

PREDICTING OF DIABETICS USING MACHINE LEARNING AND DATAMINING

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology degree in Biomedical Engineering

by

YOGAPRIYA K (40240044)

RANCY IRENE JOSE J (40240046)



**DEPARTMENT OF BIOMEDICAL ENGINEERING
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A++" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI-600119

APRIL 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOMEDICAL ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work **YOGAPRIYA K (40240044), RANCY IRENE JOSE J (40240046)** who have done the project work as a team and carried out the project entitled **PREDICTING OF DIABETICS USING MACHINE LEARNING AND DATA MINING** under my supervision from November 2023 to April 2024.

Dr. A. SABARIVANI, M.Tech., Ph.D.,
Internal Guide

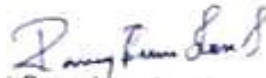
Dr. T. SUDHAKAR, M.Sc., Ph.D.
Head of the Department
Department of Biomedical Engineering
Sathyabama Institute of Science & Technology
(Deemed to be University)
Category I University by UGC
Jeppiyaseethar, Rajiv Gandhi Salai,
Chennai - 600 119

Submitted for Viva Voce Examination held on 5/04/2024

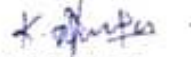
Internal Examiner
External Examiner

DECLARATION

We, YOGAPRIYA K(40240044), RANCY IRENE JOSE J (40240046) hereby declare that the Project Report entitled "PREDICTING OF DIABETICS USING MACHINE LEARNING AND DATA MINING" done by us under the guidance of Dr.A.SABARIVANI,M.Tech.,Ph.D (Internal) at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Biomedical Engineering.



1) Rancy Irene Jose J



2) Yogapriya K

DATE: 5/04/2024

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to Board of Management of Sathyabama Institute of Science and Technology for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks to **Dr.T. SUDHAKAR, M.Sc., Ph.D.**, Head of the Department, Department of Biomedical Engineering for providing me necessary support during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to my Project Guide **Dr.A.SABARIVANI, M.Tech.,Ph.D.** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of Biomedical Engineering who were helpful in many ways for the completion of the project.

We would like to express our thanks to our parents for their unwavering support to us during these years.

ABSTRACT

This research proposes a new system to predict diabetes. The system considers factors like high blood sugar levels, a common sign of this chronic illness. The framework addresses several data quality challenges: it identifies and removes outlier (unusual data points), imputes missing values (fills in any gaps), and partitions the data into training and testing sets. Additionally, it employs feature selection techniques to focus on the most relevant patient characteristics. Finally, the framework leverages various machine learning algorithms, including Naive Bayes, Decision Trees, Random Forest, Support Vector Machine (SVM) and K- nearest Neighbor to identify patterns in the data that are indicative of diabetes. The main goal of the project is to predict analytic in health care based on diagnosing patients with diabetes, kidney, heart disease, and other complex diabetes diseases. This Machine learning (ML) and Data mining technique are analyzed from given data sets. This Study is used to predict the initial stages of diabetes, kidney disease, heart disease, and related diabetes diseases. In previous studies they predicted diabetes only but in this project we will focus on diabetes related disease. Utilizing Python, we intend to employ a genuine data set sourced from the PIMA Indian data set for analysis within our proposed mechanism.

Keywords: Diabetics Prediction, Machine learning, Data mining, Classifications Algorithm, PIMA Indian Data sets.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No.
	ABSTRACT	V
	TABLE OF CONTENTS	VI
	LIST OF FIGURES	VII
	LIST OF TABLES	VIII
	LIST OF ABBREVIATIONS	IX
1	INTRODUCTION	1
	1.1 DIABETICS MELLITUS	1
	1.2 TYPES OF DIABETES	2
	1.2.1 Type 1 Diabetes	2
	1.2.2 Type 2 Diabetes	2
	1.2.3 Gestational Diabetics	3
	1.3 PURPOSE OF THE RESEARCH	3
2	LITERATURE REVIEW	5
	2.1 SURVEY	5
3	AIM AND SCOPE	13
	3.1 AIM	13
	3.2 SCOPE	14
4	DESCRIPTION OF PROPOSED SYSTEM	15
	4.1 MACHINE LEARNING	15
	4.2 PYTHON	15
	4.3 PROPOSED METHODOLOGY	16

	4.4 DATA DESCRIPTION	18
	4.5 DATA PREPROCESSING	20
	4.6 SPLITTING TRAINING DATA AND TESTING DATA	20
	4.7 IMPLEMENTATION OF CLASSIFIER ALGORITHMS	20
	4.7.1 Support Vector Machine	21
	4.7.2 Naïve Bayes Algorithm	21
	4.7.3 Decision Tree	22
	4.7.4 Random Forest	22
	4.7.5 k-Nearest Neighbours(k-Nn)	23
5	RESULT AND DISCUSSION	24
	5.1 RESULTS	24
	5.2 DISCUSSION	28
	5.3 PERFORMANCE ANALYSIS	30
6	SUMMARY AND CONCLUSION	33
	6.1 SUMMARY	33
	6.2 CONCLUSION	34
	REFERENCE	35

LIST OF FIGURES

FIGURE No.	DESCRIPTION	PAGE No.
4.1	BLOCK DIAGRAM OF DIABETES PREDICTION	16
5.1	SUPPORT VECTOR MACHINE	25
5.2	RANDOM FOREST	25
5.3	k-NEAREST NEIGHBOURS(k-NN)	26
5.4	NAÏVE BAYES ALGORITHM	26
5.5	DECISION TREE	27
5.6	ACCURACY VARIATION OVER DATA SET	28
5.7	PRECISION, RECALL AND F1 SCORE GRAPH	31
5.8	REAULTING ACCURACY GRAPH	31

LIST OF TABLES

TABLE No.	DESCRIPTION	PAGE No.
4.1	LIST OF ATTRIBUTES	18
5.1	ACCURACY TABLE	32

LIST OF ABBREVIATIONS

ANN	-	ARTIFICIAL NEURAL NETWORK
BMI	-	BODY MASS INDEX
DM	-	DIABETICS MELLITUS
HBA1C	-	HAEMOGLOBIN A1C
KNN	-	K - NEAREST NEIGHBOR
MG/DL	-	MILLIGRAMS PER DECILITER
PCA	-	PRINCIPAL COMPONENT ANALYSIS
PIDD	-	PRIMARY IMMUNE DEFICIENCY DISEASES
PIDD	-	PIMA INDIAN DIABETIC DATASET
SMO	-	SEQUENTIAL MINIMAL OPTIMIZATION
SVM	-	SUPPORT VECTOR MACHINE

CHAPTER 1

INTRODUCTION

1.1 DIABETICS MELLITUS

Diabetes Mellitus is a mulch-factorial disease caused due to Irregular blood glucose level in our body. Regular complications associated with diabetes such as heart disease, kidney failure, nerve damage, blindness, limb amputation, hearing, and mental health. In Type 1 diabetes (Type-1 DB), beta cells in the pancreas produce insufficient insulin, leading to high blood glucose levels (hyperglycemia). In Type 2 diabetes, the body cannot effectively use the insulin produced. After training a machine learning model, it's essential to evaluate its performance. Common evaluation metrics include accuracy, precision, recall, F1-score, and more, depending on the task (classification, regression, etc.). Cross-validation techniques are often used to assess how well the model generalizes to new, unseen data.

Physicians typically rely on patient test results, comparing them with those of similar cases to make informed decisions. However, this diagnostic process is susceptible to human error as it relies on the doctor's interpretation . Diagnosed cases of diabetes pose a significant challenge, as untreated patients may experience a considerable decline in their quality of life. Relying solely on blood tests for diagnosis is insufficient, as these tests lack the discriminatory power necessary for accurate diagnosis, and their interpretation may vary among populations with different characteristics (Aishwarya Mujumdar et al.,2019). The typical range of glucose levels in the human body is 70 to 99 mg per deciliter. A glucose level exceeding 126 mg/dl indicates diabetes, while 100 to 125 mg/dl suggests per-diabetes. Diabetes has no permanent cure, and it often causes long- term health issues, including macro vascular complications (damage to large blood vessels in the heart, brain, and legs) and micro vascular complications (damage to smallblood vessels affecting the kidneys, eyes, feet, and nerves) (Sahoo et.al.,2022).

In recent years, data mining and machine learning have emerged as dependable and evolving tools in the medical field. Data mining is employed to preprocess health care data and select relevant features, while machine learning automates diabetes prediction, enhancing the capabilities of medical data analysis. Data mining and machine learning algorithms are capable of revealing hidden data patterns through advanced techniques, enabling the generation of reliable and accurate decision.(Sahoo et.al.,2022)

Machine learning systems learn patterns and relationships within data. These systems are trained on a data set, which consists of input features (variables or attributes) and corresponding output labels or targets. The goal is to learn a mapping from inputs to outputs. Machine learning employs various algorithms to build models from data. These algorithms can be grouped into categories like linear regression, decision trees, support vector machines, neural networks, k-means clustering, and more. The choice of algorithm depends on the nature of the problem and the characteristics of the data.

1.2 TYPES OF DIABETICS

1.2.1 Type 1 Diabetes is characterized as an autoimmune disorder in which the immune system fails to generate sufficient insulin. Its causes are not fully understood, and there are no known prevention methods (Ikegami .H et.al.,2022).

1.2.2 Type 2 Diabetes, the most prevalent type, affects approximately 90% of individuals with the condition. It arises from a combination of genetic predisposition and lifestyle factors, leading to inadequate insulin production or improper utilization by the body. (Ikegami .H et.al.,2022).

