İSTANBUL NİŞANTAŞI UNIVERSITY FACULTY OF ENGINEERING ARCHITECTURE

SOFTWARE ENGINEERING DEPARTMENT GRADUATION PROJECT REPORT

Diabetes Disease Prediction System

HAIFA ALJUNDI

Student ID NO: 20212022209

DEPARTMENT: SOFTWARE ENGINEERING

PROJECT ADVISOR: SIBEL BORAN, Ph.D.

COURSE NAME / CODE: GRADUATION PROJECT / ESOF433

ACADEMIC YEAR / SEMESTER: 2024-2025 / FALL

PRESENTATION DATE:

Table of Contents

TABLE OF CONTENTS

ABSTRACT

Diabetes is a growing global health concern, affecting millions of individuals and placing immense pressure on healthcare systems. Early and accurate prediction of diabetes can significantly improve patient outcomes and reduce the burden on healthcare providers. This study explores the application of six machine learning algorithms—Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Gradient Boosting Decision Trees, and XGBoost (Extreme Gradient Boosting). to predict diabetes based on key health indicators. A comprehensive dataset is preprocessed and analyzed to ensure high-quality input data, and the models are evaluated using metrics such as accuracy, precision, recall, and F1-score. The comparative analysis of these algorithms highlights their strengths and weaknesses in diabetes prediction. This work aims to provide a scalable and efficient predictive model for aiding healthcare professionals in identifying individuals at risk of developing diabetes, contributing to better preventive and diagnostic strategies.

I. INTRODUCTION

Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels due to the body's inability to produce or effectively use insulin. According to the International Diabetes Federation, the prevalence of diabetes is expected to rise significantly in the coming decades, making early detection and intervention essential for reducing complications and improving quality of life. Machine learning has emerged as a transformative tool in the field of predictive healthcare, enabling the analysis of large and complex datasets to identify patterns and risk factors. This project leverages six machine learning algorithms—Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Gradient Boosting Classifier, and XGBoost (Extreme Gradient Boosting)—to develop and compare predictive models for diabetes detection.

Diabetes is one of the most common diseases worldwide, with the World Health Organization reporting an increase in the number of affected individuals from 108 million in 1980 to 422 million in 2014. It is responsible for 1.6 million deaths annually and contributes to complications such as blindness, kidney failure, heart disease, stroke, and lower limb amputations (World Health Organization, 2021). Early diagnosis of diabetes is crucial in preventing these severe outcomes. Symptoms of diabetes can appear suddenly and may take several years to be detected. Early diagnosis reduces the risks of blindness, kidney failure, heart attack, stroke, and death. Common symptoms include:

- Feeling very thirsty
- Increased urination
- Blurred vision
- Fatigue
- Unintentional weight loss

The objective of this study is twofold: first, to develop accurate and interpretable models for diabetes prediction using a variety of machine learning techniques, and second, to identify the most significant features contributing to the risk of diabetes. The dataset used in this study includes health-related attributes such as age, body mass index (BMI), glucose levels, blood pressure, and family history of diabetes. A rigorous preprocessing pipeline ensures that the data is clean, normalized, and ready for analysis. By comparing the performance of different algorithms, this study aims to determine the best approach for predicting diabetes risk and improving healthcare outcomes.

Through a comparative analysis of the models, we evaluate their performance using standard metrics, including accuracy, precision, recall, and F1-score. The results of this study aim to guide future research and practical implementations in predictive healthcare, contributing to the early diagnosis and management of diabetes.

II. RELATED WORK

In 2020, The Procedia Computer Science Journal published a research paper titled ”Prediction of Type 2 Diabetes using Machine Learning Classification Methods” by Neha Prerna Tigga and Shruti Garg from the Department of Computer Science and Engineering at Birla Institute of Technology in India. The paper applied a number of machine learning algorithms to predict the probability of Type 2 diabetes in individuals by considering their lifestyle and familial background. The authors collected data from 952 individuals through a questionnaire consisting of 18 questions. Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Gradient Boosting Decision Trees, and XGBoost (Extreme Gradient Boosting) are applied machine learning algorithms. The Random Forest Classifier was the most accurate algorithm for their dataset and the Pima Indian Diabetes database. The paper concludes that machine learning algorithms can be used to accurately predict the risk of diabetes and that individuals can use this information to self-assess their risk and take preventive action.

In 2021, a paper entitled ”Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction” was presented at the 7th International Conference on Advanced Computing and Communication Systems (ICACCS). The paper discusses the application of machine learning algorithms for the early prediction of diabetes to prevent severe consequences. The paper used two machine learning algorithms: Support Vector Machine (SVM) and Random Forest (RF), and feature selection techniques to identify the most influential factors for prediction. It also applied Principle Component Analysis (PCA) for dimensionality reduction. Various machine learning algorithms were discussed, including decision trees, Gradient Boosting Decision Trees, and XGBoost (Extreme Gradient Boosting), highlighting their strengths and applications in disease prediction.

