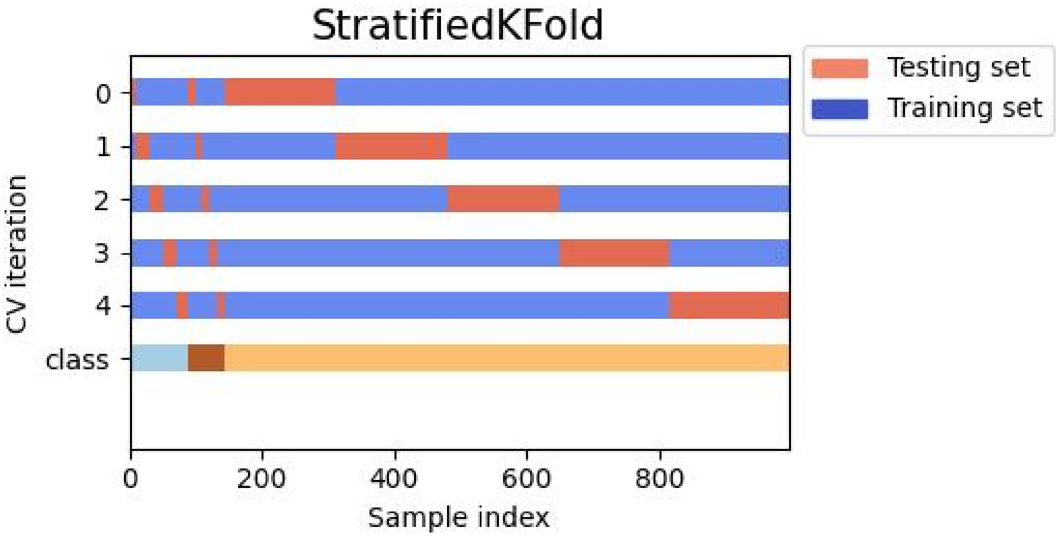
Table 3: showing the accuracy score for the different random forest hyperparameters

Experiments	**Decision tree depth**	**Number of estimators**	**Accuracy (%)**
1	3	10	93.6
2	4	20	96.9
3	5	30	96.9
4	6	50	98.3
5	10	100	98.7

4.3.1 Stratified K-fold

From the cross-fold diagram below, the dataset is divided in a stratified manner, picking instances from each class of the diabetic’s dataset. This shows that the training and test set is taken from every class and none is omitted. This method helped to reduce the overfitting of the data to the model (Fushiki, 2011). Cross-fold validation, a method frequently employed to gauge a model's generalizability, was utilized to evaluate the model's performance. The dataset was divided into 5 subsets (folds), and the model was trained and tested on various fold combinations to provide a more accurate performance estimate.

Figure 12: 5-fold split for the cross-validation



From the above model evaluation result, 5 experiment was done to obtain different accuracy with the variation in the maximum tree depth and the number of estimators in the random forest. In the section, we use the same hyperparameter values for in a stratified cross fold cross validation.

Experiment 1

Table 4: Validation score from experiment 1.

Fold	1	2	3	4	5
Score	92.4	98.9	92.9	89.9	88.4

Mean Score	92.6

Experiment 2

Table 5: Validation score from experiment 2.

Fold	1	2	3	4	5
Score	95.5	98.9	97.9	97.5	93.9
Mean Score	96.8				

Experiment 3

Table 6: Validation score from experiment 3.

Fold	1	2	3	4	5
Score	95.9	97.5	98.5	97.9	94.9
Mean Score	96.9				

Experiment 4

Table 7: Validation score from experiment 4.

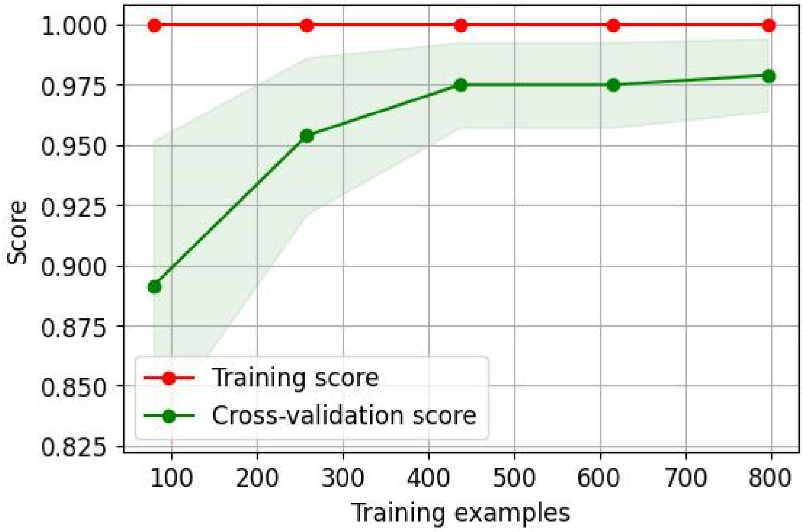
Fold	1	2	3	4	5
Score	95.5	98.5	98.9	98.5	94.5
Mean Score	97.2				

Experiment 5

Table 8: Validation score from experiment 5.

Fold	1	2	3	4	5
Score	96.9	98.9	98.9	98.9	93.9
Mean Score	97.6				

Figure 13: Showing the learning curve for the best validation score, which is experiment 5



Cross-fold validation was used to evaluate the model's robustness and consistency across several data subsets. A more accurate assessment is given by averaging performance over all folds, which minimizes the potential effects of data variability or biased splits. It reduces the possibility of overfitting or producing unduly optimistic performance estimates by considering several cycles of training and testing. Understanding the stability and variation of the model requires examination of the performance measures over several folds. The performance of the model is heavily dependent on the precise subset of data used for training and testing if there is a notable variation in performance across folds. The model may be able to generalize effectively to new data, however, if performance is constant between folds. Comparing various models or parameter combinations is also made easier with cross-fold validation. It offers a fair and unbiased performance comparison, assisting in the choice of the best-performing model, by evaluating each model using the same folds.

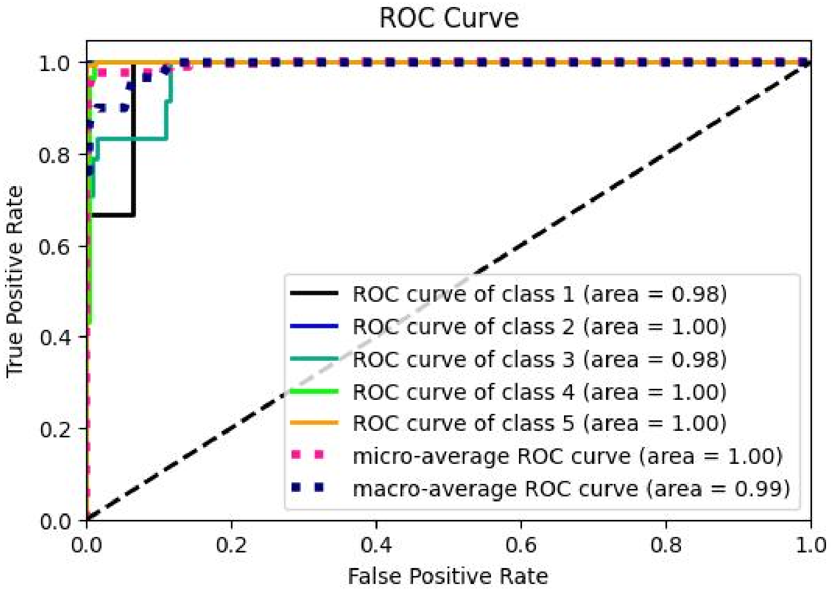
4.3.2 ROC for AUC

Impressive Area Under the Curve (AUC) scores were obtained for the five classes in the classification model. The outstanding discrimination abilities of the model across the various classes are indicated by the AUC values of 0.98, 1, 0.98, 1, and 1.

The high AUC values indicate that RF model has achieved a nearly perfect balance between true positive and false positive rates, proving its capability to precisely distinguish between the various classes. These findings show a high level of confidence in the predictions made by the model and its success in categorizing cases. Additionally, the robustness of the model's performance is shown by the consistency of the AUC values across the bulk of the classes. This consistency shows that the model can make accurate predictions in all situations without being biased towards any particular classes.

However, it is crucial to consider the dataset's imbalance, particularly if some classes have a disproportionately small number of instances compared to others. The model's performance on minority classes should be thoroughly understood by analyzing the precision, recall, and F1-score metrics for each class separately, even though the high AUC values indicate good overall performance.

Figure 14: Receiver operating characteristic (ROC) curve.