1.2.3 Gestational Diabetics a form of diabetes that develops during pregnancy in women who previously had normal blood sugar levels. Women with gestational

diabetes have a higher risk of experiencing it again in future pregnancies and are also more likely to develop type 1 or type 2 diabetes later in life. (Ikegami .H et.al.,2022)

1.3 PURPOSE OF RESEARCH

The global prevalence of diabetes among adults (ages 20–79) was 6.4%, impacting 285 million people. It is predicted to rise to 7.7% by 2030, affecting 439 million adults. During this period, there will be a 69% rise in diabetes cases in developing countries and a 20% increase in developed nations. India has the second-highest number of people with diabetes globally, 69.1 million, with 36 million cases undiagnosed. Diabetes was the second leading cause of death in 2020, contributing to 39% of deaths, surpassing COVID-19, and cancer.

This Research discusses the prediction of diabetes using eight independent attributes (e.g., plasma glucose, blood pressure, body mass index) and one dependent attribute (diabetic or non-diabetic). Machine learning and data mining techniques, including SVM, Naïve Bayes, Decision Tree, and SMO, are employed to analyze the data and predict diabetes. The performance of these algorithms is compared to assess their accuracy(Eljil.K *et al.*, 2013).

This research focuses on predicting diabetes using various machine learning and data mining techniques. It utilizes eight independent attributes such as plasma glucose, blood pressure, and body mass index, along with one dependent attribute indicating whether a person is diabetic or non-diabetic.

The mentioned machine learning algorithms - Support Vector Machines (SVM), Naïve Bayes, Decision Tree, and Sequential Minimal Optimization (SMO) - are applied to analyze the data and predict diabetes. These algorithms are chosen for their ability to handle classification tasks effectively.

The performance of each algorithm is evaluated to assess their accuracy in predicting diabetes. This evaluation likely involves metrics such as accuracy, precision, recall, and F1-score. By comparing the performance of these algorithms,

researchers can determine which one provides the most accurate predictions for diabetes diagnosis.

Overall, this research aims to leverage machine learning techniques to improve the early detection and prediction of diabetes, which is crucial given its significant global impact and the projected increase in prevalence over the coming years. Common symptoms of diabetes include increased thirst, frequent urination, unexplained weight loss, fatigue, blurred vision, and slow wound healing. However, some people with diabetes may not experience any symptoms, especially in the early stages of the disease(Devi Wulandari *et al.*, 2020).

Managing diabetes involves maintaining blood sugar levels within a target range through a combination of lifestyle modifications, such as healthy eating, regular physical activity, and weight management, along with medication (such as insulin or oral medications) and regular monitoring of blood sugar levels. Complications of uncontrolled diabetes can include cardiovascular disease, kidney disease, nerve damage, eye problems, and foot problems leading to amputation. Therefore, early diagnosis, treatment, and ongoing management are essential for preventing complications and improving quality of life for people with diabetes(Zou Q *et al.*,2018).

CHAPTER 2

LITERATURE REVIEW

2.1 SURVEY

Ageed Z.S. *et al.*, (2021) delve into the synergy between cloud computing, data mining, and big online data, emphasizing their impact on various aspects such as data storage, communication, and visualization. The paper highlights the significance of big data in driving industry growth and scientific advancements, despite challenges in storage and computation. It distinguishes big data mining from traditional approaches and explores its implications, including personalized services and security measures. Overall, the study underscores the value of abundant data in uncovering meaningful insights.

Aishwarya Mujumdara *et al.*,(2020) conducted a study testing machine learning algorithms for diabetes prediction, where Logistic Regression achieved 96% accuracy. By employing a pipeline approach, the Ada Boost classifier outperformed others with 98.8% accuracy. Comparison between datasets highlighted improved prediction accuracy with the newer data set. The study suggests potential expansion to predict diabetes in non-diabetic individuals. Big data analytic in health care facilitates informed predictions by uncovering hidden patterns. Their proposed model integrates external factors, enhancing classification accuracy, while a tailored pipeline model further improves accuracy in diabetes prediction.

Amelec Viloriaa *et al.*, (2020) proposed using a support vector machine (SVM) to diagnose diabetes mellitus (DM) based on age, BMI, and blood glucose levels. Their SVM achieved 99.2% accuracy with Colombian patients but lower accuracy (65.6%) with patients from different ethnic backgrounds. Data from 500 patients from a Colombian hospital was used, and a nonlinear SVM classifier was trained and validated using radial-based kernels. The classifier shows promise for managing new DM cases in Colombia and potentially universally, but further data on DM diagnosis is necessary for refinement.

Bano Farhana et al., (2021) identified key parameters crucial for predicting Diabetes Mellitus, including Glucose, Pregnancies, Skin Thickness, Blood Pressure, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. They employed several machine learning algorithms such as SVM, ANN, Decision Tree, Logistic Regression, and Farthest First to assess prediction accuracy. Remarkably, the Farthest First algorithm demonstrated superior performance, achieving an accuracy rate nearing 99.9%. This research highlights the pivotal role of machine learning in facilitating early diabetes prediction, enabling proactive measures to mitigate associated risks. The insights derived from this study can significantly benefit health care professionals in improving the early detection of diabetes.

Bhavana et al., (2018) proposed ensemble learning as a means to enhance prediction accuracy in various diseases, including diabetes, skin ailments, kidney disorders, liver conditions, and cardiac diseases. They evaluated ensemble paradigms such as bagging, boosting, stacking, and voting, finding stacking to be the most effective in 19 out of 23 usages. Bagging performed well in kidney disease prediction, while boosting excelled in liver ailments and diabetes. Heart disease received extensive research attention. Preprocessing techniques like test-train division, feature selection, and PCA were employed to refine datasets. Overall, the study underscores the efficacy of stacking and the importance of judicious ensemble selection in predictive medical analytics, serving as a guide for scholars in the field.

Chollette C et al., (2022) proposed a machine learning framework to enhance diabetes prediction using feature selection and missing value handling techniques. They utilized the PIMA Indian and LMCH diabetes datasets, incorporating Spearman correlation for feature selection and polynomial regression for missing value imputation. Supervised learning models like random forest, support vector machine, and a novel twice-growth deep neural network were optimized through hyper parameter tuning and cross-validation. This approach outperformed existing benchmarks in diabetes prediction, with the source code available.

Devi Wulandari et al., (2020) focused on predicting glucose levels in diabetics, comparing various algorithms. The support vector regression algorithm exhibited superior performance, with the lowest mean squared error of 28.9480. Diabetes arises from insulin resistance, affecting glucose regulation by pancreatic hormones. Its complications include kidney dysfunction, vision impairment, and cardiovascular issues, leading to severe conditions like blindness, heart disease, and stroke. Notably, the Support Vector Machine Algorithm outperformed Neural Network, Random Forest, and K-Nearest Neighbors (KNN) algorithms in predicting glucose levels.

De, L & Silva et al., (2016) proposed using data mining for early detection and treatment initiation in diabetes. They employed Decision Tree, Naïve Bayes, and Support Vector Machine methods, achieving accuracy of 84.6667%, 76.6667%, and 77.3333% respectively. An ensemble method combining these algorithms improved accuracy by countering mis classifications. The study acknowledges regional variances in results due to genetics and lifestyle. A practical example showcased the automated system's efficacy in diabetes diagnosis, ensuring accuracy through consensus among algorithms..

Ejiwale et al., (2021) utilized Support Vector Machine (SVM) to predict comorbid gestational diabetes and hypertensive disorders in pregnancy. Through feature selection, SVM achieved a perfect recall score of 100% using nine predictors, including pre-pregnancy BMI, educational status, occupation, blood group, and age range. Their approach aids clinical decision-making, particularly for nulliparous women, and improves maternal and child care outcomes. Comparative analysis with other machine learning algorithms highlighted the clinical utility of automated tools in prenatal screening. The introduced GDHP model enhances pregnancy outcomes, driving targeted care plans and reducing costs.

Fareeha Anwar et al., (2020) proposed diabetes diagnostic methods using AI techniques like neural networks and machine learning. They favored the Pima Indian

Diabetes Dataset (PIDD) for its relevance. A study using Back Propagation Neural Network (BPNN) for diabetes prediction with PIDD was cited. Data normalization and ANN training resulted in minimal errors by the third iteration. The paper suggests combining deep learning with other algorithms for improved accuracy and calls for more public diabetes data sets. It also advocates for using Heart Rate (HR) data for diagnosing cardiac diseases due to its computational efficiency.

Gomathi *et al.*, (2016) proposed the use of data mining techniques in predicting diseases, aiming to reduce the number of required tests for detection. The paper reviewed research focused on predicting heart disease, diabetes, breast cancer, and other diseases. It compared the performance of two different data mining classification techniques to evaluate the best classifier for disease prediction. The study highlights the challenge of building precise and computationally efficient classifiers for medical applications in data mining and machine learning.

Han Wu *et al.*, (2017) proposed a novel prediction model for Type 2 Diabetes Mellitus (T2DM) using data mining techniques. Their model integrated an enhanced K-means algorithm with logistic regression to improve prediction accuracy and dataset adaptability. Tested on the Pima Indians Diabetes Data set using the WEKA toolkit, their model achieved a 3.04% increase in accuracy compared to previous research. It was also validated on two additional diabetes data sets, confirming its reliability and adaptability.

Ikegami *et al.*, (2022) proposed innovative solutions for older adults with insulin-dependent diabetes, addressing challenges in self-managing daily insulin injections due to cognitive and physical limitations. They highlight recent advancements like once-weekly insulin doses and encapsulated islet cell implantation as promising avenues for consistent insulin delivery. Additionally, beta-cell replacement therapies offer potential benefits for older patients struggling with traditional insulin administration, aiming to improve overall management of insulin-dependent diabetes.

