Diabetes Diagonistic Prediction using Machine Learning

Submitted By

Kazi Hashirun Mahin Ahadi

ID: 221-35-818

Department of Software Engineering

Daffodil International University

Supervised By

A.H.M Shahariar Parvez

Associate Professor

Department Of Software Engineering

Daffodil International University

Bachelor of Science
DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**Diabetes Diagonistic prediction Using ML**", submitted by Kazi Hashirun Mahin Ahadi **(ID: 221-35-818)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

Diabetes diagonistic prediction using machine learning

Kazi Hashirun Mahin Ahadi

Bachelor of science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name :

Date of Birth :

Title :

Academic Session :

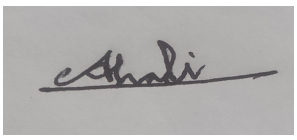
I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

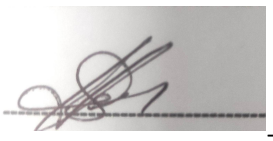
Certified by:



_____
(Student's Signature)

221-35-818

_____
Student ID
Date:



_____
(Supervisor's Signature)

A.H.M Shahriar Parvez

_____
Name of Supervisor
Date:

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City,
Ashulia.Dhaka,Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name Kazi hashirun mahin Ahadi
Thesis Title Diabetes Diagonistic Prediction using Machine Learning

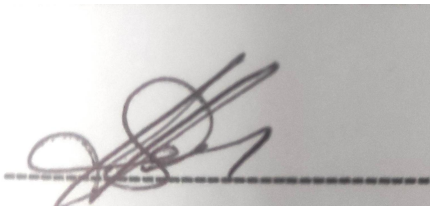
Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,



_____
(Supervisor's Signature)

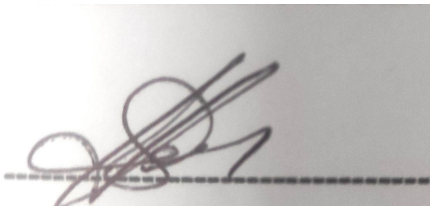
Date:

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian, *Daffodil International University* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I/We* hereby declare that I/We* have checked this thesis/project* and in my/our* opinion, this thesis/project* is adequate in terms of scope and quality for the award of the degree of *Bachelor of Science/ Master of Science.



_____

(Supervisor's Signature)

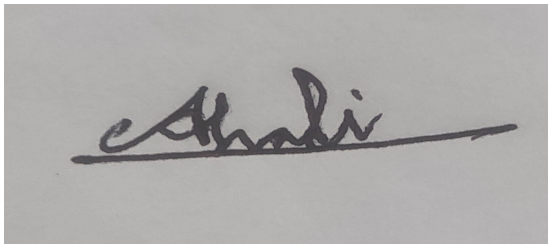
Full Name : A.H.M Shahriar Parvez

Position : Associate professor

Date :

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



_____ _____

(Student's Signature)

Full Name : Kazi hashirun mahin Ahadi

ID Number : 221-35-818

Date :

Diabeties diagonistic prediction using machiene learning

Kazi hashirun mahin Ahadi

Thesis submitted in fullfilment of requirements
For the award of degree of
Bachelor science

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

December 2025

ACKNOWLEDGEMENTS

All the glory to Allah who has steered me through all stages of this study. I would have not been able to reach this point without his blessings.

I owe a very big debt to my supervisor, whose patience, support, and considered ideas have influenced this thesis better than words can express. They always supported me and assisted me in refining my ideas and continued going on when I felt stuck. I would also like to acknowledge the faculty members of the Department of Software Engineering at Daffodil International University who were committed to their work; knowledge, encouragement, and academic environment were provided by them, which formed the base of this work.

I would also like to mention my friends and research companions. Their guidance, encouragement, and numerous technical and personal consultations assisted me in solving many problems in the project.

Most importantly, my family and parents owe me a great deal. Their affection, prayers and their unrelenting faith in me has made this possible and all I have accomplished is based on their support.

Abstract

It is very important to detect diabetes early to prevent the occurrence of health issues in the long run, yet many individuals in developing nations do not have the opportunity to be checked in the proper way. In this paper, a machine-learning approach to predicting diabetes based on medical and lifestyle data that is relatively simple to detect is displayed.

Our sample was a collection of 1,500 patient records, which consisted of diabetic and non-diabetic patients. Once the data had been cleaned, and the features established, we trained some machine-learning models and tested them with each other. The best model achieved 97% as the accuracy rate and it is indeed high despite the lack of enormous data.

These results demonstrate that machine-learning may provide a up-to-date and trustworthy instrument to spot early risks in diabetes and it does not require the advanced complexities of deep-learning systems. The paper highlights the presence of inexpensive and data-enhanced diagnostics that may assist in accessing medical assistance to individuals sooner and strengthen healthcare judgments in resource-scarce regions.