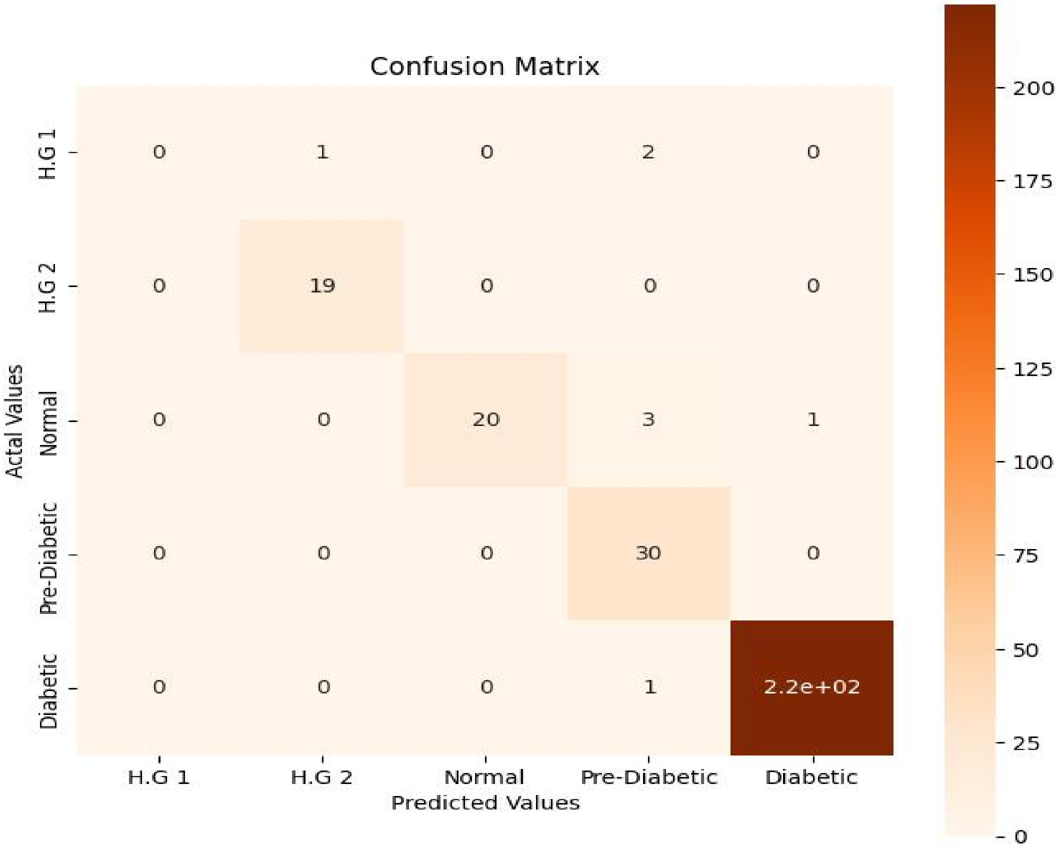
The ROC curve of a hypothetical classifier is depicted in figure 14, along with the diagonal reference line (random guessing). The model performs better when the curve is closer to the top-left corner. The area under the curve (AUC-ROC) value evaluates the model's overall discriminative capacity.

In conclusion, the obtained AUC results show the model's high ability to discriminate between instances belonging to the five classes. Its efficacy and robustness are indicated by the consistently high values. The model's performance will be better understood with further examination of class-specific measures, particularly for classes with imbalances. Overall, these findings imply that the model is capable of discriminating between the various classes accurately and shows potential for real-world use in the field of this work.

4.3.3 Confusion matrix

The confusion matrix offers a thorough analysis of a classification model's performance, particularly in terms of the predictions made for each class. A confusion matrix would reveal how well the model performed for each of the five classes in the 5-class classification model. The true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class can be found by analyzing the confusion matrix. These metrics provide a thorough evaluation of the model's effectiveness during the whole categorization operation. The TP are the diagonal and the errors are the remaining boxes.

Figure 15: Confusion matrix used to represent the performance of the model



	H.G 1	H.G 2	Normal	Pre-Diabetic	Diabetic
H.G 1	0	1	0	2	0
H.G 2	0	19	0	0	0
Normal	0	0	20	3	1
Pre-Diabetic	0	0	0	30	0
Diabetic	0	0	0	1	2.2e+02

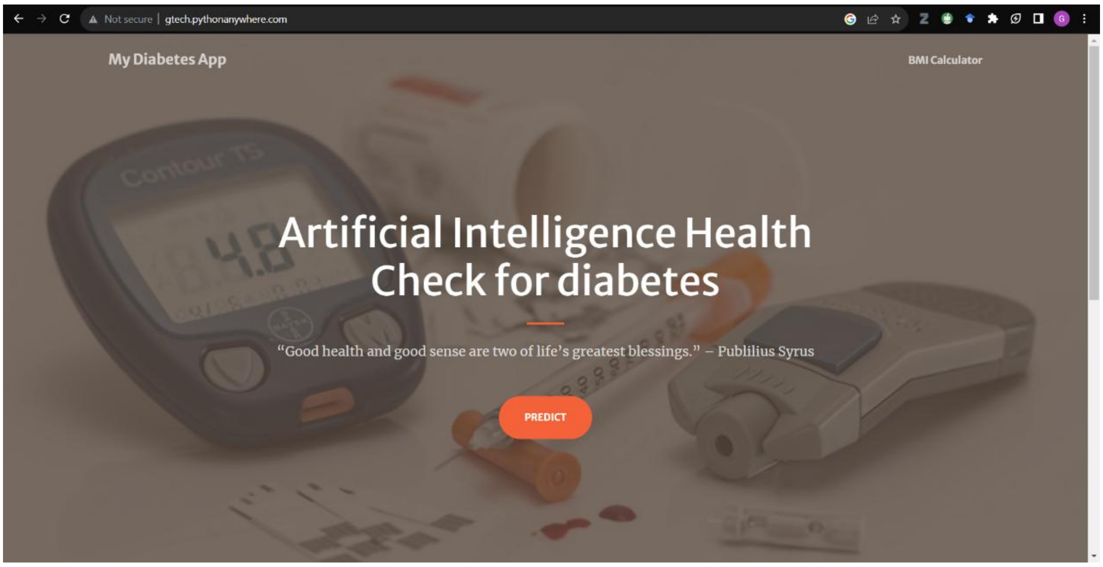
It is possible to spot any patterns or trends in misclassifications by looking at the confusion matrix. By figuring out whether particular classes are routinely misclassified as others, for instance, you might find possible areas where the model can be improved. A more thorough assessment of the model's performance for each class is possible thanks to the confusion matrix's capability to calculate metrics like precision, recall, and F1-score on a per-class basis. A comprehensive picture of your model's performance can be achieved by interpreting the confusion matrix combined with the AUC data you obtained. It enables you to evaluate the model's strengths and shortcomings for each individual class in addition to its overall accuracy.

In conclusion, the confusion matrix offers a thorough analysis of the model's predictions for each class, allowing for a fine-grained assessment of its effectiveness.

You can learn more about class-specific performance data and pinpoint opportunities for development by evaluating the matrix.

4.4 Web application results

Figure 16: Home page view



Not secure | gtech.pythonanywhere.com

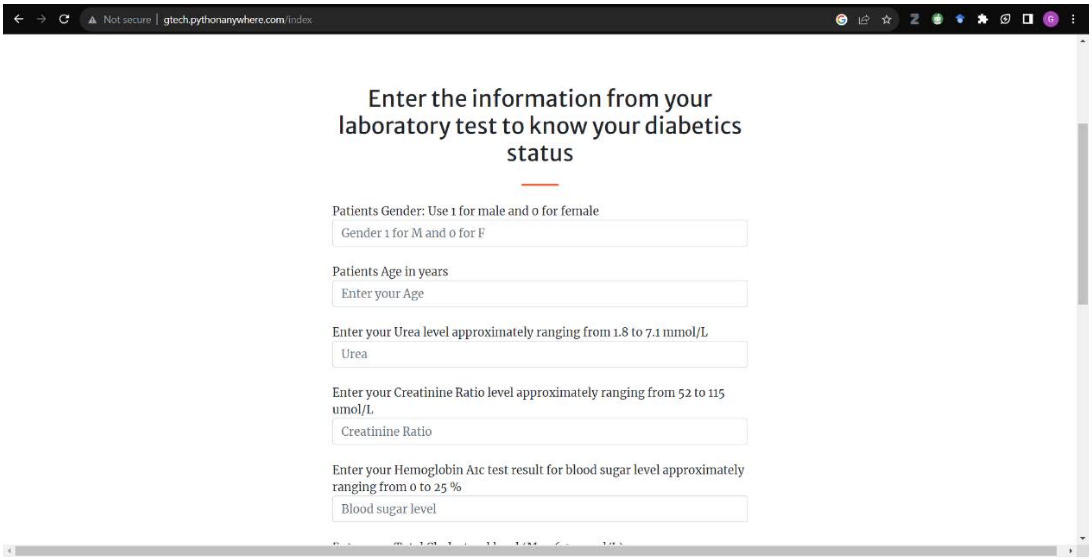
My Diabetes App

BMI Calculator

Artificial Intelligence Health Check for diabetes

“Good health and good sense are two of life’s greatest blessings.” – Publilius Syrus