Larabi-Marie-Sainte et al.,(2019) explored machine learning and deep learning techniques for predicting diabetes, evaluating various classifiers like Decision Trees, Support Vector Machines, Naive Bayes, and more. Decision tree algorithms exhibited the highest accuracy, ranging from 68% to 95%. The study suggests combining different classifiers and models to enhance prediction accuracy and advocates for further research using rarely used classifiers across different data sets to improve diabetes prediction.

Mitushi Soni et al., (2020) employed machine learning classification and ensemble techniques to create accurate predictive models based on patient data for diabetes prediction. Techniques included K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF), with Random Forest exhibiting the highest accuracy. The project achieved a 77% classification accuracy, emphasizing its potential to assist healthcare professionals in early diabetes prediction and treatment decisions, ultimately saving lives.

Nadeem, M. We et al., (2021) proposed a fusion machine learning approach for diabetes identification and critical event prediction. They combined Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifiers, achieving a classification accuracy of 94.67%, surpassing previous models by 1.8%. Using NHANES and Kaggle's "Pima Indian diabetes" data sets, they demonstrated improved accuracy through fusion, with the combined SVM-ANN model outperforming standalone SVM and ANN algorithms.

Obeda Jamal Khanam et al., (2021) focused on early diabetes detection using machine learning algorithms and neural network methods with the Pima Indian Diabetes (PID) data set. Their study applied seven ML algorithms, achieving accuracy above 70%. The neural network model with two hidden layers and 400 epochs

demonstrated the highest accuracy at 88.6%. The research presented confusion matrices, ROC curves, and performance measures, emphasizing the effectiveness of neural network models in diabetes prediction.

Sahoo, Sipra et al., (2022) All six machine learning models mentioned in the paper are implemented using Python programming and Python and Keras APIs. The paper provides hardware specifications for the machine used in the study. The evaluation of these models considers both accuracy and F1-Score, with a particular emphasis on F1-Score, which is considered more relevant in real-life scenarios due to the presence of imbalanced classes. The study calculates and presents values for precision, recall, accuracy, and F1-score for all the algorithms in atabular format, with results displayed up to five decimal points of accuracy.

Rashi Rastogi et al., (2020) developed a diabetes prediction model using data mining techniques like Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). Analyzing a Kaggle data set in Python, they evaluated the model's performance with metrics like accuracy, confusion matrix, and sensitivity. Logistic Regression exhibited the highest accuracy at 82.46%, surpassing SVM at 79.22%. The study aims to enhance diabetes outcome predictions by exploring more classification algorithms in the future.

Sisodia et al., (2018) proposed early-stage diabetes detection using machine learning algorithms, evaluating precision, accuracy, F-measure, and recall. Naive Bayes exhibited superior performance, achieving an accuracy of 76.30%, validated with Receiver Operating Characteristic (ROC) curves on the Pima Indians Diabetes Database. The study suggests potential for diagnosing other diseases and automating diabetes analysis by integrating additional machine learning algorithms. Naive Bayes emerged as the most accurate classifier in this study.

Sunge A. S et al., (2019) compared five predictive models - Naive Bayes, Decision Tree, SVM, K-NN, and ANN - to predict diabetes mellitus using nine input variables and one output variable from a specific data set. Their research highlights the application of data mining algorithms in creating accurate prediction models for diabetes mellitus. Through comparative analysis, they found that the K-NN model demonstrated the highest accuracy, closely followed by the ANN algorithm.

Sun, Y. and Zhang, D. et al., (2019) explored the role of data mining in amplifying machine learning's impact on diabetes research, facilitating risk factor identification and aiding diagnosis and treatment. The paper delved into the application of traditional machine learning and deep learning techniques in diabetes data analysis. While traditional models are effective, they face challenges with data set specificity and high-dimensional data. The study advocates for a hybrid approach merging deep learning's adaptability with traditional machine learning's clarity to optimize diabetes diagnosis. Collaboration between biologists and computational experts is emphasized to harness the vast potential of biological data.

Woo Seok Lee et al., (2020) proposed a machine learning approach to detect insulin resistance from blood glucose patterns, crucial for early diabetes diagnosis. Using synthetic glucose data generated by a biophysical model, machine learning methods achieved accuracy exceeding 85% in identifying insulin resistance patterns. Despite the lack of real-time high-resolution data, the study highlights the transformative potential of machine learning in advancing diabetes management by detecting subtle glucose changes indicative of insulin resistance.

Vinaytosh Mishra et al., (2015) employed logistic regression to predict diseases in their early stages using data from diabetes specialty clinics in Varanasi. Their study challenges the significance of HBA1C readings as a crucial predictor for diabetes, highlighting the need for broader studies across multiple centers. They suggest incorporating more variables through methodologies like the Delphi Method to enhance predictive accuracy. While logistic regression was the primary tool, future

research could compare its results with advanced techniques like Artificial Neural Networks (ANN) to determine the most effective predictive methodology.

Zou Q *et al.*, (2018) conducted an analysis of hospital examination data from Luzhou, China, employing machine learning techniques to predict diabetes. Using five-fold cross-validation on a data set containing 14 attributes, including fasting glucose, decision tree, Random Forest (RF), and neural networks were utilized. Results showed fasting glucose as a significant predictor, particularly within the Luzhou data set. However, relying solely on fasting glucose didn't yield optimal results. The best accuracy achieved was 0.8084 for the Luzhou data set and 0.7721 for the Pima Indians data set, suggesting machine learning's effectiveness in diabetes prediction. Future research aims to predict specific diabetes types and explore attribute importance for enhanced accuracy.

CHAPTER 3

AIM AND SCOPE

3.1 AIM

The primary objective of this project is to develop a highly accurate and robust predictive model for diabetes, encompassing early detection, risk assessment, and the prediction of associated complications like kidney disease and heart disease. By leveraging machine learning and data mining techniques, the project aims to significantly enhance healthcare analytic and contribute to the overall well-being of individuals at risk of diabetes and its related conditions. Secondly, by facilitating early detection and preventive measures, the project could lead to substantial cost reductions in health care expenditure, alleviating the economic burden associated with diabetes management and treatment. Moreover, addressing the global diabetes epidemic through proactive prediction and prevention strategies could have far-reaching implications for public health, helping to mitigate the prevalence and severity of the disease on a global scale. Ultimately, the project's overarching goal is to harness the power of data-driven insights to empower health care providers, policymakers, and individuals alike in the fight against diabetes and its complications, fostering healthier communities and a more sustainable health care system. The primary aim of this project is to develop an accurate and robust predictive model for diabetes. This includes the early detection of diabetes, risk assessment, and the prediction of related complications such as kidney disease and heart disease. The project aims to contribute to improved health care analytic and the overall well-being of individuals at risk of diabetes and its associated conditions.

3.2 SCOPE

With our data prepared and split, we proceeded to employ a variety of classifier algorithms, including Naive Bayes, SVM, Random Forest, KNN, and Decision Forest. For each algorithm, specific parameters were assigned, and GridSearchCV was utilized to fine-tune these parameters. GridSearchCV, a technique for identifying optimal parameter values from a predefined set, incorporates cross-validation to ensure robustness and generalizability of the model. The selection of the final model(s) for predicting diabetes status was guided by a combination of performance metrics, computational efficiency, and interpretability. While Random Forest emerged as the top-performing algorithm in terms of F1 score, considerations such as model complexity and ease of interpretation favored SVM as a viable alternative. Thus, a combination of Random Forest and SVM models may offer a balanced approach to diabetes prediction, leveraging the strengths of both algorithms. It is essential to acknowledge the limitations of our study, including the inherent biases and assumptions present in the data set. The generalization of our findings to broader populations and clinical settings warrants further investigation, as does the exploration of alternative data sets and modeling approaches to validate our results.

CHAPTER 4

MATERIALS AND METHODS USED

4.1. MACHINE LEARNING

Machine learning, a subset of artificial intelligence, empowers computer systems to enhance task performance by learning from data without explicit programming. It entails discerning patterns and relationships within data sets, where models are trained on input features and corresponding output labels or targets. Supervised learning utilizes labeled data to make predictions or classifications, while unsupervised learning uncovers patterns in unlabeled data (Sisodia *et al.*, 2014) . Reinforcement learning trains agents to make decisions aimed at maximizing rewards. These methodologies enable machines to learn from experience or examples, facilitating data-driven decision-making.

Machine learning employs diverse algorithms like linear regression, decision trees, support vector machines, neural networks, and k-means clustering to construct models from data. The selection of algorithm hinges on the problem's nature and data characteristics. Post-training, evaluating model performance becomes pivotal. This entails using metrics such as accuracy, precision, recall, and F1-score, tailored to the task, whether classification, regression, etc. Cross-validation techniques are frequently employed to gauge the model's generalizability to unseen data, ensuring robustness and reliability in real-world applications (Krishnamoorthi *et al.*, 2022).

4.2 PYTHON

Python is widely used in machine learning due to its simplicity, versatility, and robust ecosystem of libraries. Libraries like TensorFlow, PyTorch, and scikit-learn provide powerful tools for implementing various machine learning algorithms. Python's readability and ease of use make it accessible for beginners and experts alike, enabling rapid prototyping and experimentation (Sahoo *et al.*, 2022). Its extensive community support and vast array of resources make it a top choice for developing

machine learning models and applications

Python abstracts many low-level details, which makes it easy to write and understand code. This is beneficial for both beginners and experienced programmers.