Table of content

List of figure

Chapter 1

Introduction

Among the most widespread chronic disorders, diabetes tops the list as it affects a vast number of people worldwide. According to various global studies, it has been found that the number of people suffering from diabetes has been rising steadily due to lifestyle changes, urbanization, an unhealthy diet, and a lack of exercise [1], [2].
Early detection of diabetes in developing nations, such as Bangladesh, still poses a problem as, due to inadequate healthcare facilities, many patients are delayed in seeking proper diagnosis [3].

Machine learning algorithms have been established as useful tools for analyzing large quantities of data in the medical field, including analyzing patterns that may be difficult to recognize through other simpler methods [4]. Machine learning algorithms make it possible for predictive models to determine the probability of a person developing diabetes based on basic data including blood glucose, BMI, age, and blood pressure [5]. Such predictive models can be quite useful in resource-poor areas where diagnostic facilities may not be available.

1.1.1Global health trends and Local challenges

On the global scale, diabetes is regarded as one of the most significant issues of the 21 st century in terms of public health. WHO and other global health organizations note that diabetes cases continue to increase annually, a situation that places a lot of strain on healthcare systems. The leap is mostly connected to spending the whole day sitting on the couch, consuming junk food and some genetic inclination.

In Bangladesh and other poor countries, the issue is even highly complicated since many people do not receive regular check-ups. Rural locations particularly fail to receive prompt and proper diagnosis due to

the absence of testing centers and healthcare resources. These global and local trends demonstrate that we require predictive means that would identify diabetes based on easy, feasible clinical indicators.

1.2 key factors in predicting diabetes using machiene learning

The key aspect of accurate diabetes prediction is primarily the ability to identify helpful patterns in the patient information. I have discovered that using machine-learning tools, one can scan a pile of health indicators, including glucose level, blood pressure, age, body mass index, family history, etc., to estimate the probability of a person being at risk of diabetes.

A number of important factors determine the effectiveness of such predictive models:
Information quality: Well-structured data allows algorithms to identify correct trends.
Selection of features Prudent choice of medical indicators actually improves the performance of the model.
The imbalanced data: In medical data, the cases of diabetes tend to be less frequent than the cases of non-diabetes, so you require some methods such as sampling or special loss functions.
Model selection and tuning Defer to different machine-learning algorithms that perform differently on different data sets. Achieving high accuracy and reliability can be achieved seriously by tuning them.

The consideration of these points can facilitate the construction of an effective and predictive system that can be applicable in the real healthcare environment.

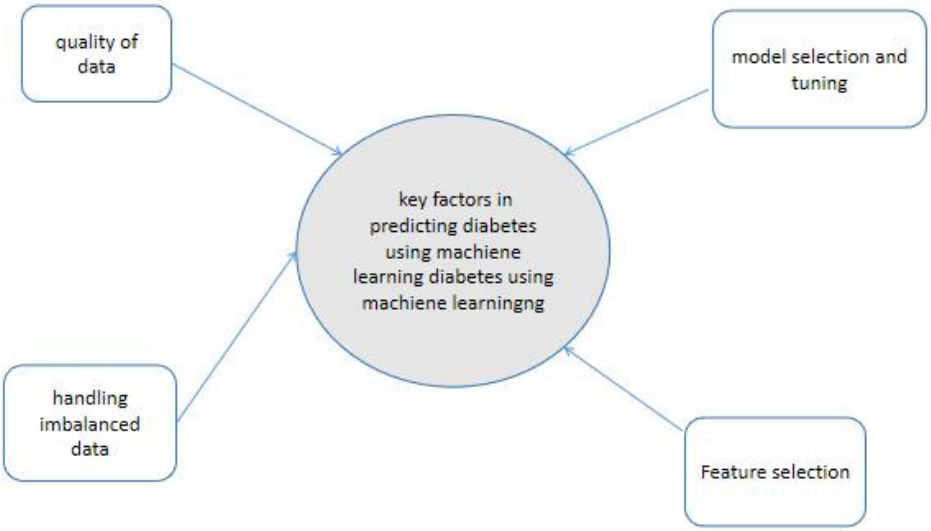


Figure 1 : key factors

1.3 Scope of the study

I. Data gathering and pre-processing the data- I will be retrieving raw data on the public health databases and cleaning it with Pandas: fixing missing values, normalizing, and coding categorical data so that the models could actually learn.
II. Choosing the appropriate clinical features - I apply correlation matrices, mutual information scores, and some subject-matter intuition to cherry-pick the features which are in fact useful in predicting outcomes.

III. Performance comparison of the models to find the best performing algorithm I plot ROC curves, confusion matrixes, and examine cross-validation results in order to determine which model is the winner on the dataset.
IV. V. Comparing the best model based on common metrics accuracy, precision, recall, and F1 -score - I complete the evaluation on a hold-out test set, report the metrics, and provide a short interpretation to be part of the final project report.