Figure 17: Form page to enter patients test results (a)



Enter the information from your laboratory test to know your diabetics status

Patients Gender: Use 1 for male and 0 for female

Gender 1 for M and 0 for F

Patients Age in years

Enter your Age

Enter your Urea level approximately ranging from 1.8 to 7.1 mmol/L

Urea

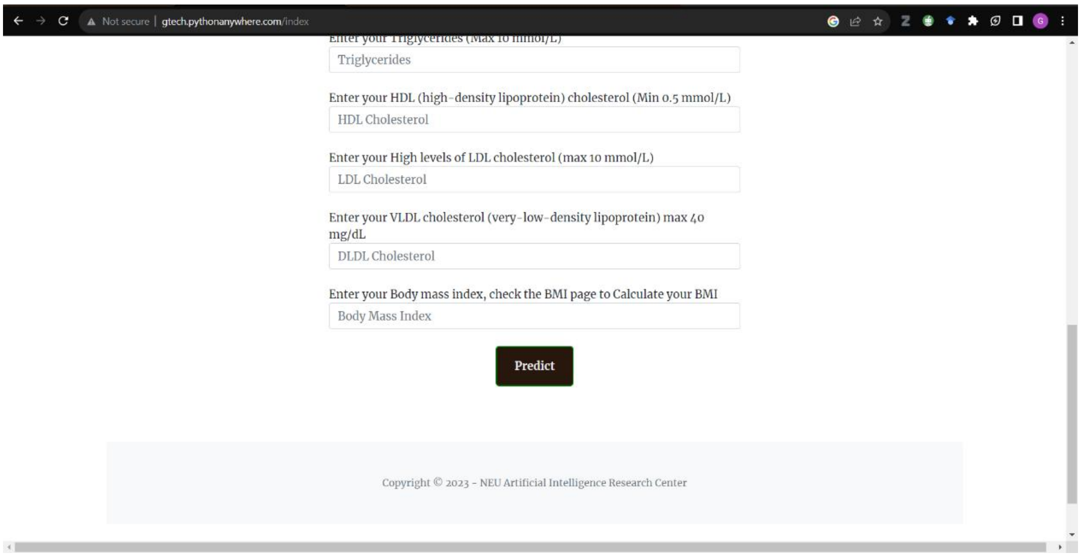
Enter your Creatinine Ratio level approximately ranging from 52 to 115 umol/L

Creatinine Ratio

Enter your Hemoglobin A1c test result for blood sugar level approximately ranging from 0 to 25 %

Blood sugar level

Figure 18:Form page to enter patients test results (b)



Enter your Triglycerides (Max 10 mmol/L)

Triglycerides

Enter your HDL (high-density lipoprotein) cholesterol (Min 0.5 mmol/L)

HDL Cholesterol

Enter your High levels of LDL cholesterol (max 10 mmol/L)

LDL Cholesterol

Enter your VLDL cholesterol (very-low-density lipoprotein) max 40 mg/dL

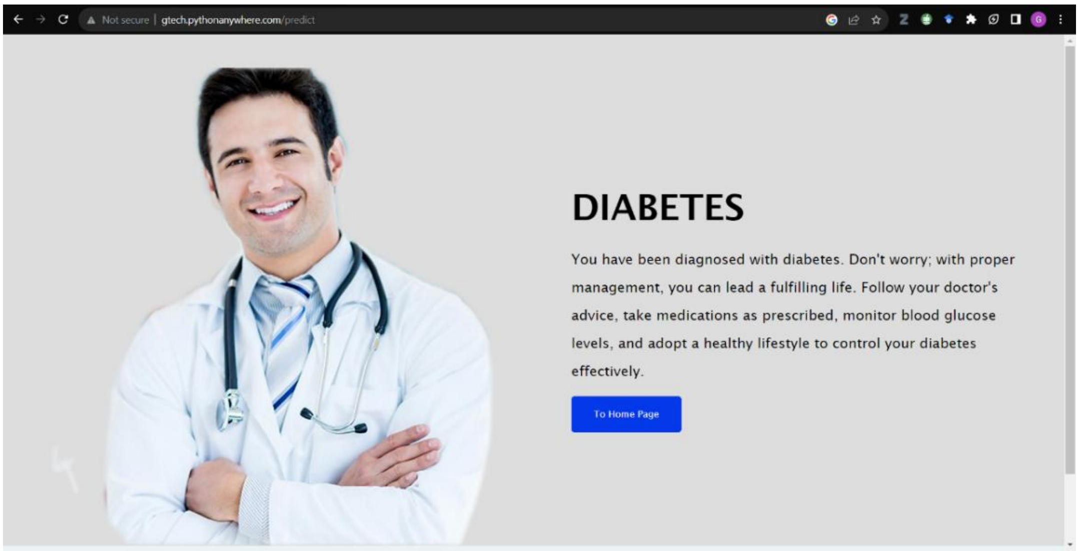
DLDL Cholesterol

Enter your Body mass index, check the BMI page to Calculate your BMI

Body Mass Index

Copyright © 2023 - NEU Artificial Intelligence Research Center

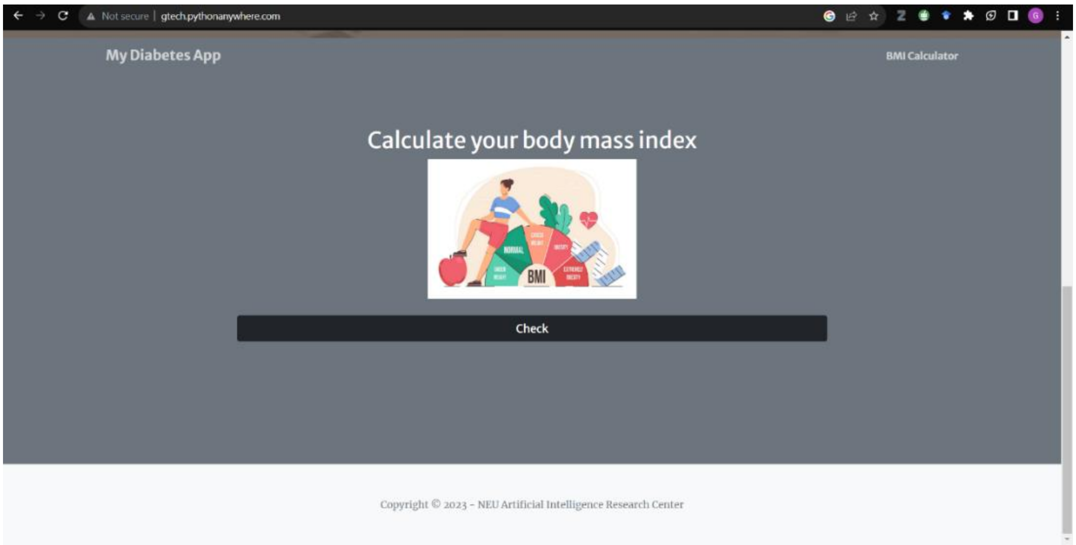
Figure 19: Prediction result page



DIABETES

You have been diagnosed with diabetes. Don't worry; with proper management, you can lead a fulfilling life. Follow your doctor's advice, take medications as prescribed, monitor blood glucose levels, and adopt a healthy lifestyle to control your diabetes effectively.