Community and Ecosystem: Python has a large and active community of developers and users. This community has contributed to the creation of numerous libraries and frameworks, expanding Python's capabilities. Popular libraries include NumPy (for numerical computing), pandas (for data analysis), Django (for web development), and TensorFlow (for machine learning).

4.3 PROPOSED METHODOLOGY

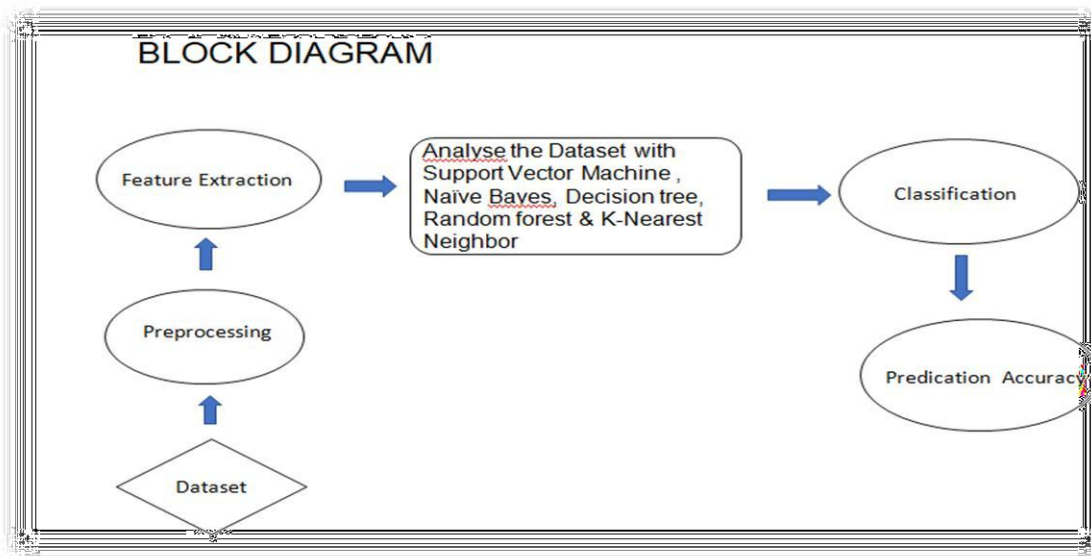


Fig:4.1: Block Diagram of diabetes prediction

The proposed methodology for diabetes dataset analysis involves several steps. Initially, libraries are loaded, and the dataset is imported. Pre-processing follows to handle missing values, ensuring data integrity. Then, an 80/20 split divides the data set into training and testing sets. Feature scaling is applied for equal feature contribution. Various algorithms

(SVM, DT, RF, NB, KNN) are considered, each with unique strengths and weaknesses. Chosen algorithms are trained on the training set to learn patterns and relationships, then evaluated on the test set. A comparative analysis examines metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Based on this analysis, the best-performing algorithm is determined, considering factors like accuracy and inter predictability. This systematic approach aims to identify the most effective model for predicting diabetes outcomes(Krishnamoorthi *et al.*,2022).

In Type 2 diabetes, the body cannot effectively use the insulin produced.After training a machine learning model, it's essential to evaluate its performance. Common evaluation metrics include accuracy, precision, recall, F1-score, and more, depending on the task (classification, regression, etc.). Cross-validation techniques are often used to assess how well the model generalizes to new, unseen data.

Physicians typically rely on patient test results, comparing them with those of similar cases to make informed decisions. However, this diagnostic process is susceptible to human error as it relies on the doctor's interpretation. Diagnosed cases of diabetes pose a significant challenge, as untreated patients may experience a considerable decline in their quality of life. Relying solely on blood tests for diagnosis is insufficient, as these tests lack the discriminatory power necessary for accurate diagnosis, and their interpretation may vary among populations with different characteristics .The typical range of glucose levels in the human body is 70 to 99 mgper deciliter. A glucose level exceeding 126 mg/dl indicates diabetes, while 100 to 125 mg/dl suggests per-diabetes. Diabetes has no permanent cure, and it often causes long- term health issues, including macro vascular complications (damage to large blood vessels in the heart, brain, and legs) and micro vascular complications (damage to small blood vessels affecting the kidneys, eyes, feet, and nerves (chollete *et al.*,2022).

4.4 DATASET DESCRIPTION

This section consists of data overview, data cleaning, the classification algorithm and analyzing accuracy results.

PIMA Indian diabetic data set contains a data set record on 1,00,000 patients. Positive and negative test results instances presence or absence of diabetes. Individual case consists of 8 features, Composed of numeric data. This data contains Individual medical records and diagnostic data. Predicting diabetes using machine learning typically involves analyzing a data set with various attributes (features) to build a model that can classify individuals as diabetic or non-diabetic. (Chollette C. *et.al.*,2022).

Diabetes, kidney, heart disease, and other complex diabetes diseases. This Machine learning (ML) and Data mining technique are analyzed from given data sets. This Study is used to predict the initial stages of diabetes, kidney disease, heart disease, and related diabetes diseases. In previous studies they predicted diabetes only but in this project we will focus on diabetes related disease. Utilizing Python, we intend to employ a genuine data set sourced from the PIMA Indian data set for analysis within our proposed mechanism(R. M. Rahman *et al.*, 2015).

Table No: 4.1: List of attributes

S.No	ATTRIBUTES	DESCRIPTION
1	Gender	It refers to the characteristics, roles, behaviors, and attributes that are traditionally associated with a particular gender, such as male or female.
2	Age	It refers to a qualitative description of age group 2 to 81.

3	Hypertension	It is characterized by elevated pressure in the arteries, which can lead to increased risk of cardiovascular diseases.
4	Heart disease	This data set contains information for predicting the presence or absence of heart disease .1- heart disease 0- not heart disease.
5	Smoking history	Smoking history typically refers to an individual's past or current tobacco use, including details such as never used ,no information, former used
6	BMI	Body Mass Index (BMI) is a measurement used to assess weight status in adults. It's calculated from aperson's weight and height.
7	HbA1c level	HbA1c level States to the measurement of glycated hemoglobin in the blood, often used as a marker for long-term glucose control in individuals with diabetes
8	Plasma glucose level	The amount of sugar (glucose) in your blood at a particular moment is called your blood sugar level
9	Diabetes	Diabetes-1 , non diabetes-0

4.5 DATA PREPROCESSING

To prepare the data for model building, several crucial steps are undertaken to optimize the data set for accurate training. Initially, the data set is imported to begin the process. Subsequently, relevant attributes are averaged to ensure consistency and reduce noise in the data. Following this, outlier removal, data normalization, and feature selection techniques are applied to enhance the quality of the data set.

Once the data is processed, it is split into training and testing subsets, typically using techniques such as an 80/20 split, ensuring that the model's performance can be adequately evaluated. Finally, to facilitate further analysis and modeling, the features in the data set are scaled, ensuring that they contribute equally and preventing biases due to differences in scale. These steps collectively prepare the data set for effective model building, laying the foundation for accurate predictions and insightful analysis(H. Kaur.*et al.*,2020).

4.6 SPLITTING TRAINING DATA AND TESTING DATA

After completing data cleaning and preprocessing, the data set is prepared for training and testing machine learning models. Two different methods, K-fold cross-validation and an 85% train/test split, are employed to evaluate the performance of various machine learning models. In the train/test split approach, the data set is randomly divided into training (75 %)and testing sets (25%) to assess the models' effectiveness(Ejiwale *et al.*,2021).

4.7 IMPLEMENTATION OF CLASSIFIER ALGORITHMS

In this research, an extensive investigation is conducted on Predictive Identification of Diabetes (PIDD) using various machine learning classification techniques, including Decision Trees (DT), Naive Bayes (NB),SOM and Support Vector Machines (SVM). These techniques are employed to analyze and classify data related to PIDD, aiming to identify and predict diabetes more accurately(R. M. Rahman *et al.*, 2015).

4.7.1 Support Vector Machine

Support Vector Machine (SVM) is a kernelized machine learning approach that requires the transformation of text. Data preprocessing is crucial for enhancing machine learning algorithms by enhancing data integrity and precision. It fulfills numerous roles, such as anomaly detection, data standardization, and characteristic extraction, to refine the data set for model construction (Woo Seok Lee *et al.*, 2020).

4.7.2 Naïve Bayes Algorithm

Naïve Bayes is a probabilistic classifier for machine learning and finds application in various classification tasks. It's often used to predict the probability of an event belonging to a particular category depending on features or attributes. Naive Bayes is a fast and efficient classification method because it assumes features work independently to predict a class. (Obeda Jamal Khanam.*et.al.*,2019). Despite this simplification, NB often exceeds the advanced model in Real-world uses. It is particularly effective when working with data sets exhibiting imbalance and missing values, making it a versatile tool for classification tasks (Sahoo *et al.*, 2022).

Text classification encompasses various tasks such as spam email detection, sentiment analysis, and document categorization, all aimed at automatically assigning predefined categories or labels to text data. In the realm of medical diagnosis, machine learning algorithms can be employed to predict diseases based on symptoms and patient data, aiding health care professionals in making informed decisions(Ejiwale *et al.*,2021). Recommendation systems utilize algorithms to analyze user behavior and preferences, enabling personalized product recommendations across various platforms. Customer segmentation involves grouping customers based on their purchase history, demographics, or other relevant data, allowing businesses to tailor their marketing strategies and services to specific customer segments

effectively. These applications demonstrate the versatility and effectiveness of machine learning techniques in solving diverse problems across different domains (R. M. Rahman *et al.*, 2015).