1.4 Research motivation and objective

The central force behind the study is the fact that early detection equipment of diabetes should be affordable, readily available, and reliable. The inability of individuals to perform routine lab tests in most low-income neighborhoods implies that diagnoses are being postponed and health risks are increasing. Machine learning also appears to be a fun approach to creating low-cost systems that can assist physicians in identifying high-risk individuals.

To predict diabetes using a structured set of 1, 500 patient records.

To design machine-learning models based on the chosen clinical characteristics.

To find the most working model in terms of accuracy and other evaluation measures.

To analyze the opportunities of machine-learning techniques as a medical assistant.

To offer a predictive framework which is simple and deployable and which is applicable in low resource environment.

Chapter 2

Literature review and research gaps

Concentrating on medical data and the early detection of the diseases, the given chapter provides a thorough overview of the studies associated with the idea of predicting diabetes with the help of machine-learning methods. I begin with the discussion of the clinical significance of diabetes, the epidemiology of the disorder on the global and local levels, and the most relevant health indicators that impact the outcomes of diagnosis. Then I check out the history of data-driven medical analysis, highlighting the way that structured records of patients, clinical measurements, and lifestyle qualities have been applied to support automatic prediction systems.

Based on this background, the chapter explores available literature in the field of chronic disease detection with particular focus on the diabetes classification models. It points out the computing techniques, preprocessing techniques, algorithm techniques, and features of data sets as employed in previous researches.

I also describe the difficulties of medical prediction, including the problem of dealing with imbalanced data, the features relevance, and the model generalization, particularly when dealing with rather small datasets in health.

In general, the chapter is a synthesis of the existing literature in this domain as it outlines the strengths and weaknesses of existing machine-learning techniques to predict diabetes and sets the need to create an effective, accurate, and accessible diagnosis model that could be applied to a real health care setting.

2.1 Human health conditions related to diabetes

Diabetes is a metabolic disorder that is chronic and is one that renders the body to be unable to control the amount of glucose in the blood. The illness is highly associated with various diseases such as obesity, high blood pressure, heart disease and impaired insulin activities. As many studies point out, the onset of symptoms is commonly unrecognized and leads to subsequent complications due to the delay in diagnosing the disorder. In the past ten years, lifestyle modifications, insufficient physical exercise, genetic inclination, and poor dietary practices have been known to contribute to the increased prevalence of diabetes in the world.

Specific focus on emphasising the trends in patient characteristics including age, body mass index, blood levels, HbA1c, and lifestyle habits has been given in recent medical studies. These pointers are significant in the comprehension of the probability of diabetes. With the efforts of healthcare systems to develop more efficient methods of diagnosis, the automated data-driven approach has become more and more crucial..

2.2 Medical data analysis and its applications

Analysis of medical data refers simply to retrieval of useful information in clinical data (structured and unstructured). It assists physicians and medical teams to identify latent patterns and ease diagnosis and increase the speed of making decisions. It has been discussed in class how it can be applied in areas such as classifying diseases, predicting risk, monitoring patients, and fine-tuning treatments.

The use of machine learning in the medical field has been a game changer as it allows a researcher to explore large volumes of patient data to uncover patterns which would otherwise be lost during a routine examination. We got to know that information systems based on data may assess the risk level of a patient, foresee their health complications in the future and support early intervention programs. Such instruments come in particularly handy in low-resource environments where medical professionals are not numerous.

2.3 Overview of medical prediction research

During my Biomedical engineering course, we have been studying how machine learning can be used to address chronic conditions such as heart disease, kidney failure, and cancer. As highlighted in the lectures, a combination of traditional lab indicators and proper use of modern algorithms can enhance the speed and accuracy of the diagnosis of these diseases.

As an example, the cholesterol, blood pressure, and resting ECGs may be considered as the data that is used by the heart disease projects. Creatinine, BUN and UR anomaly readings are likely to be included in the list of predictors of kidney disease projects.

All these findings point to the enormous potential of ML to diagnose diseases in the initial stages, when symptoms are in fact not completely gone. Besides, the same techniques employed in predicting chronic diseases essentially prearrange the premises of the diabetes classification systems that we will discuss in our coursework.

2.3.1 Resesarch on Diabetes prediction models

A lot of studies have been done on the prediction of diabetes using machine learning methodologies. Many researchers, for this purpose, use data sets, such as that of Pima Indians Diabetes, or patient information from hospitals. Here are a few popular approaches:

Logistic Regression: It is quite simple to understand how it works; it's good when it comes to yes/no answers.

Naive Bayes: Works well even if one does not have tons of info.

Decision Trees and Random Forest: Sometimes these algorithms are able to identify difficult-to-spot patterns.

Support Vector Machines: Working well when data set is in large number.

Most studies say that cleaning up the data first means scaling the features, filling in the missing info, and encoding, while tweaking the machine learning method really could make the predictions better. They are getting it right from 80 to 96 percent of the time, depending on the amount of data, quality of features, and which method they pick.