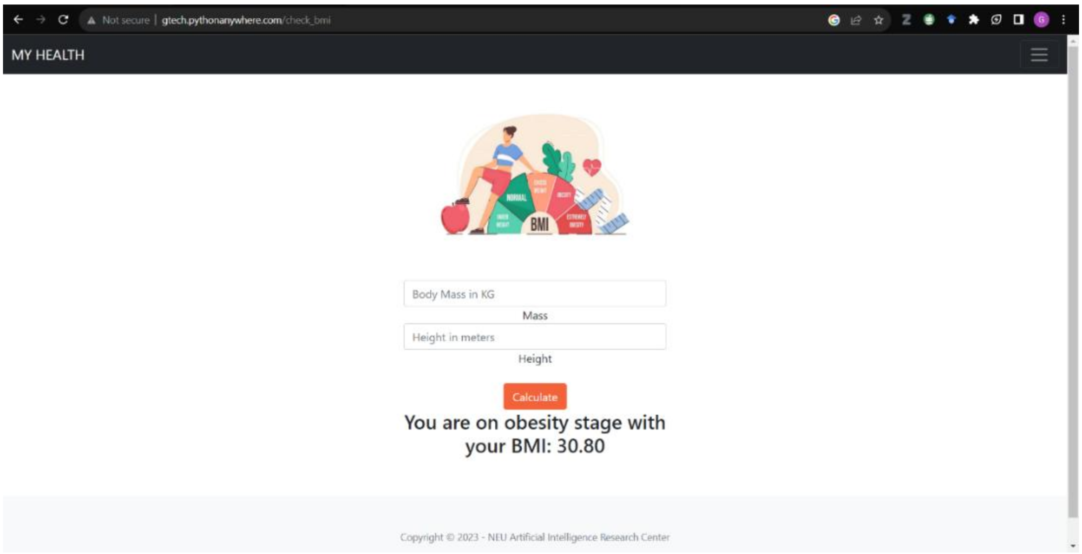
Figure 20: Page for Body mass index calculation



Calculate your body mass index

Copyright © 2023 - NEU Artificial Intelligence Research Center

Figure 21: Page for Body mass index calculation result page



Body Mass in KG

Mass

Height in meters

Height

You are on obesity stage with your BMI: 30.80

Copyright © 2023 - NEU Artificial Intelligence Research Center

This section shows the results and the overview of the applicable web app designed to server the model and to create an easy way for patient or the user to predict their diabetes status.

Here is a breakdown of how a user can use your web app to predict diabetes:

4.4.1 Home page

The user is welcomed by the home page, which serves as the main interface, when they access the web application as shown in figure 16.

The home page is created to be both aesthetically pleasing and user-friendly, with distinct navigation options for various functionalities.

4.4.2 BMI calculation

The user is taken to the BMI calculation page after clicking the "BMI Calculation" link on the homepage. A simple form with input boxes for height and weight may be found on the BMI page as shown in figure 21. The user fills out the appropriate areas with their height in centimeters and weight in kilos. The user enters the data and then selects "Calculate BMI" from the menu. If the user is underweight, normal weight, overweight, or obese, the BMI result is immediately calculated using the formula (BMI = weight / height ^{2}) and shown on the same page. Without leaving the page, the visitor can simply make several BMI estimations.

4.4.3 Diabetes prediction

The "Diabetes Prediction" link is located on the home page, and users can click it to determine their likelihood of developing diabetes. They will be taken to the diabetes prediction website, where a thorough form is offered. The diabetes risk assessment prediction form has a number of sections to gather pertinent data, including age, gender, Urea, and other pertinent risk factors. Additionally, the form has tooltips and clear instructions to help users fill out the form with the necessary and precise information. The user starts the process of assessing their risk of developing diabetes by filling out the form and clicking the "Predict" button.

4.4.4 Displaying the Prediction

From figure 19, the web app's back-end processes the data using the Random Forest model API when the user submits the prediction form. The Random Forest model evaluates user inputs and makes a prediction in light of the supplied data. On the prediction page, the prediction outcome, which may be "Diabetic", "Pre-diabetic", "Normal", "Hypoglycemia stage 1" and "Hypoglycemia stage 2", is displayed right away along with a confidence level or likelihood. The user also gets helpful comments on the forecast, including any potential risk factors found in the supplied data. The prediction page also provides explicit instructions on how to interpret the outcomes and, if necessary, seek expert medical help.

4.4.5 Responsive Design

The web app has a responsive structure, which ensures that customers may access and utilize the app on a variety of platforms, including desktops, laptops, tablets, and smartphones.

CONCLUSION

In this study, we investigated a diabetes dataset in-depth with the goal of enhancing classification precision and offering more thorough insights into the disease spectrum. We added two more classes to the original three-class classification—stages 1 and 2 of hypoglycemia—by including the HbA1c characteristic.

Through the evaluation of our model, we were able to appropriately classify cases into the five categories with noteworthy accuracy. Cross validation, confusion matrix, and AOC were some of the performance indicators that gave a comprehensive picture of the model's efficiency. The model's stability and generalizability were established by cross-fold validation, which considered fluctuations in the dataset. The HbA1c attribute's inclusion turned out to be essential because it offered useful data for risk classification and individualized treatment plans. With the use of this improved classification system, medical personnel can recognize patients in various hypoglycemic phases and tailor their therapies accordingly.

The outcomes also highlight how important feature engineering and data pretreatment are for enhancing classification outcomes. The dataset's quality and interpretability were improved by the use of relevant approaches, including data cleaning and encoding categorical categories. The suggested paradigm showed promising results in properly predicting diabetes in individuals after considerable testing and experimentation. The Random Forest algorithm performed well, differentiating between various glucose tolerance or diabetes stages with excellent accuracy.

The created web application improved the framework's usability and accessibility by giving users a simple way to enter their data and get individualized risk forecasts.

The software also provided graphics and insights into the underlying causes, enabling users to grasp the prediction model and their individual health concerns on a deeper level.

By offering a useful tool for the early identification of people at risk, the study's findings benefit the area of diabetes care. Early detection of prediabetics and diabetics enables the implementation of focused therapies and preventive measures to slow the disease's course and enhance patient outcomes.

Although the framework has shown encouraging outcomes, there is still room for development and further study. To improve prediction accuracy and interpretability, future research may investigate the use of additional machine learning methods or feature engineering approaches. Additionally, adding longitudinal data and carrying out validation studies in other groups would support the framework's generalizability and efficiency.

Overall, this study integrates cutting-edge machine learning methods with a user-friendly online application to give a comprehensive framework for diabetes prediction. By enabling early identification, targeted therapies, and improved management techniques for people at risk of developing diabetes, the findings of this study have the potential to have a significant influence on healthcare.

REFERENCE

Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Zarei, M. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review. *Diabetology & Metabolic Syndrome*, *14*, 196. https://doi.org/10.1186/s13098-022-00969-9

Ahamed, B. S., Arya, M. S., & Nancy V, A. O. (2022). Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques. *Frontiers in Computer Science*, *4*. https://www.frontiersin.org/articles/10.3389/fcomp.2022.835242

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(none), 40–79. https://doi.org/10.1214/09-SS054

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-022-07049-z

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1), 131–156. https://doi.org/10.1016/S1088-467X(97)00008-5

Enomoto, M., Yoshii, H., Mita, T., Sanke, H., Yokota, A., Yamashiro, K., Inagaki, N., Gosho, M., Chie, O., Kudo, K., Watada, H., & Onuma, T. (2015). Relationship between dietary pattern and cognitive function in elderly patients with type 2 diabetes mellitus. *The Journal of International Medical Research*, *43*. https://doi.org/10.1177/0300060515581672

Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C. B., Martínez-González, M. A., Salas-Salvadó, J., & Hu, F. B. (2016). Metabolomics in Prediabetes and

Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care*, *39*(5), 833–846. https://doi.org/10.2337/dc15-2251

Hasan, Md. K., Alam, Md. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2020.2989857

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-09954-8

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*, *47*, 45–51. https://doi.org/10.1016/j.procs.2015.03.182

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

Kumari, S., & Singh, A. (2013). A data mining approach for the diagnosis of diabetes mellitus. *2013 7th International Conference on Intelligent Systems and Control (ISCO)*, 373–375. https://doi.org/10.1109/ISCO.2013.6481182

Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science*, *112*, 2519–2528. https://doi.org/10.1016/j.procs.2017.08.193

Rashid, A. (2020). *Diabetes Dataset* [Data set]. Mendeley. https://doi.org/10.17632/WJ9RWKP9C2.1

Roglic. (n.d.). *WHO Global report on diabetes: A summary*. Retrieved April 6, 2023, from https://www.ijncd.org/article.asp?issn=2468-8827;year=2016;volume=1;issue=1;spage=3;epage=8;aulast=Roglic;type=3

Samant, P., & Agarwal, R. (2018). Machine learning techniques for medical diagnosis of diabetes using iris images. *Computer Methods and Programs in Biomedicine*, *157*, 121–128. https://doi.org/10.1016/j.cmpb.2018.01.004

Woldaregay, A. Z., Årsand, E., Botsis, T., Albers, D., Mamykina, L., & Hartvigsen, G. (2019). Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes. *Journal of Medical Internet Research*, *21*(5), e11030. https://doi.org/10.2196/11030

A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sensing*, *11*(2), 185. https://doi.org/10.3390/rs11020185