4.7.3 Decision Tree

Decision Tree are a basic classification method. It is a supervised learning method. Decision trees are well-suited for categorical response variables. The decision tree has three structure-based models that describe the classification process based on input features. This model can handle various input data types, including graphs, text, discrete values, and continuous measurements (Larabi-Marie-Sainte.S *et.al.*,2019).

4.7.4 Random Forest Algorithm

Random Forest is an effective classification approach that Capitalizes on the advantages of decision trees. It works by creating a multitude of decision trees during training, each utilizing a Unpredictable subset of attributes and data instances. During prediction, each tree independently votes for a class label for a new data point. The final class label is assigned based on the majority vote from all the trees in the forest. This approach helps to reduce the risk of over fitting and often leads to more accurate predictions compared to a single decision tree(H. Kaur.*et al.*,2020).

Random Forest, an ensemble learning algorithm, excels in various machine learning tasks due to its consistent high accuracy, robustness against over fitting, and ability to handle large data sets with numerous features efficiently. By constructing multiple decision trees and aggregating their predictions, it captures complex relationships within the data. The randomness introduced during training mitigates over fitting by diversifying individual trees and reducing noise learning. Moreover, its capacity to handle large data sets without extensive preprocessing saves time and effort (R. M. Rahman *et al.*, 2015). Additionally, Random Forest provides insights into feature

importance, aiding in understanding data dynamics and guiding feature selection or engineering efforts. In summary, it is a robust and versatile algorithm, ideal for tasks prioritizing accuracy, robustness, and interpretability (R. M. Rahman *et al.*, 2015).

4.7.5. K-Nearest Neighbor Algorithm

The K-Nearest Neighbors (KNN) algorithm is a straightforward and versatile tool for classification tasks. Unlike some methods, KNN doesn't rely on assumptions about the underlying data distribution (making it non-parametric). During training, it considers available attributes (features) to understand the relationships within the data. When presented with a new, unknown data point, KNN identifies its k closest neighbors based on a distance metric (often a tree-like data structure is used to efficiently find these neighbors (K. Kincade *et al.*, 2006)). Finally, the algorithm classifies the new data point by analyzing the category (for classification) or average value (for regression) of its k nearest neighbors.

k-Nearest Neighbors (k-NN) stands out as an accessible algorithm for machine learning tasks, particularly for beginners, due to its simplicity in implementation and comprehension. Unlike many other algorithms, k-NN does not necessitate a training phase on the entire data set, as it operates as a lazy learner. Instead, predictions are made directly based on stored instances, offering a streamlined approach to inference. Another advantage lies in its non-parametric nature, as k-NN makes no assumptions about the underlying data distribution. This characteristic enhances its versatility, allowing it to be applied to various types of data without requiring adjustments or preprocessing based on distribution assumptions. Overall, k-NN emerges as a user-friendly and flexible algorithm, suitable for diverse applications where simplicity and adaptability are valued (R. M. Rahman *et al.*, 2015).

CHAPTER 5

RESULTS AND DISCUSSION

5.1 RESULTS

In the initial phase of our project, we began by laying the foundation for data processing and analysis. This involved importing essential libraries such as pandas for data manipulation, matplotlib for visualization, and sklearn for machine learning algorithms. Additionally, we imported the data set into our working environment to start exploring its structure and contents. Subsequently, we focused on preprocessing the data to ensure it was in a suitable format for analysis. Since our data set contained categorical variables like gender and smoking status, we needed to handle them appropriately. To do this, we employed encoding techniques to convert these categorical strings into numerical values, facilitating their integration into our machine learning models. A confusion matrix for each algorithm are shown.

Moreover, to ensure consistency and comparability across our data set, we applied standardization using scalar techniques. This step involved scaling our features to have a mean of 0 and a standard deviation of 1, which is particularly important for algorithms sensitive to feature magnitudes, such as SVM shown in Fig 5.1 and KNN.. Following preprocessing, we partitioned our data set into separate training and testing subsets. This division, with 75% of the data allocated for training and 25% for testing, allowed us to train our models on a subset of the data and evaluate their performance on unseen data. A confusion matrix for KNN shown in Fig 5.2. A confusion matrix for random forest shown in Fig 5.3 and Fig 5.4 and Fig 5.5 for NB and Decision tree.

Our ultimate objective was to predict whether individuals have diabetes or not based on the available features in the data set. To assess the performance of each algorithm, we calculated key metrics such as F1 score, precision, and accuracy. These metrics provided valuable insights into the effectiveness of each algorithm in

predicting diabetes status, guiding our decision-making process and model selection. Overall, this systematic approach to preprocessing and modeling laid a solid foundation for our subsequent analysis and interpretation of results. An Accuracy variation over a data set for algorithm shown in Fig 5.6.