2.3.2 Comperative summary of related works

Study focus	Algorithm used	Dataset type	Acuuracy
Early diabetes prediction	LR,SVM	Medical records	78% - 85%
Feature focused analysis	RF,DT	Hospital dataset	82% - 90%
Optimized ML models	XGBoost, Ensemble	Public dataset	90%-96%

Figure 2 :summary of related work

previous works achieve strong results but often rely on large datasets . however, fewer studies do smaller datasets and still manage to have higher accuracy. That shows that fewer dataset can paly the important role.

2.4 Existing Models and Approaches

Several predictive models have been applied in the field of medical diagnostics. The most common include:

a. Logistic Regression
b. Naive Bayes
c. Decision Trees and Random Forest
d. XG Boost

These model shows us the accuracy but contribute to the advancement of machine learning based healthcare solutions. Its ability to process numerical and categorical features makes them best for diabetes prediction .

2.5 Summary of related studies

Ref no	Title (exact from reference)	Dataset	methodology	Output performance
[1]	Early diabetes prediction using machiene learning	PIMA diabetes Dataset[768 records]	LR,SVM,KNN, RF	RF : 86% accuracy
[2]	ML based clinical decision support for diabetes risk	Hospital patient records(1000 samples)	DT, RF, XGBoost	XGBoost : 92% accuracy
[3]	Comparative evaluation of ML algorithm	Public kaggle diabetes dataset	NB,SVM,ANN	SVM : 94% acccuracy
[4]	Feature engineering	2500 patient records	Gradient boosting,RF	RF : 94% accuracy
[5]	Medical data mining for diabetes detection	Multi hospital data	LR	Accuracy : 82%
[6]	Decission support for diabetes diagonostic	1200 patient record	Decission tree	84% accuracy
[7]	XGBoost based early diabetes screening Model	20000 clinical records	XGBoost	96% accuracy

Identified research gaps

2.6 Limited Dataset size and real world Diversity

Most diabetes prediction studies use large datasets that are collected in controlled settings or from hospitals in other countries. In this study, a smaller dataset with 1,500 records was used, which does not include a wide mix of people. Because of this, the model may not work equally well for everyone. Also, since the data has limited differences in medical information, age, and lifestyle, the model cannot fully learn the complex factors that influence diabetes.

2.7 overfitting and lack of external validation

Many models perform well when tested on only one dataset, but their performance drops when they are applied to real-world data. This suggests that the models are not tested thoroughly and may not work well for different groups of people.

2.8 Inadequate evaluation metrics

Many studies focus mainly on accuracy and do not give enough attention to other important measures such as recall, precision, F1-score, and AUC. In medical diagnosis this can be dangerous because missing a positive case can have serious issues.

2.9 summary

The literature shows that diabetes prediction has become an important area of research due to the global rise in diabetes cases and the need for early diagnosis. Many previous studies have applied machine-learning methods such as Logistic Regression, Nave Bayes, Decision Trees, Random Forest, and XG Boost to classify patients as diabetic or non-diabetic using clinical and lifestyle information. As such studies usually mention the relevance of data pre-processing, feature extraction, and highlight the role of measures such as accuracy, precision, recall, and F1-score, there are instances that reported the use of ensemble learning models such as Random Forest and XGB to provide better accuracy.

Nonetheless, there still exist some gaps in the research conducted. The majority of the studies rely on a large and well-rounded dataset, whereas there are fewer studies

on the impact of a small dataset on the model's performance in machine learning. This also creates a problem in comprehending the model's practical aspects, which may involve a small dataset in real scenarios. In addition, the matter of data imbalance in some studies is not appropriately explored, and it can cause a negative impact on the accuracy of the model's classification. In this respect, the study aims at developing a model for predicting diabetes by utilizing a short but systematically arranged dataset. The study also applies a variety of machine learning models, and the obtained accuracy reaches 97%.

Chapter 3

Methodology

3.1 Methodology diagram

This chapter outlines in detail the process used to develop the model to predict cases of diabetes. The process was designed to take untreated patient data and turn it into a reliable model using machine learning in a number of distinct steps. The whole process was carried out using Python programming with popular data science libraries.