Ahmad, I., Suwarni, E., Borman, R. I., Asmawati, Rossi, F., & Jusman, Y. (2021). Implementation of RESTful API Web Services Architecture in Takeaway Application Development. *2021 1st International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 132–137. https://doi.org/10.1109/ICE3IS54102.2021.9649679

Al-Turjman, F., Hussain, A. A., Alturjman, S., & Altrjman, C. (2022). Vehicle Price Classification and Prediction Using Machine Learning in the IoT Smart Manufacturing Era. *Sustainability*, *14*(15), 9147. https://doi.org/10.3390/su14159147

Azad, C., Bhushan, B., Sharma, R., Shankar, A., Singh, K. K., & Khamparia, A. (2022). Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. *Multimedia Systems*, *28*(4), 1289–1307. https://doi.org/10.1007/s00530-021-00817-2

Bartolome, A., & Prioleau, T. (2022). A computational framework for discovering digital biomarkers of glycemic control. *Npj Digital Medicine*, *5*(1). https://doi.org/10.1038/s41746-022-00656-z

Bhuiyan, M. F. U., Rahman, M. T., Anik, M. A., & Khan, M. (2021). A Framework for Type-II Diabetes Prediction Using Machine Learning Approaches. *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*. https://doi.org/10.1109/ICCCNT51525.2021.9580158

Bidari, I., Chickerur, S., Kulkarni, A., Mahajan, A., Nikkam, A., & Thm, A. (2021). Deploying Machine Learning Inference on Diabetic Retinopathy in Binary and Multi-class Classification. *ICIERA 2021 - 1st International Conference on Industrial Electronics Research and Applications, Proceedings*. https://doi.org/10.1109/ICIERA53202.2021.9726533

Buyrukoğlu, S., & Akbaş, A. (2022). Machine Learning based early prediction of type 2 diabetes: a new hybrid feature selection approach using correlation matrix with heatmap and SFS. *Balk. J. Electr. Comput. Eng.*, *10*(2), 110–117.

Connie, T., Tan, Y. F., Goh, M. K. O., Hon, H. W., Kadim, Z., & Wong, L. P. (2022). Explainable health prediction from facial features with transfer learning. *Journal of Intelligent and Fuzzy Systems*, *42*(3), 2491–2503. https://doi.org/10.3233/JIFS-211737

Cook, N. R. (2007). Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation*, *115*(7), 928–935. https://doi.org/10.1161/CIRCULATIONAHA.106.672402

Cujilan, Y. C., Pérez, D. P., Gutiérrez, C. G., Gutiérrez, Á. G., Tobar, M. B., Mora, C. M., & Rivera, D. I. (2022). Supervised Learning Techniques for the Optimization of Diagnosis Processes of Diabetes in Public Health Centers | Técnicas de Aprendizaje

Supervisado para la Optimización de Procesos de Diagnostico de Diabetes en los Centros de Salud Públicos. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*, *2022-July*. https://doi.org/10.18687/LACCEI2022.1.1.779

Daghistani, T., & Alshammari, R. (2020). Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology*, *11*(2), 78–83. https://doi.org/10.12720/jait.11.2.78-83

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., & Bellazzi, R. (2017). Machine learning methods to predict diabetes complications. *J. Diab. Sci. Technol*, *12*.

Das, U., Yakin Srizon, A., Ansarul Islam, M., Sikder Tonmoy, D., & Al Mehedi Hasan, M. (2020). Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers. *2020 2nd International Conference on Sustainable Technologies for Industry 4.0, STI 2020*. https://doi.org/10.1109/STI50764.2020.9350498

de Hond, A. A. H., Steyerberg, E. W., & van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, *4*(12), e853–e855. https://doi.org/10.1016/S2589-7500(22)00188-1

Deebak, B. D., & Al-Turjman, F. (2023). EEI-IoT: Edge-Enabled Intelligent IoT Framework for Early Detection of COVID-19 Threats. *Sensors*, *23*(6), 2995. https://doi.org/10.3390/s23062995

Del Giorno, S., D’Antoni, F., Piemonte, V., & Merone, M. (2023). A New Glycemic closed-loop control based on Dyna-Q for Type-1-Diabetes. *Biomedical Signal Processing and Control*, *81*. https://doi.org/10.1016/j.bspc.2022.104492

Dimililer, K., Teimourian, H., & Al-Turjman, F. (2022). Radio galaxies classification system using machine learning techniques in the IoT Era. *Journal of Experimental & Theoretical Artificial Intelligence*, 1–13. https://doi.org/10.1080/0952813X.2022.2080277

Elseddawy, A. I., Karim, F. K., Hussein, A. M., & Khafaga, D. S. (2022). Predictive Analysis of Diabetes-Risk with Class Imbalance. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3078025

Fahim, F., Al Farabi, A., Hasan, M. S., & Hasan, M. M. (2022). Diagnosis of Diabetes using Clinical Features: An Analysis based on Machine Learning Techniques. *3rd International Informatics and Software Engineering Conference, IISEC 2022*. https://doi.org/10.1109/IISEC56263.2022.9998257

Felizardo, V., Garcia, N. M., Pombo, N., & Megdiche, I. (2021). Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction – A systematic literature review. *Artificial Intelligence in Medicine*, *118*. https://doi.org/10.1016/j.artmed.2021.102120

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906–914. https://doi.org/10.1093/bioinformatics/16.10.906

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Gneiting, T., & Vogel, P. (2022). Receiver operating characteristic (ROC) curves: equivalences, beta model, and minimum distance estimation. *Machine Learning*, *111*(6), 2147–2159. https://doi.org/10.1007/s10994-021-06115-2

Goel, R., & Satish, C. J. (2023). Precision Monitoring of Health-Care Using Big Data and Java from Social Networking and Wearable Devices. *2023 3rd International Conference on Intelligent Communication and Computational Techniques, ICCT 2023*. https://doi.org/10.1109/ICCT56969.2023.10075744

Gopal, V. N., Al-Turjman, F., Kumar, R., Anand, L., & Rajesh, M. (2021). Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement*, *178*, 109442. https://doi.org/10.1016/j.measurement.2021.109442

Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing*, *12*(3), 3715–3723. https://doi.org/10.1007/s12652-019-01652-0

Hassan, M. M., & Amiri, N. N. (2019a). Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms. *IV.International Conference on Theoretical and Applied Computer Science and Engineering*.

Hassan, M. M., & Amiri, N. N. (2019b). Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms. *IV.International Conference on Theoretical and Applied Computer Science and Engineering*.

Hu, F., & Li, H. (2013). A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*, *2013*, 1–10. https://doi.org/10.1155/2013/694809

Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., Fan, L., Yin, H., Xu, Y., & Li, J. (2020). Sample imbalance disease classification model based on association rule feature selection. *Pattern Recognition Letters*, *133*, 280–286. https://doi.org/10.1016/j.patrec.2020.03.016

Huckvale, E., & Moseley, H. N. B. (2023). kegg_pull: a software package for the RESTful access and pulling from the Kyoto Encyclopedia of Gene and Genomes. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05208-0

Hussain, A. A., & Al-Turjman, F. (2022). Hybrid Genetic Algorithm for IOMT-Cloud Task Scheduling. *Wireless Communications and Mobile Computing*, *2022*, 1–14. https://doi.org/10.1155/2022/6604286

Irkham, I., Ibrahim, A. U., Nwekwo, C. W., Al-Turjman, F., & Hartati, Y. W. (2022). Current Technologies for Detection of COVID-19: Biosensors, Artificial Intelligence and Internet of Medical Things (IoMT): Review. *Sensors*, *23*(1), 426. https://doi.org/10.3390/s23010426

Jahanur Rahman, Md., Ahammed, B., Maniruzzaman, Md., & Menhazul Abedin, Md. (2020). Classification and predictiosn of diabetes diseases using machine learning paradigms. *Health Information Science & Systems*.

Jakka, A., & Rani, J. V. (2019). Performance evaluation of machine learning models for diabetes prediction. *Int J Innov Technol Explor Eng*, 8.

Junaid, M., Sohail, A., Turjman, F. Al, & Ali, R. (2022). Agile Support Vector Machine for Energy-efficient Resource Allocation in IoT-oriented Cloud using PSO. *ACM Transactions on Internet Technology*, *22*(1), 1–35. https://doi.org/10.1145/3433541

Krstinic, D., Seric, L., & Slapnicar, I. (2023). Comments on ‘MLCM: Multi-Label Confusion Matrix’. *IEEE Access*, 1. https://doi.org/10.1109/ACCESS.2023.3267672

Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting. *Annals of Data Science*, *10*(1), 183–208. https://doi.org/10.1007/s40745-021-00344-x

Lyman, D., Natale, D., Schriml, L., Anton, K., Crichton, D. C., & Mazumder, R. (2021). Analysis of Biomarker Data Towards Development of a Molecular Biomarker Ontology. *CEUR Workshop Proceedings*, *3073*, 99–103.