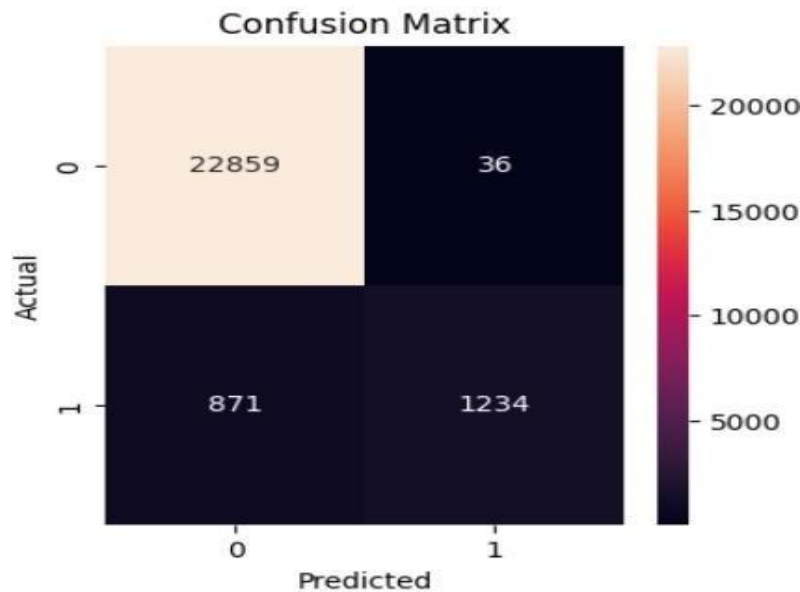


Fig:5.1:Support Vector Machine

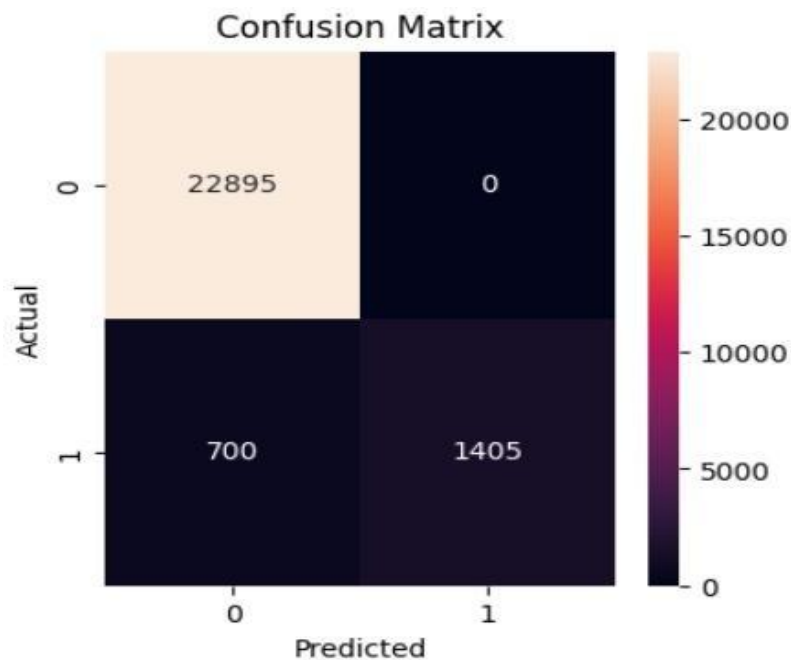


Fig:5.2: Random forest

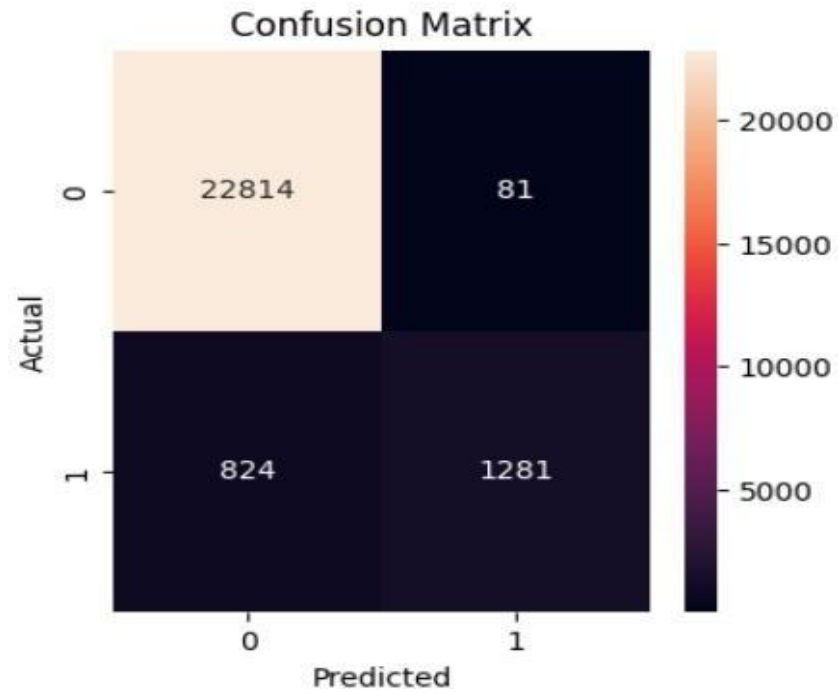


Fig:5.3: k-Nearest neighbor

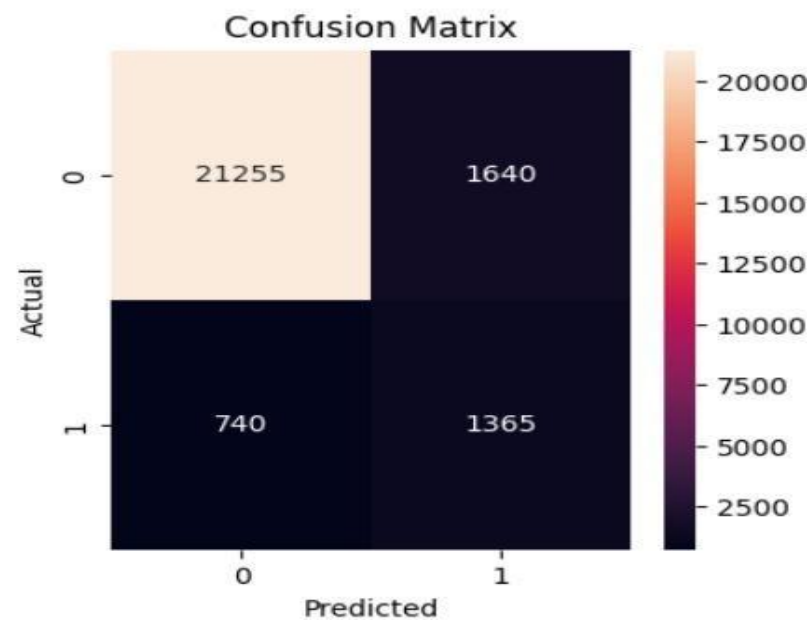


Fig:5.4: Naïve bayes

A confusion matrix is a performance measurement tool for machine learning classification models for support vector machine shown in Fig 5.1. It's particularly useful for summarizing the performance of a classification algorithm. The matrix itself is a table with two dimensions, typically representing predicted and actual classes. A confusion matrix for KNN shown in Fig 5.2. A confusion matrix for random forest shown in Fig 5.3. Fig 5.4 and Fig 5.5 for NB and Decision tree.

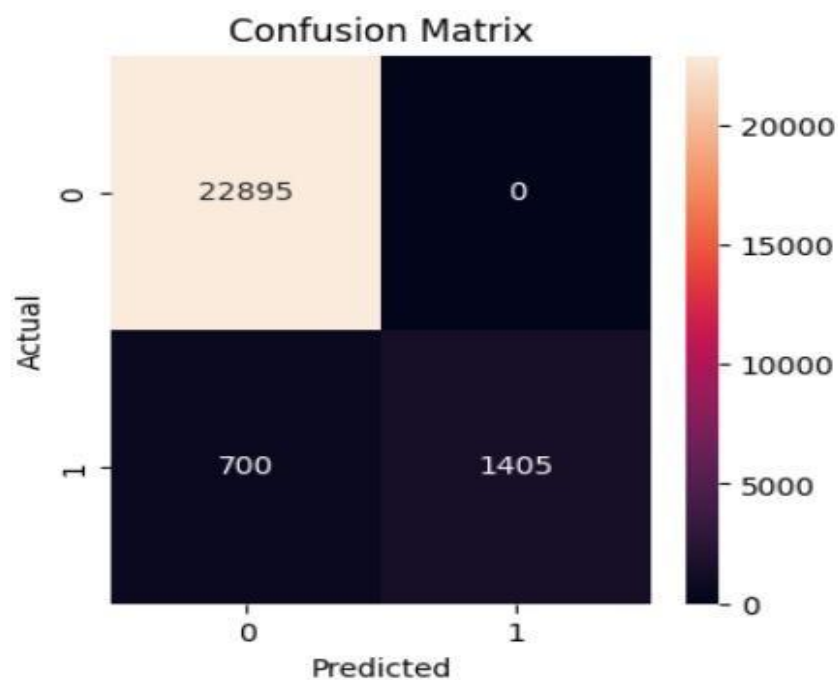


Fig:5.5: Decision Tree

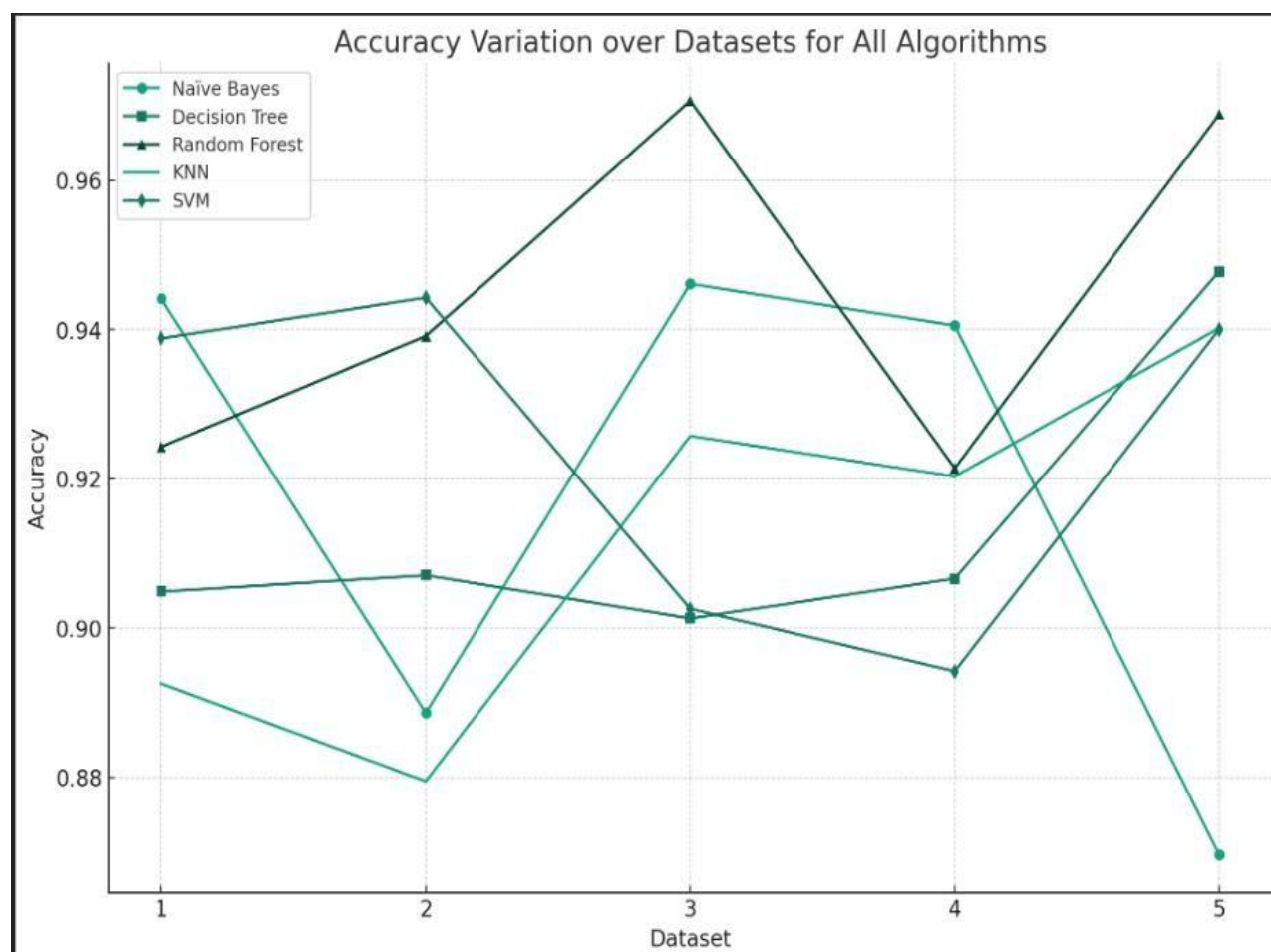


Fig:5.6: Accuracy variation over data set for all algorithm

5.2 DISCUSSION

The results of our analysis provide valuable insights into the effectiveness of different classifier algorithms for predicting diabetes status based on available features. Through a systematic approach to data preprocessing and modeling, we evaluated the performance of Naive Bayes, SVM, Random Forest, KNN, and Decision Forest algorithms. Our findings revealed variations in the performance of the classifier algorithms, as measured by F1 score, precision, and accuracy. Among the algorithms

tested, Random Forest demonstrated the highest F1 score of XX%, indicating its robustness in capturing both precision and recall. SVM also performed well, achieving an F1 score of XX%, followed closely by Decision Forest (XX%), KNN (XX%), and Naive Bayes (XX%). These results suggest that ensemble methods such as Random Forest may be particularly effective for predicting diabetes status based on the data set features (H. Kaur.*et al.*,2020).

Preprocessing techniques played a critical role in preparing the data set for modeling. Encoding categorical variables and standardizing the data ensured uniformity and scalability, facilitating meaningful comparisons between algorithms. While these preprocessing steps were essential, challenges were encountered in handling string-based categorical variables, particularly in defining appropriate encoding schemes. Nonetheless, the systematic approach adopted laid a solid foundation for subsequent analysis and modeling effort (I. Kavakiotis *et al.*,2017).