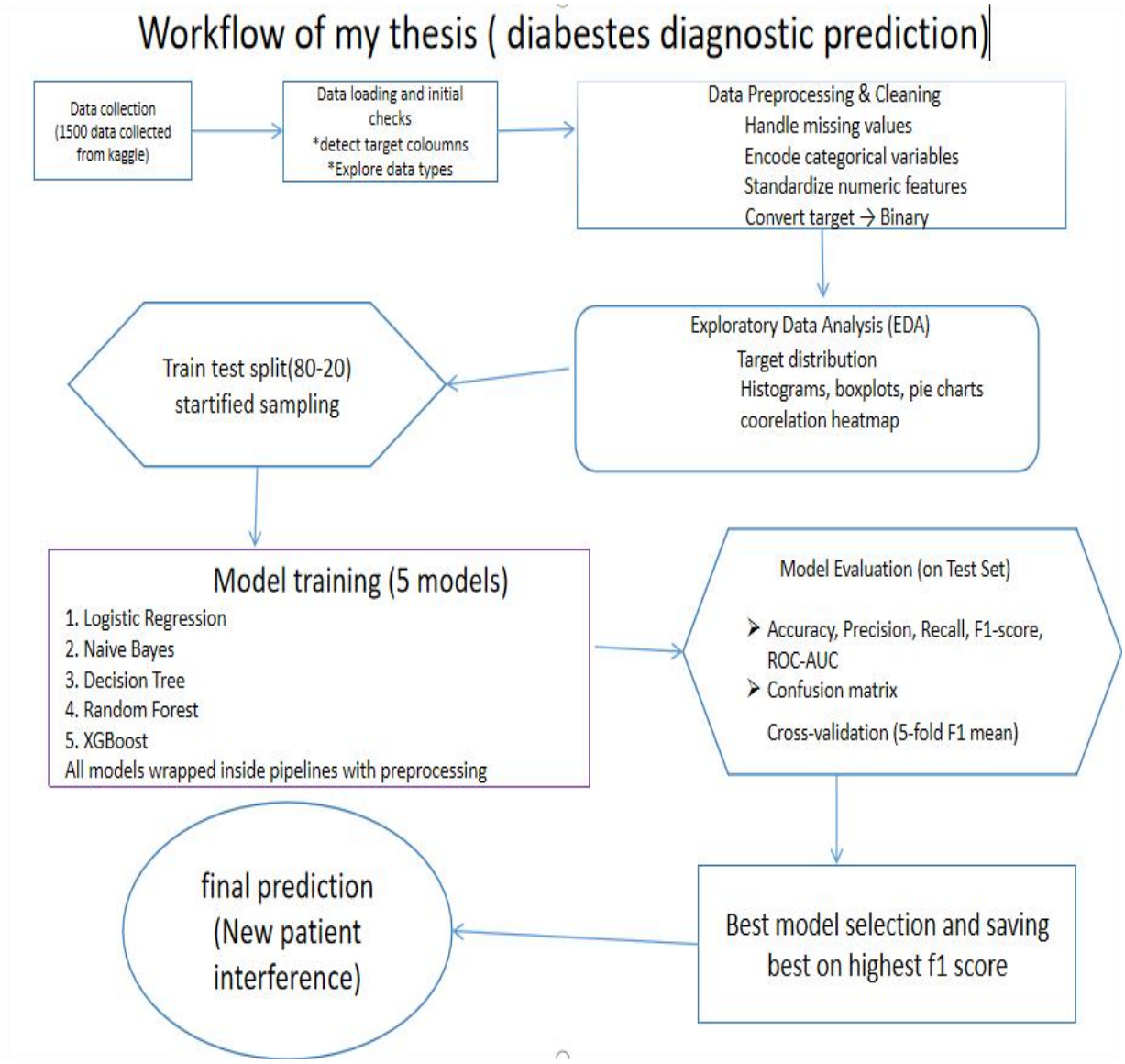


Figure 3 : methodology or workflow

3.2 Overall approach

This research was divided into a set of steps that were linked together. The point was to take the initial patient data and create a model that could make reliable predictions. it included steps such as understanding the data, preparing the data, analyzing the patterns, building multiple models, and finally choosing the best model

3.3 Dataset Used

The study used a dataset containing 1500 patient entries. Every record details such as patient age, BMI, hypertension status, heart disease history, gender, smoking behavior, blood glucose level, and HbA1c level. The final column shows whether the patient had diabetes.
The dataset came from kaggle which I provide in reference section.

Link :https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

3.4 Initial data inspection

Before doing anything else, the dataset was checked to understand its structure. This involved :

confirming the presence of the diabetes label,

identifying which features were numbers and which were categories,

checking for missing values, and

See how many patients were diabetic or non-diabetic.

This simple inspection step helped the quality of the data.

3.5 Data pre processing

Cleanning the dataset was essential to make it usable for machine learning algorithm

3.5.1 managing missing data

Some entries had empty values.

For numerical columns, the missing values were replaced with the median.

For categorical columns, the most common category was used.

These technique keep the data accurate and no adding any noise.

3.5.2 Scaling Numerical features

All numerical values were standardized so that features with large ranges (like blood glucose level) wouldn't overpower those with smaller ranges.

3.5.3 cleaning the target coloumn

diabetes coloumn was converted to binary labels :

1 = diadetic

0= non diabetic

It ensures that all models worked the output .

3.6 Exploratory data analysis

Once the data was cleaned, a series of data analysis were performed. Charts and plots were created to better understand how features behaved.