Maheshwari, D., Garcia-Zapirain, B., & Sierra-Soso, D. (2020). Machine learning applied to diabetes dataset using quantum versus classical computation. In Proceedings of the 2020 IEEE international symposium on signal processing and information technology (ISSPIT), Louisville, KY. *USA*, 2020.

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Mani, S., Chen, Y., Elasy, T., Clayton, W., & Denny, J. (2012). Type 2 diabetes risk forecasting from EMR data using machine learning. *AMI*.

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*, *9*(4), 81. https://doi.org/10.3390/technologies9040081

Mary, X. A., Raj, A. P. W., Evangeline, C. S., Neebha, T. M., Kumaravelu, V. B., & Manimegalai, P. (2023). Multi-class Classification of Gastrointestinal Diseases using Deep Learning Techniques. *Open Biomedical Engineering Journal*, *17*(1). https://doi.org/10.2174/18741207-v17-e230215-2022-HT27-3589-11

Mastoli, M. M., Pol, U. R., Kulkarni, R. V., & Patil, R. (2022). Artificial intelligence and machine learning techniques for diabetes health-care. In *Blockchain for 5G Healthcare Applications: Security and privacy solutions*.

Mishra, S., Mallick, P. K., Tripathy, H. K., Bhoi, A. K., & González-Briones, A. (2020). Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. *Applied Sciences (Switzerland)*, *10*(22), 1–35. https://doi.org/10.3390/app10228137

Munna, M. T. A., Alam, M. M., Allayear, S. M., Sarker, K., & Ara, S. J. F. (2020). Prediction model for prevalence of type-2 diabetes complications with ANN approach combining with K-fold cross validation and K-means clustering. In *Lecture Notes in Networks and Systems* (Vol. 69). https://doi.org/10.1007/978-3-030-12388-8_71

Ogwo, O., Turabieh, H., Sheta, A., & King, S. (2019). Medical Data Classification Using Binary Brain Storm Optimization Algorithm. *ACM International Conference Proceeding Series*, 44–52. https://doi.org/10.1145/3388218.3388224

Pan, C., Poddar, A., Mukherjee, R., & Ray, A. K. (2022). Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction. *Biomedical Signal Processing and Control*, *76*. https://doi.org/10.1016/j.bspc.2022.103666

Pankaj, C., Singh, K. V., & Singh, K. R. (2021). Artificial Intelligence enabled Web-Based Prediction of Diabetes using Machine Learning Approach. *Proceedings of IEEE International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications, CENTCON 2021*, 60–64. https://doi.org/10.1109/CENTCON52345.2021.9688236

Parajuli, S., Parajuli, S., & Guragai, M. K. (2022). A data-driven approach to predict the risk of readmission among patients with Diabetes Mellitus. *2022 2nd International Conference on Artificial Intelligence and Signal Processing, AISP 2022*. https://doi.org/10.1109/AISP53593.2022.9760601

Park, G., Chandrasegar, V. K., & Koh, J. (2023). Accuracy Enhancement of Hand Gesture Recognition Using CNN. *IEEE Access*, *11*, 26496–26501. https://doi.org/10.1109/ACCESS.2023.3254537

Patra, N., Pramanik, J., Samal, A. K., & Pani, S. K. (2022). Machine Learning Application in Primitive Diabetes Prediction—A Case of Ensemble Learning. In *Lecture Notes in Networks and Systems* (Vol. 375). https://doi.org/10.1007/978-981-16-8763-1_64

Rabiha, S. G., Wibowo, A., Lukas, & Heryadi, Y. (2021). Diabetes Classification Using Support Vector Machine: Binary Classification Model. *ICOIACT 2021 - 4th International Conference on Information and Communications Technology: The Role of AI in Health and Social Revolution in Turbulence Era*, 280–284. https://doi.org/10.1109/ICOIACT53268.2021.9563994

Rjoob, K., McGilligan, V., Bond, R., Watterson, S., Chemaly, M., McAlister, R., De Melo Malaquias, T., Leslie, S. J., Knoery, C., Iftikhar, A., Bjourson, A., & Peace, A. (2020). Improving the Detection of Acute Coronary Syndrome Using Machine Learning of Blood Biomarkers. *Computing in Cardiology*, *2020-Septe*. https://doi.org/10.22489/CinC.2020.337

Roy, A., & Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety*, *233*, 109126. https://doi.org/10.1016/j.ress.2023.109126

Roy, S., Bhateja, G., Gulati, G., & Saxena, S. (2022). Physiological Parameter Analysis for Type-1 Diabetes and ML Approach for Insulin Prediction. *PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing*, 606–611. https://doi.org/10.1109/PDGC56933.2022.10053195

Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., & Shapi'i, A. (2021). A systematic literature review on obesity: Understanding the causes &amp; consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, *136*. https://doi.org/10.1016/j.compbiomed.2021.104754

Sampath, P., Packiriswamy, G., Kumar, N. P., Shanmuganathan, V., Song, O.-Y., Tariq, U., & Nawaz, R. (2020). IoT based health—related topic recognition from emerging online health community (Med help) using machine learning technique. *Electronics (Switzerland)*, *9*(9), 1–15. https://doi.org/10.3390/electronics9091469

Sangkatip, W., & Phuboon-Ob, J. (2020). Non-Communicable Diseases Classification using Multi-Label Learning Techniques. *InCIT 2020 - 5th International Conference on Information Technology*, 17–21. https://doi.org/10.1109/InCIT50588.2020.9310978

Sania Febriani, & Fitri Purwaningtias. (2022). Implementasi Platform As A Service (PAAS) Pada Aplikasi Getfix Berbasis Cloud Computing. *Jurnal Sains Dan Informatika*, *8*(2), 86–95. https://doi.org/10.22216/jsi.v8i2.1653

Schepart, A., Burton, A., Durkin, L., Fuller, A., Charap, E., Bhambri, R., & Ahmad, F. S. (2023). Artificial intelligence–enabled tools in cardiovascular medicine: A survey of current use, perceptions, and challenges. *Cardiovascular Digital Health Journal*. https://doi.org/10.1016/j.cvdhj.2023.04.003

Severeyn, E., Velásquez, J., Herrera, H., Wong, S., & Cruz, A. L. (2021). Analysis of Receiver Operating Characteristic Curve Using Anthropometric Measurements for Obesity Diagnosis. In *Advances in Intelligent Systems and Computing: Vol. 1273 AISC*. https://doi.org/10.1007/978-3-030-59194-6_7

Shah, J., & Patel, R. (2019). Classification techniques for Disease detection using Big-data. *4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, ICEECCOT 2019*, 140–145. https://doi.org/10.1109/ICEECCOT46775.2019.9114589

Shrivastava, A. K., Karthikeyan, V., Kaushik, S., & Sudagar, M. (2022). Early Diabetes Prediction using Random Forest. *3rd International Conference on Electronics and Sustainable Communication Systems, ICESC 2022 - Proceedings*, 1154–1159. https://doi.org/10.1109/ICESC54411.2022.9885683

Sreenivasu, S. V. N., Gupta, S., Vatsa, G., Shrivastava, A., Vashisht, S., & Srivastava, A. (2022). Carbohydrate Recommendation for Type-1 Diabetics Patient Using Machine Learning. *Proceedings of 5th International Conference on Contemporary Computing and Informatics, IC3I 2022*, 600–604. https://doi.org/10.1109/IC3I56241.2022.10072919

Stephan, T., Al-Turjman, F., Ravishankar, M., & Stephan, P. (2022). Machine learning analysis on the impacts of COVID-19 on India's renewable energy transitions and air quality. *Environmental Science and Pollution Research*, *29*(52), 79443–79465. https://doi.org/10.1007/s11356-022-20997-2

Tao, X., Jiang, M., Liu, Y., Hu, Q., Zhu, B., Hu, J., Guo, W., Wu, X., Xiong, Y., & Shi, X. (2022). Predicting three-month fasting blood glucose and glycated hemoglobin of patients with type 2 diabetes based on multiple machine learning algorithms. *Research Square*.