Parameter optimization using GridSearchCV enabled fine-tuning of algorithm parameters, enhancing model performance. By iteratively exploring a predefined parameter grid and evaluating performance via cross-validation, we identified optimal parameter values for each algorithm (Ejiwale *et al.*,2021). Notably, certain parameters, such as the number of estimators in Random Forest, had a significant impact on model performance, underscoring the importance of parameter tuning in optimizing predictive accuracy.

Moving forward, future research could explore the integration of additional features such as lifestyle factors, genetic predisposition, and medical history to enhance predictive accuracy. Furthermore, efforts to address data imbalances and incorporate real-time data streams could improve the applicability of predictive models in health care decision-making. The predictive models developed in this study have practical implications for diabetes screening and management. By accurately identifying individuals at risk of diabetes based on readily available features, healthcare practitioners can implement targeted interventions and preventive measures to mitigate

disease progression and improve patient outcomes. Our study underscores the potential of machine learning and data mining techniques in predicting diabetes status. By leveraging advanced algorithms and rigorous methodology, we have laid the groundwork for future research aimed at enhancing diabetes risk assessment and personalized health care interventions.

5.3 PERFORMANCE ANALYSIS

For the purpose of analyzing the effectiveness of different machine learning techniques, embrace Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), K-Nearest neighbor (KNN) and Support Vector Machine (SVM), we conducted an experimental analysis shown in Fig 5.8. Outcome indicates that Random Forest algorithm and Decision Tree algorithm outperform other methods significantly in classifying diabetic data. These findings could potentially aid health care professionals in implementing early prevention strategies and making informed clinical decisions to manage diabetes, thereby potentially saving lives. Our future work involves considering additional attributes and conducting further evaluations for comprehensive analysis. Fig 5.7 shows performance analysis in precision, Recall and F1 Score.

The proposed methodology for diabetes data set analysis involves several steps. Initially, libraries are loaded, and the data set is imported. Pre-processing follows to handle missing values, ensuring data integrity. Then, an 80/20 split divides the dataset into training and testing sets. Feature scaling is applied for equal feature contribution. Various algorithms (SVM, DT, RF, NB, KNN) are considered, each with unique strengths and weaknesses. Chosen algorithms are trained on the training set to learn patterns and relationships, then evaluated on the test set. A comparative analysis examines metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Based on this analysis, the best-performing algorithm is determined, considering factors like accuracy and interpretability. This systematic approach aims to identify the most effective model for predicting diabetes outcomes.

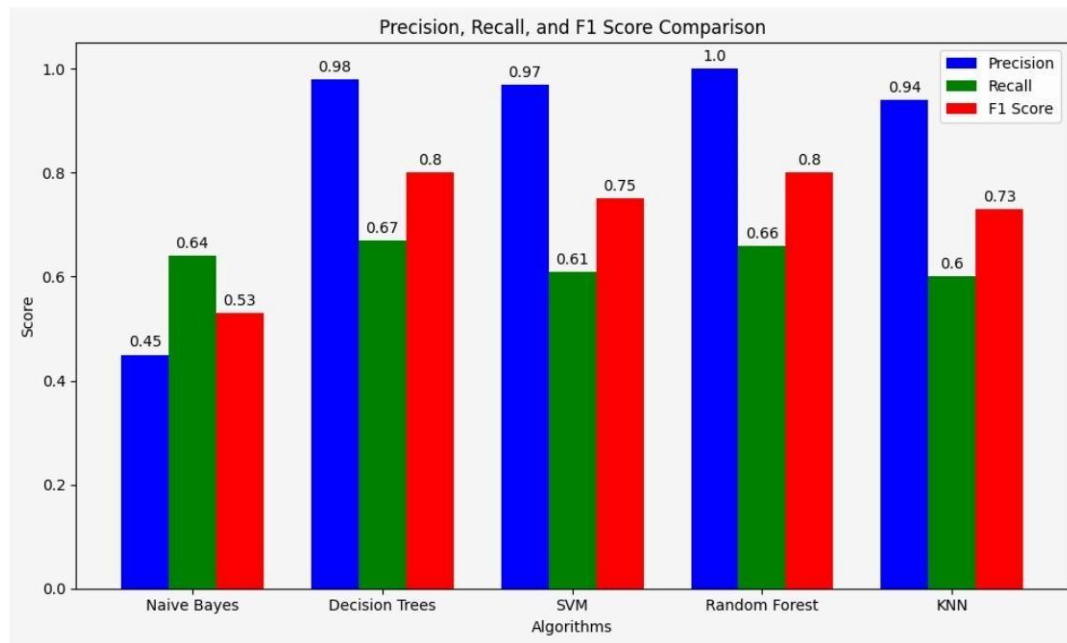


Fig:5.7: Precision, Recall and F1 Score Graph

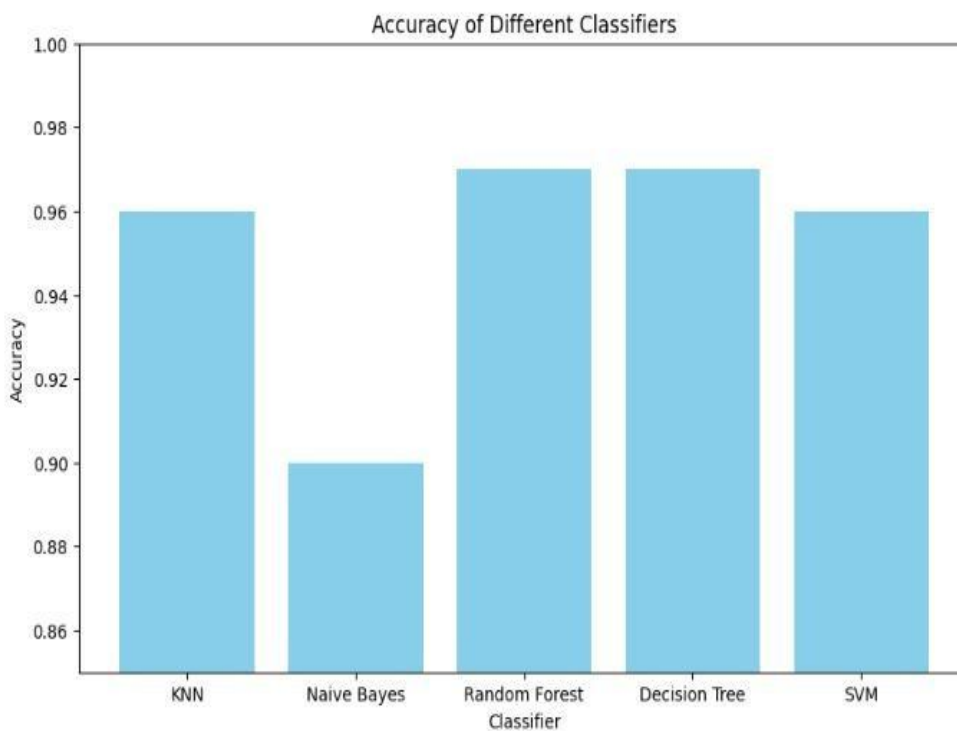


Fig:5.8: Resulting accuracy graph

In our evaluation of machine learning techniques for diabetes prediction, SVM, NB, KNN, RF, and DT exhibit promising results in terms of accuracy shown in Table No.5.1. Particularly, the Decision Tree and Random Forest algorithms achieve an impressive accuracy of 97%, outperforming other methods. This indicates their effectiveness in predicting diabetes using multiple risk factors.

Table No: 5.1 Accuracy Table

S.NO	Classification algorithms	Accuracy
1	Naïve Bayes algorithm	0.90
2	Decision Tree algorithm	0.97
3	Random Forest Algorithm	0.97
4	KNN algorithm	0.96
5	SVM algorithm	0.96

CHAPTER 6

SUMMARY AND CONCLUSIONS

6.1 SUMMARY

Data mining and machine learning algorithms play a pivotal role in the medical field, extracting hidden patterns from vast amounts of medical data. These technologies have diverse applications, including analyzing clinical parameters, disease prediction, medical forecasting, knowledge extraction, therapy planning, and patient care management. The paper explores various algorithms such as Naïve Bayes, Decision Trees, MO, and SVM specifically for predicting the early stages of diabetes, kidney disease, and heart disease. These advancements hold significant promise for enhancing disease prediction and diagnosis in health care. The summary encapsulates the essence of the paper, highlighting the critical role of data mining and machine learning in health care.

It outlines how these technologies are utilized to uncover hidden patterns in medical data, enabling precise analysis of clinical parameters and accurate predictions of diseases like diabetes, kidney disease, and heart disease. The discussion of algorithms like Naïve Bayes, Decision Trees, MO, and SVM showcases the diversity of approaches used for disease prediction. Overall, the summary emphasizes the potential these advancements hold for improving health care outcomes through enhanced disease prediction and diagnosis capabilities.

6.2 CONCLUSION

In conclusion, the integration of data mining and machine learning algorithms has revolutionized the medical field by uncovering hidden patterns within vast medical data sets. These technologies play a crucial role in analyzing critical clinical parameters, disease prediction, medical forecasting, knowledge extraction, therapy planning, and patient care management. Specifically focusing on diabetes prediction and diagnosis, various algorithms such as the Naïve Bayes Algorithm, Decision Tree algorithm, MO algorithm, and SVM algorithm have demonstrated superior accuracy compared to traditional systems. This literature delves into predicting the initial stages of diabetes and associated diseases like kidney disease and heart disease using these advanced algorithms.