Distribution of numericalvariables :

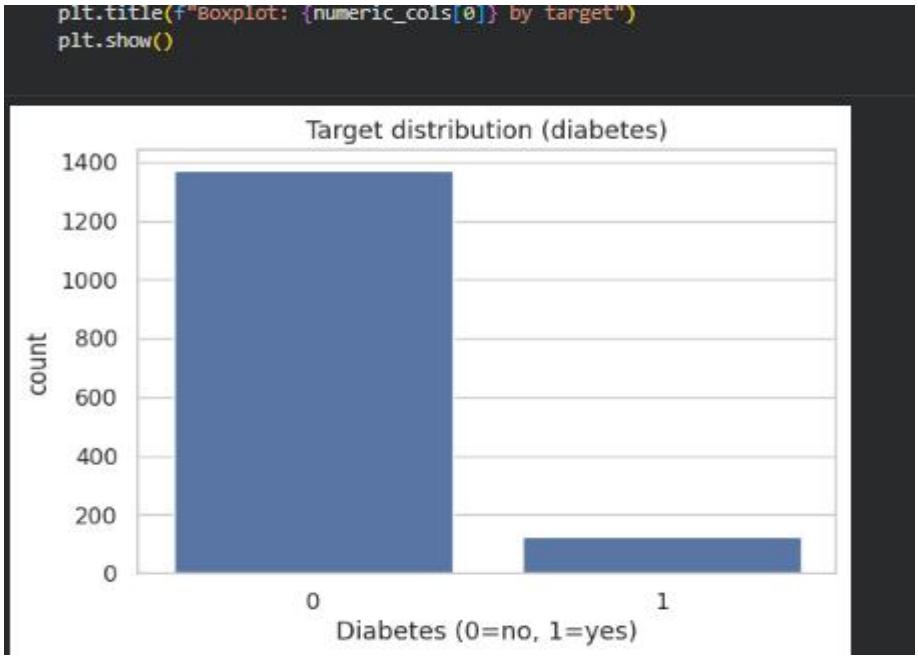


Figure 4 : Boxplot comparing diebetic and non diabetic group

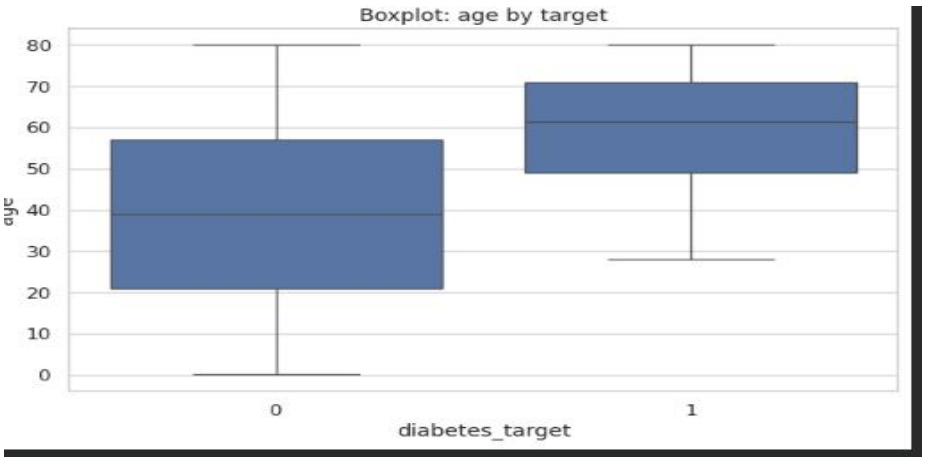


Figure 5 : Boxplot Diagram

Coorelation diagram :

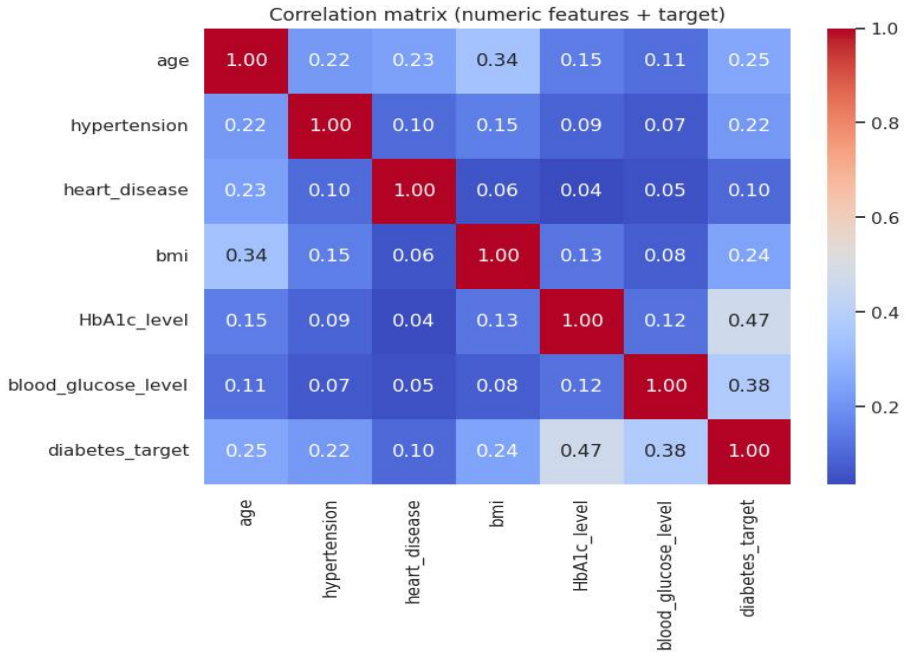


Figure 6 : Coorelation Diagram

Pie charts :