Thirunavukkarasu, U., & Umapathy, S. (2020). Classification of Prediabetes and Healthy Subjects in Plantar Infrared Thermal Imaging Using Various Machine Learning Algorithms. In *Lecture Notes in Networks and Systems* (Vol. 106). https://doi.org/10.1007/978-981-15-2329-8_9

Tiwari, S., Gupta, N., & Yadav, P. (2021). Diabetes Type2 Patient Detection Using LASSO Based CFFNN Machine Learning Approach. *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*, 602–608. https://doi.org/10.1109/SPIN52536.2021.9565965

Umar Ibrahim, A., Al-Turjman, F., Ozsoz, M., & Serte, S. (2022). Computer aided detection of tuberculosis using two classifiers. *Biomedical Engineering / Biomedizinische Technik*, *67*(6), 513–524. https://doi.org/10.1515/bmt-2021-0310

Umar Ibrahim, A., Pwavodi, P. C., Ozsoz, M., Al-Turjman, F., Galaya, T., & Agbo, J. J. (2023). Crispr biosensing and Ai driven tools for detection and prediction of Covid-19. *Journal of Experimental & Theoretical Artificial Intelligence*, *35*(4), 489–505. https://doi.org/10.1080/0952813X.2021.1952652

Usman, T. M., Saheed, Y. K., Ignace, D., & Nsang, A. (2023). Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. *International Journal of Cognitive Computing in Engineering*, *4*, 78–88. https://doi.org/10.1016/j.ijcce.2023.02.002

Wang, X., Yang, Y., Xu, Y., Chen, Q., Wang, H., & Gao, H. (2020). Predicting hypoglycemic drugs of type 2 diabetes based on weighted rank support vector machine. *Knowledge-Based Systems*, *197*. https://doi.org/10.1016/j.knosys.2020.105868

Yadav, S., Maravi, Y. P. S., Agrawal, J., & Mishra, N. (2021). A Neural Network based Diabetes Prediction on Imbalanced Data. *Proceedings - 2021 IEEE 10th International Conference on Communication Systems and Network Technologies, CSNT 2021*, 515–521. https://doi.org/10.1109/CSNT51715.2021.9509732

Yang, W., Li, C., & Jiang, L. (2023). Learning from crowds with robust support vector machines. *Science China Information Sciences*, *66*(3), 132103. https://doi.org/10.1007/s11432-020-3067-8

Ying, Z., Cao, S., Xu, S., Liu, X., Lyu, L., Chen, C., & Wang, L. (2021). Privacy-preserving optimal insulin dosing decision. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2021-June*, 2640–2644. https://doi.org/10.1109/ICASSP39728.2021.9414807

Zhang, Y. (2012). *Support Vector Machine Classification Algorithm and Its Application* (pp. 179–186). https://doi.org/10.1007/978-3-642-34041-3_27

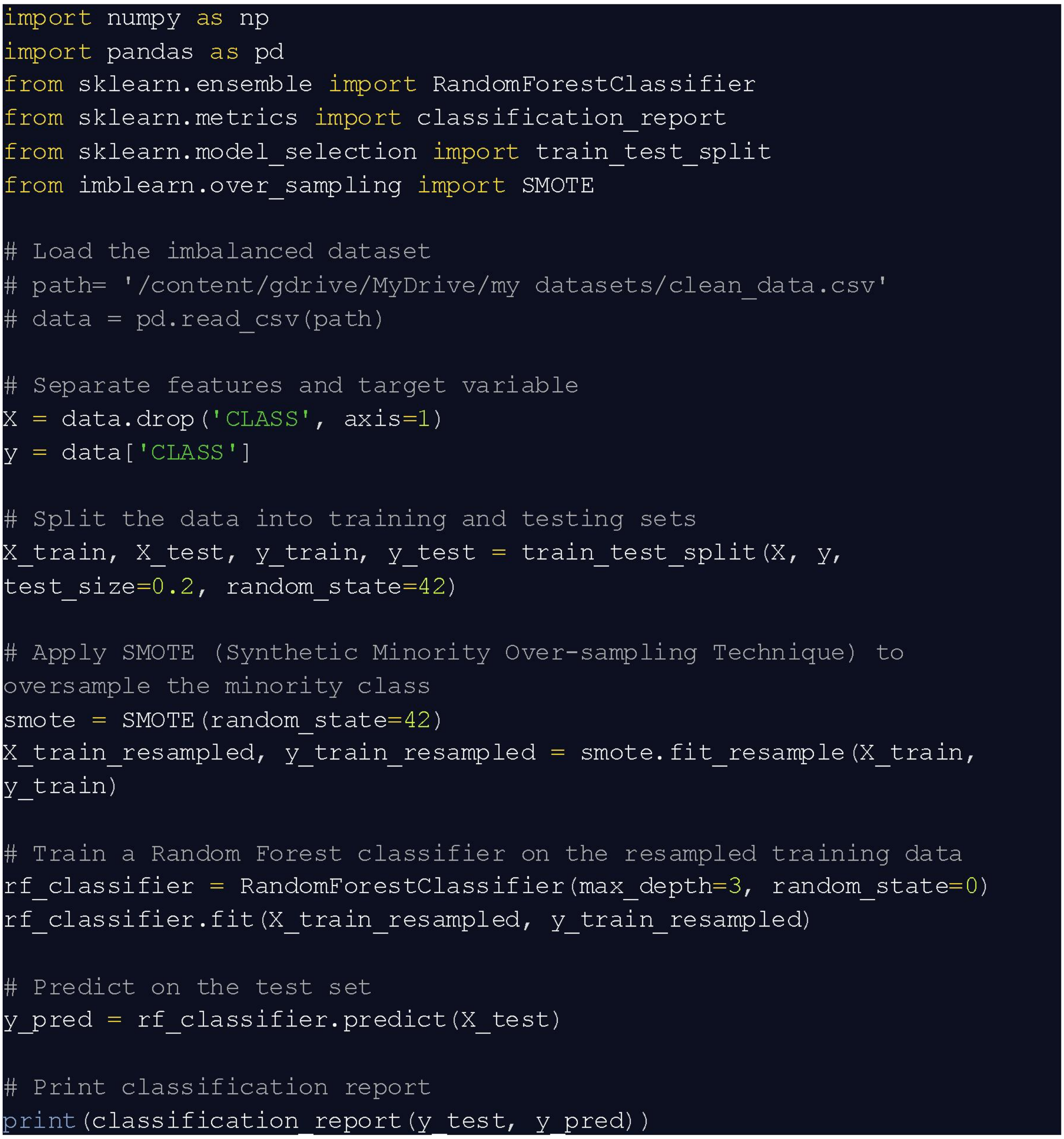
Zhang, Y., Hu, Y., Jiang, N., & Yetisen, A. K. (2023). Wearable artificial intelligence biosensor networks. *Biosensors and Bioelectronics*, *219*. https://doi.org/10.1016/j.bios.2022.114825

Zhou, W., Jiang, H., Cheng, Y., Pei, L., & Ding, S. (2023). Predicting seasonal patterns of energy production: A grey seasonal trend least squares support vector machine. *Expert Systems with Applications*, *213*, 118874. https://doi.org/10.1016/j.eswa.2022.118874

Zhu, T., Kuang, L., Daniels, J., Herrero, P., Li, K., & Georgiou, P. (2023). IoMT-Enabled Real-Time Blood Glucose Prediction With Deep Learning and Edge Computing. *IEEE Internet of Things Journal*, *10*(5), 3706–3719. https://doi.org/10.1109/JIOT.2022.3143375