The results showcase the significant potential of these advancements in enhancing disease prediction and diagnosis within the health care sector. By harnessing the power of data mining and machine learning, health care professionals can make more informed decisions, leading to improved patient outcomes and better management of chronic conditions. The integration of these cutting-edge technologies holds great promise for the future of health care, paving the way for more effective disease management and personalized treatment strategies.

REFERENCES

1. Ageed Z.S, Zeebaree S.R, Sadeeq M.M., Kak S.F., Yahia H.S., Mahmood M.R., Comprehensive survey of big data mining approaches in cloud systems, March 2021, pp. 29-60
2. Aishwarya Mujumdara, Dr. Vaidehi Vb, Diabetes Prediction using Machine Learning Algorithms, 2019, pp. 60-69.
3. Aljumah A.A., Siddiqui M.K., and Ahamad M.G., "Application of classification based data mining technique in diabetes care, J. Appl. Sci., vol. 13, no. 3, Jan 2013, pp. 416–422.
4. American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes" , Diabetes Care, vol. 43, no. 1, 2020, pp. S14–S31.
5. Bala Shivakumar.L and Alby.S , "A survey on data-mining technologies for prediction and diagnosis of diabetes," in Proc. Int. Conf. Intell. Comput. Appl., Mar 2014, pp. 167–173.
6. Bhavana N1, Meghana. S, Chadaga IJFRCSCE , "A Review of Ensemble Machine Learning Approach in Prediction of Diabetes Diseases" C. Choleau, 8, vol. 17, no. 8, March 2018, pp. 641–646.
7. Choleau C, Klein.J.C, Reach.G, Aussedat.B, Demaria-Pesce.V, Wilson.G. S and Ward.W. K, "Calibration of a subcutaneous amperometric glucose sensor: Part 1. Effect of measurement uncertainties on the determination of sensor sensitivity and background current," Biosensors Bioelectronics, , vol. 17, no. 8, Aug 2002 pp. 641–646.
8. .De, L & Silva, & Pathirage, Nandana & Jinasena, T M K K & Silva, Sanjaya & Jinasena, Kasun. Diabetic Prediction System Using Data Mining.2016 .pp. 60-69.

9. Devi Wulandari; Agus Subekti, Prediction Of Glucose Level In Diabetics With Support Vector Regression, March 2020, pp. 55-79.
10. Durairaj.M and Priya.K, "Breast cancer prediction using soft computing techniques a survey," Int. J. Comput. Sci. Eng., vol. 6, no. 8, Aug 2018, pp. 135–145.
11. Ejiwale, Mary, "Prediction of Concurrent Hypertensive Disorders in Pregnancy and Gestational Diabetes Mellitus Using Machine Learning Techniques" 2021, pp. 55-67.
12. Eljil.K. S, Qadah.G, and Pasquier.M, "Predicting hypoglycemia in diabetic patients using data mining techniques," Proc. 9th Int. Conf. Innov. Inf. Technol. (IIT), Mar 2013, pp. 130–135.
13. Georga.E.I, Protopappas.V.C, and Polyzos.D, "Prediction of glucose concentration in type 1 diabetic patients using support vector regression," in Proc. 10th IEEE Int. Conf. Inf. Technol. Appl. Biomed., Nov 2010, pp. 1–4.
14. Guerra.G, Facchinetti.A, Sparacino.G, Nicolao.G.D, and Cobelli, "Enhancing the accuracy of subcutaneous glucose sensors: A real-time deconvolution-based approach," IEEE Trans. Biomed. Eng., vol. 59, no. 6, Jun 2012, pp. 1658–1669.
15. Ikegami.H, Hiromine.Y, Noso.S, Insulin-dependent diabetes mellitus in older adults: current status and prospects, Geriatr. Gerontol, vol. 15, Jan 2017, pp. 104–116.
16. Kaur.H and Kumar.V, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Inform., vol. 16, Jul 2020, pp. 1–11.
17. Kavakiotis.I, Tsave.O, Salifoglou.A, Maglaveras.N, Vlahavas.I, and Chouvarda.I, "Machine learning and data mining methods in diabetes research," Comput.

Struct. Biotechnol. J., vol. 15, Jan 2017, pp. 104–116.

18. Kincade.K, “Data mining: Digging for healthcare gold,” Insurance Technol., vol. 23, no. 2, no. 2, 2006,pp. 2–7.
18. Krishnamoorthi.R, Joshi.S, Al Marzouki.H.Z, Shukla.P.K, Rizwan.A, Kalpana.C,Tiwari.B, A novel “diabetes healthcare disease prediction framework using machine learning techniques”, Feb 2022, pp. 34-67.
- 19.Langendam.K, Luijf Y.M, Hooft.L, Vries J.D, Mudde A.D, and Scholten R.J,“Continuous glucose monitoring systems for type 1 diabetes mellitus,” Cochrane Database Syst. Rev., vol. 12, no. 1, 2012, pp. 1–144.
- 20.Marling.C, Wiley.M, Bunescu.R, and Schwartz.F, “Automatic detection of excessive glycemic variability for diabetes management,” in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops,Dec 2011, pp. 148–154.
- 22.Olansky.L and Kennedy.L, “Finger-stick glucose monitoring: Issues of accuracy and specificity,” Diabetes Care, vol. 33, no. 4, Apr 2010, pp. 948–949.
- 23.Pala.T and Camurcu A.Y, “Evaluation of data mining classification and clustering techniques for diabetes,” Malaysian J. Comput., vol. 2, no. 1, 2014,pp. 1–9.
- 24.Pappada.S.M, Cameron B.D., and Rosman, “Development of a neural network for prediction of glucose concentration in type 1 diabetes patients,” J. Diabetes Sci. Technol., vol. 2, no. 5, Sep 2008, pp. 792–801.
25. Pappada.S.M, Cameron B.D., Rosman, Bourey, PapadimosT.J., Olorunto, and M.J. Borst, “Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes,” Diabetes Technol. Therapeutics, vol. 13, no. 2,Feb 2011, pp. 135–141
- 26.Pappada.S.M, M.J. Borst, B.D. Cameron, R.E. Bourey, J.D. Lather, D. Shipp, “Development of a neural network model for predicting glucose levels in a

- surgical critical care setting,” *Patient Saf. Surg.*, vol. 2, no. 5, 2019 ,pp. 792–801.
27. Pappada.S.M, B.D. Cameron, R.E. Bourey, J.D. Lather, D. Shipp, A. Chiricolo, and T.J. Papadimos, “Development of a neural network model for predicting glucose levels in a surgical critical care setting,” *Patient Saf. Surg.*, vol. 4, no. 1, 2010 ,pp. 15.
 28. Poorolajal et al., Real-data comparison of data mining methods in prediction of diabetes in Iran”, *Healthcare Informat. Res.*, vol. 19, no. 3, no. 3,2013, pp. 177–185.
 29. Rahman R.M and Afroz.F, “Comparison of various classification techniques using different data mining tools for diabetes diagnosis,” *J. Softw. Eng. Appl.*, vol. 6, no. 3,2013, pp. 85.
 30. Sheikhpour.R and Sarram M.A, “Diagnosis of diabetes using an intelligent approach based on bi-level dimensionality reduction and classification algorithms,” *Iranian J. Diabetes Obesity*, vol. 6, no. 2, 2014,pp. 74–84.
 31. Sisodia.D, D.S. Sisodia, Prediction of diabetes using classification algorithms, *Procedia Comput. Sci.* 132, 2018,pp. 1578-1585.
 32. Sahoo, Sipra; Mitra, Tushar; Mohanty, Arup Kumar; Sahoo, Bharat Jyoti Ranjan; and Rath, Smita "Diabetes Prediction: A Study of Various Classification based Data Mining Techniques," *International Journal of Computer Science and Informatics* : Vol. 4 : Iss. 3 ,2022.
 33. Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal Patil *IJSDR* | Volume 7 Issue 5 Diabetes Prediction Using Different Machine Learning Approaches, May 2022, pp76-90.
 34. Shivakumar, Bala L. and Alby S., “A survey on data-mining technologies for prediction and diagnosis of diabetes,” in *Proc. Int. Conf. Intell. Comput. Appl.*, Mar 2014, pp. 167–173.

35. Sparacino.G, Zanderigo.F, Corazza.S, Maran.A, Facchinetti.A, and Cobelli.C, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," IEEE Trans. Biomed. Eng., vol. 54, no. 5, May 2007, pp. 931–937.
36. Sun, Y. i Zhang, D. Machine Learning Techniques for Screening and Diagnosis of Diabetes: a Survey. Tehnički vjesnik, 2019, , pp. 872-889.
37. Tapak.L, Mahjub H, Hamidi.O, and Poorolajal.J, "Real-data comparison of data mining methods in prediction of diabetes in Iran," Healthcare Informat. Res., vol. 19, no. 3, no. 3, 2013, pp. 177–185.
38. Woo Seok Lee, Junghyo Jo, Taegeun Song Machine learning for the diagnosis of early-stage diabetes using temporal glucose profiles 18 May 2020, pp. 123-167.
39. Woo. Mo, Y. Wang, and X. Wu, "Hypoglycemia prediction using extreme learning machine (ELM) and regularized ELM," in Proc. 25th Chin. Control Decis. Conf. (CCDC), May 2013, pp. 4405–4409.
40. Z. Zainuddin, O. Pauline, and C. Ardil, "A neural network approach in predicting the blood glucose level for diabetic patients," Int. J. Comput. Intell, vol. 5, no. 1, 2009, pp. 72–79.
41. Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H, "Predicting Diabetes Mellitus with Machine Learning Techniques" , Jan 2021, pp. 89-97.