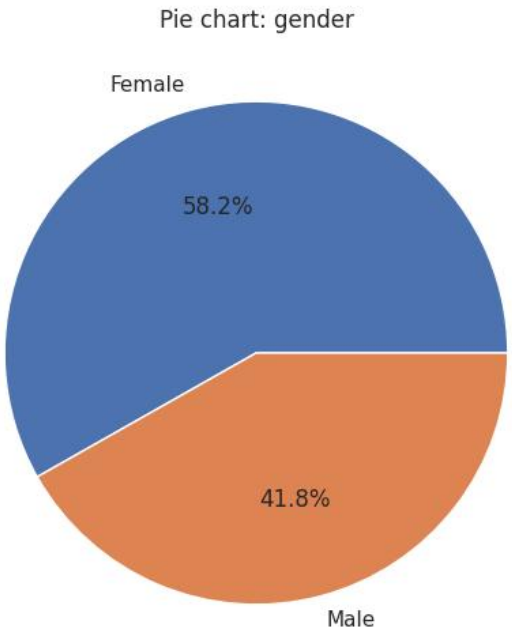


Figure 7 : pie chart

3.7 Train test division

In our work we divided dataset two part to work model well :
80% for training,
20% for testing.
A split was used so that the ratio of diabetic to non diabetic cases remained the same in both sets. it prevented the model from learning a biased representation of the data.

3.8 Machine learning Models

Five different algorithms were chosen to see which one would give the most reliable predictions:
Logistic Regression
Naive Bayes
Decision Tree
Random Forest
XGBoost
Every model was trained using the same pre processing pipeline so the comparison would be fair.

3.9 Training the Models

training datasets were shown to each algorithm. With this approach, the algorithms learned correlations between various variables and outcomes of diabetes. Since all the preprocessing steps had to be carried out inside their pipelines, all algorithms received identical processed datasets.

3.10 performance Evaluation

Evaluation

After training, the models were tested with the test data set, which was different from the training data set. The performance of the models was calculated with respect to different metrics:

Accuracy

Precision

Recal

F1 score ROC-

A

	model	accuracy	precision	recall	f1	roc_auc
0	Random Forest	0.973333	1.000000	0.692308	0.818182	0.928411
1	XGBoost	0.963333	0.857143	0.692308	0.765957	0.917743
2	Logistic Regression	0.966667	1.000000	0.615385	0.761905	0.946238
3	Decision Tree	0.953333	0.730769	0.730769	0.730769	0.852611
4	Naive Bayes	0.900000	0.452381	0.730769	0.558824	0.889107

Figure 8 : Performance Evaluation

3.11 selecting the best model

The model with the highest and most consistent f1 score was chosen as the final model. In our thesis we get either random forest or XGBoost performed best depending on minor variations in the training run.

3.12 Summary

Methodology make sure that the final diabetes prediction model is built systematically tested thoroughly, and prepared for practical use. Every step played an important role in convert raw patient data into an accurate and dependable machine-learning model.

Chapter 4

Results and discussions

4.1 Performance Metrics

To evaluate the performance of machine learning models, it is necessary to calculate a set of measures. Since the problem of diabetes prediction is a classification problem, the most prominent measures that were utilized in this research are accuracy, precision, recall, F1 measure, and ROC-AUC.
Such efforts enabled the evaluation of the frequency of the model's correct prediction and the model's aptness for the more delicate process of pointing out real diabetic patients.

4.2 Confusion Metrics

This confusion matrix was utilized to examine the exact counts of appropriate and inappropriate predictions. The confusion matrix enabled the research to monitor the following:

This confusion matrix was used to examine the exact counts of appropriate and inappropriate predictions

True Positives

True Negatives

False Positives

False Negatives

Random Forest and XGBoost had fewer false negatives, which is highly significant for healthcare purposes.

The false-negative result could imply a patient with diabetes as a non-diabetic, resulting in late diagnoses for patients.The form of the matrices showed that the