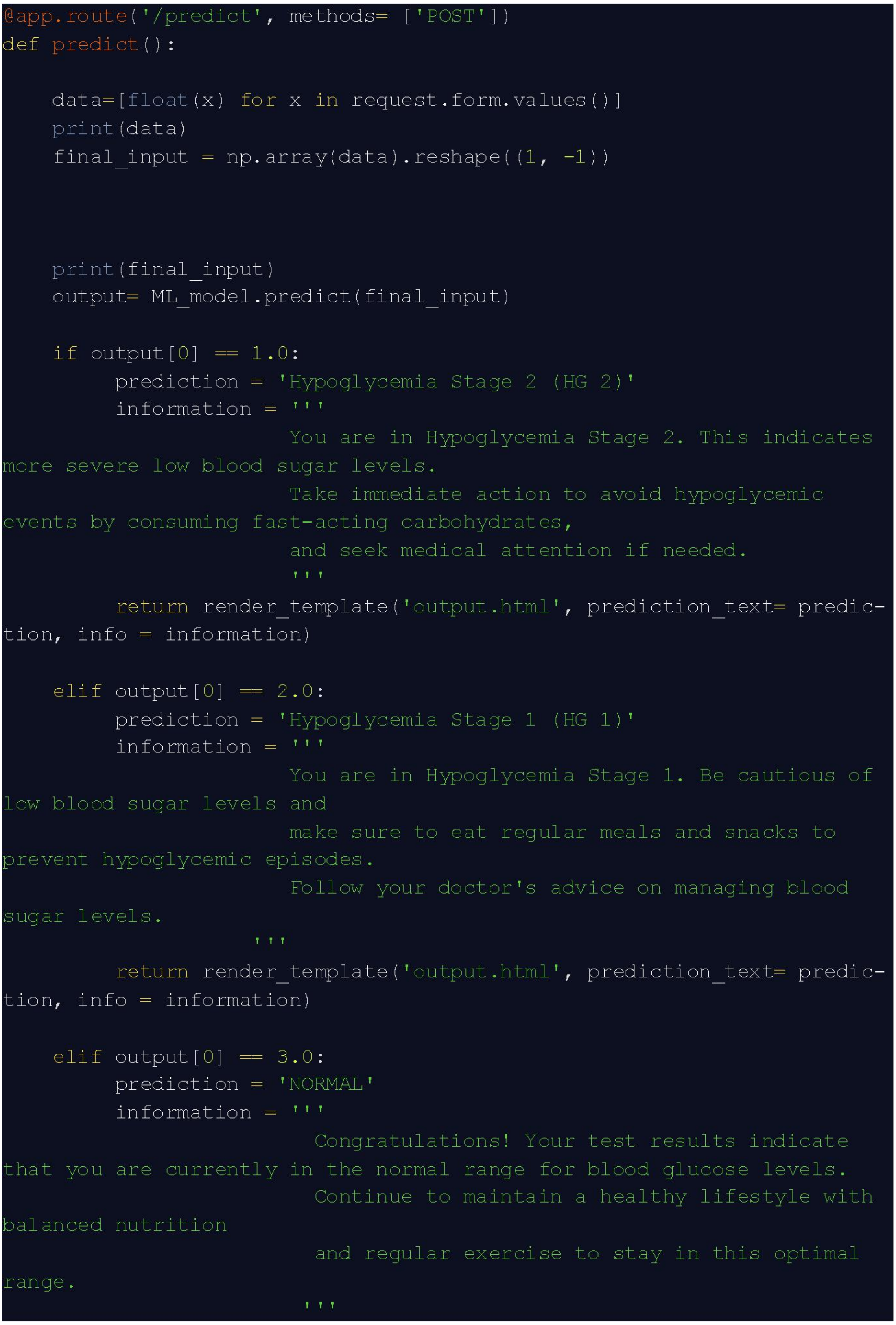
APPENDICES

Appendix A: Code showing how the module were installed during the experiment and model training



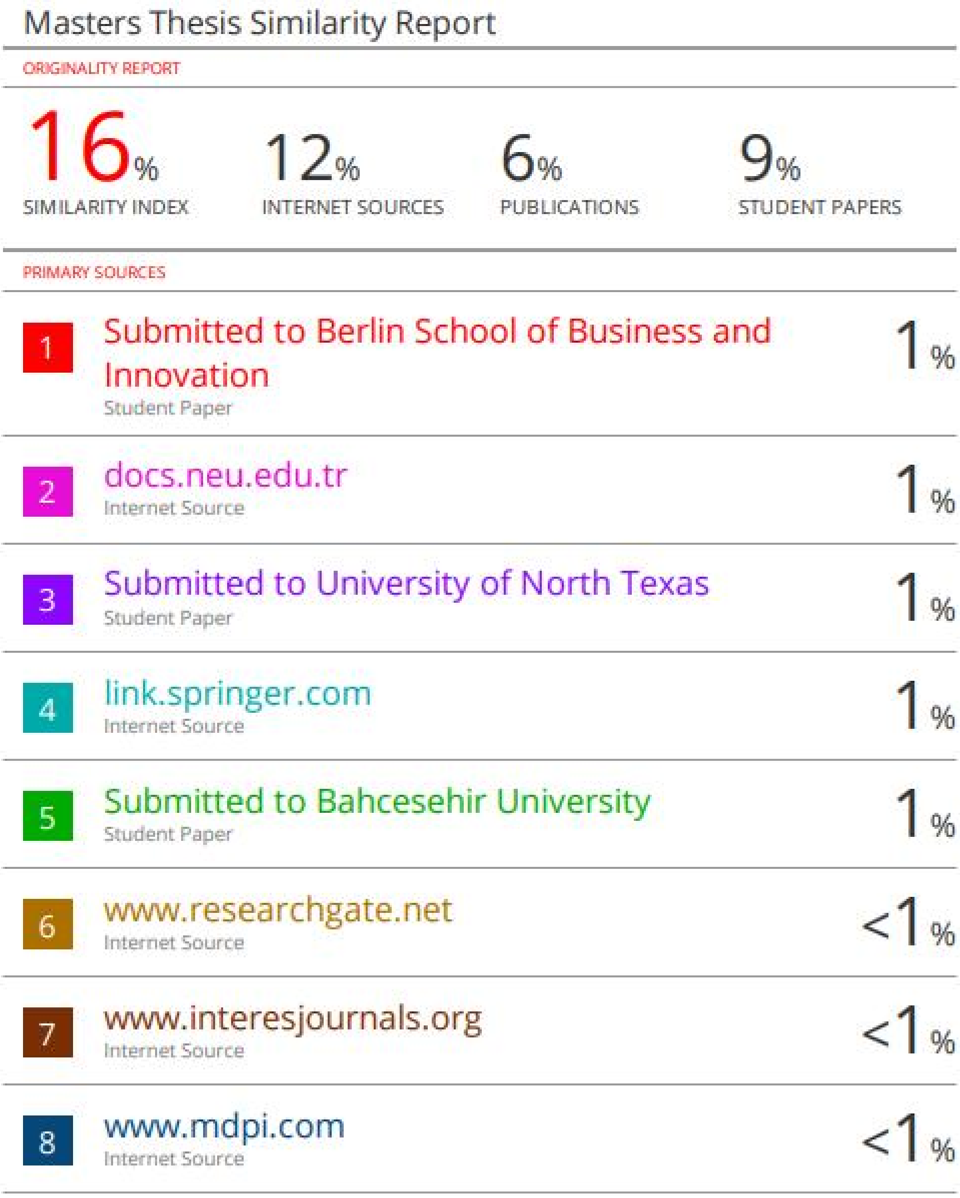
```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE

# Load the imbalanced dataset
# path= '/content/gdrive/MyDrive/my datasets/clean_data.csv'
# data = pd.read_csv(path)

# Separate features and target variable
X = data.drop('CLASS', axis=1)
y = data['CLASS']

# 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)

# Apply SMOTE (Synthetic Minority Over-sampling Technique) to
oversample the minority class
smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train,
y_train)

# Train a Random Forest classifier on the resampled training data
rf_classifier = RandomForestClassifier(max_depth=3, random_state=0)
rf_classifier.fit(X_train_resampled, y_train_resampled)

# Predict on the test set
y_pred = rf_classifier.predict(X_test)

# Print classification report
print(classification_report(y_test, y_pred))
```

Appendix B: **Code for the prediction API, rendering the random forest model as an API for prediction and showing for 5 classes**



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

    data=[float(x) for x in request.form.values()]
    print(data)
    final_input = np.array(data).reshape((1, -1))



    print(final_input)
    output= ML_model.predict(final_input)

    if output[0] == 1.0:
         prediction = 'Hypoglycemia Stage 2 (HG 2)'
         information = '''
                       You are in Hypoglycemia Stage 2. This indicates
more severe low blood sugar levels.
                       Take immediate action to avoid hypoglycemic
events by consuming fast-acting carbohydrates,
                       and seek medical attention if needed.
                       '''
         return render_template('output.html', prediction_text= predic-
tion, info = information)

    elif output[0] == 2.0:
         prediction = 'Hypoglycemia Stage 1 (HG 1)'
         information = '''
                       You are in Hypoglycemia Stage 1. Be cautious of
low blood sugar levels and
                       make sure to eat regular meals and snacks to
prevent hypoglycemic episodes.
                       Follow your doctor's advice on managing blood
sugar levels.
                    '''
         return render_template('output.html', prediction_text= predic-
tion, info = information)

    elif output[0] == 3.0:
         prediction = 'NORMAL'
         information = '''
                         Congratulations! Your test results indicate
that you are currently in the normal range for blood glucose levels.
                         Continue to maintain a healthy lifestyle with
balanced nutrition
                         and regular exercise to stay in this optimal
range.
                        '''
```

Appendix C: Similarity Report



Masters Thesis Similarity Report

ORIGINALITY REPORT

16%	12%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Berlin School of Business and Innovation Student Paper	1%
2	docs.neu.edu.tr Internet Source	1%
3	Submitted to University of North Texas Student Paper	1%
4	link.springer.com Internet Source	1%
5	Submitted to Bahcesehir University Student Paper	1%
6	www.researchgate.net Internet Source	<1%
7	www.interesjournals.org Internet Source	<1%
8	www.mdpi.com Internet Source	<1%