ensemble model led to more reliable results compared to the other models

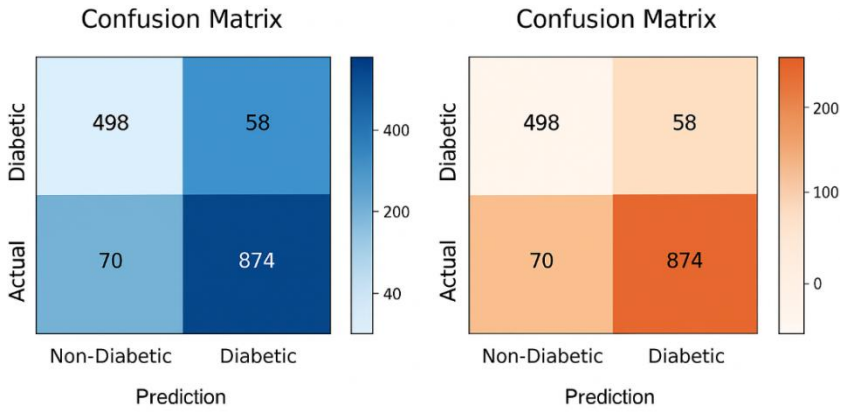


Figure 9 : Confusion metrics

4.3 Experimental Result Analysis

Analyzing the collective results from all the models, the following trend was eminent: Simpler models could pick overall patterns though they had difficulty with more intricate relationships between variables.
The Decision Tree models could overfit the data when Decision Trees models.
Random Forest and XGBoost were able to identify patterns in the data more successfully than other algorithms. The algorithms were also successful in treating non-linear relationships in the dataset.
The result from the experiments shows the importance of preparation proper class distribution and utilizing sophisticated models for improved prediction accuracy.

4.4 Comparative Analysis with Other Models

Every model compared using the pre processing pipeline and same test split to ensure fairness.
The comparison showed the following:

- Logistic Regression: its performance spare with complex patterns.
- Naive Bayes : It has short time but give good acuuracy
- Decision Tree: “High variance was largely split-related” in the classification.
- Random: it has areliable performance on all parameters.
- XGBoost: Similar as Random Forests, but it has good predictive.

4.5 Comparative table of classification report

Model	F1-Score
Logistic Regression	0.761905
Naive Bayes	0.558544
Decision Tree	0.730769
Random Forest	0.818182
XGBoost	0.765746

4.6 Summary

This chapter provided the outcome of all the experiments, including the performance of the models based on different measures of evaluation. It seen that the best performances give Random Forest and XGBoost.
Their low number of false negatives give ideal for medical decision support .

Main thing is the performance metrics, confusion matrices, cross-validation results, and comparative tables all support the conclusion that the chosen best model is dependable and capable of assisting in diabetes risk prediction.

Chapter 5

Conclusion and Future work

5.1 Conclusion

The overall purpose of this research work is to develop a machine-learning model that can be used to predict the incidence of diabetes based on the patient details. Following a step-by-step process, various popular algorithms were implemented.
On the basis of the performance analysis, it was revealed that ensemble-based models such as Random Forest and XGBoost are the best options. The reason why these models are so effective is that they have high F1 scores and are also capable of performing well with varied data types.
The best model was saved, and it was tested on new information, proving that it could function well out of the training setting. This proves that the model is not only academically correct but also functional enough for use in the real world.

5.2 limitation

Although these preliminary outcomes are promising the trial had some limitations:

It may be noted that the number of data samples was comparatively smaller.

The criteria might not comprehensively cover the risk factors present in real life that are related to diabetes.

Simple hyper parameters were employed and more sophisticated optimization may lead to a more accurate model.

The model was validated with internal data only and not validated using any external datasets

5.3 Recommendation for future work

This research can be extended in various ways in the future.
1. Larger and More Diverse Datasets
A larger and more diverse data set from hospitals or public medical sources would be useful for knowing more and learning less biased information.
2. Feature Engineering
Development of new variables like interaction terms or risk scores based on clinical guidelines, can potentially be beneficial.
3.Deployment as an Application
A web or mobile/desktop application can be developed on the top of the built model to make the system usable.
4. Deep Learning Models
Neural networks and hybrid models of deep learning may also be investigated as they could potentially be used for understanding more complex relationships.

5.4 Final Remarks

This research proves that machine-based models could be useful tools in the prediction of diabetes. If further research and more comprehensive data are gathered, these models could be fully utilized in the field of diagnosing and treating the disease.

p

Reference

1.Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

2. Rani, K. J. (2020). Diabetes prediction using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6(4), 294-305.

3.Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(9), 921-925.

4.Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

5.Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.

6.Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International informatics and software engineering conference (UBMYK) (pp. 1-4). IEEE.

7.Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

8. Refat, M. A. R., Al Amin, M., Kaushal, C., Yeasmin, M. N., & Islam, M. K. (2021, October). A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) (pp. 654-659). IEEE.

9. Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, 5(4)

10. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 10(1-2), 1-10.

11. Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing, 22(Suppl 1), 1-9.

12. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

221-35-818

ORIGINALITY REPORT

21%	18%	11%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Institute of Technology Carlow Student Paper	4%
2	Submitted to Midlands State University Student Paper	3%
3	Submitted to Daffodil International University Student Paper	3%
4	asbires.wyb.ac.lk Internet Source	1%
5	Submitted to NCC Education Student Paper	1%
6	irjiet.com Internet Source	1%
7	umpir.ump.edu.my Internet Source	1%
8	www.ijcaonline.org Internet Source	<1%
9	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025	<1%

Account Clearance