In 2022, the Mathematical Biosciences and Engineering (MBE) journal published a research paper titled ”Risk Prediction of Diabetes and Pre-diabetes Based on Physical Examination Data”. The paper discusses the importance of early diagnosis and prediction of diabetes in order to prevent and control the disease and its complications. The paper collected physical examination data from the Beijing, China, Physical Examination Center and divided the population into three groups: normal fasting plasma glucose (NFG), mildly impaired fasting plasma glucose (IFG), and type 2 diabetes mellitus(T2DM). Four classification models were constructed to distinguish between the three groups. including eXtreme Gradient Boosting (XGBoost), random forest (RF), logistic regression (LR), and fully connected neural networks (FCN). Additionally, binary classification models were established to discriminate between the three groups.

All previous studies used supervised machine learning algorithms, which are commonly used in disease prediction for several reasons:

1) They require labeled training data.
2) Provide insights into feature importance.

3) Aim to achieve high prediction accuracy.
4) Handle both classification and regression tasks.
5) Offer interpretability in some cases.

On the other side, the Prediction of disease in general and diabetes, in particular, requires historical data with known disease outcomes to train the model. It also requires identi- fying which features are most influential in predicting the disease outcome. predicting the disease outcome must be high accuracy to be reliable. When the goal is to assign a discrete label to a data instance, such as classifying a patient as either having a disease or not. And when the target variable is continuous, such as predicting disease severity or estimating the progression of a disease, Regression al- gorithms are used. Providing interpretability is sometimes required, allowing researchers and clinicians to understand the decision-making process of the model.

In the first and the second papers, the highest accuracy was achieved by Random Forest, and in the last one, the best accuracy was achieved by XGBoost.

III. DIABETES PREDICTION ALGORITHMS

We used six supervised machine learning algorithms, which were chosen from a group of algorithms used in the Related work, including Random Forest and XGBoost. In the end, the accuracy was compared to find out which of them will achieve the highest accuracy on our datasets.

A. Random Forest Classifier (RF)

Random Forest is one of the most popular and effective classification algorithms. It operates by building an ensemble of decision trees and aggregating their predictions through majority voting for classification tasks. In this project, RF was utilized to classify the data and predict outcomes. Its ability to handle large datasets with higher dimensionality and its resistance to overfitting made it a valuable choice for our analysis. The RF algorithm demonstrated robust performance in handling classification problems effectively.

B. Gradient Boosting Classifier

Gradient Boosting is a machine learning technique that builds models incrementally from decision trees. Each subsequent tree minimizes the residual errors of the prior trees, optimizing the loss function iteratively. In this project, the Gradient Boosting Classifier was employed to

enhance the predictive accuracy. Its ability to focus on correcting errors from previous iterations contributed significantly to improved model performance, making it an excellent option for complex datasets.

C. Decision Tree Classifier (DT)

Decision Trees are a widely used algorithm due to their simplicity and interpretability. They are capable of handling both classification and regression tasks. A decision tree consists of nodes, where each decision-node represents a test, and the leaf-nodes represent the outcomes. In this project, DT was implemented as a baseline model to assess its performance in prediction tasks. Despite being straightforward, the algorithm effectively captured the patterns in the data.

D. Logistic Regression (LG)

Logistic Regression is a statistical model commonly used for binary classification tasks. It calculates the probability of an instance belonging to a particular class using a logistic function. LG was applied in this project to predict outcomes based on input features. Its simplicity, efficiency, and effectiveness in predicting probabilities made it a useful algorithm for comparison with more complex models.

E. Support Vector Machine (SVM)

Support Vector Machines are supervised learning models designed for classification and regression tasks. They create hyperplanes to separate classes in a high-dimensional space, maximizing the margin between data points of different categories. SVM was employed in this project to classify data and predict outcomes. Its ability to perform well on smaller datasets and maintain accuracy with non-linear relationships made it a valuable component of our modeling efforts.

F. Extreme Gradient Boosting (XGBoost)

XGBoost is a commonly used machine learning algorithm for regression and classification tasks. It is based on the gra- dient boosting framework, which involves combining several weak models (such as decision trees) to generate a strong model.The XGBoost algorithm emerged from a research project conducted at the University of Washington and was introduced in 2016 by

Tianqi Chen and Carlos Guestrin. XGBoost works by recursively training decision trees on the residuals (or errors) of previous trees, with the goal of minimizing the loss function. XGBoost also includes several advanced features that make it particularly effective, such as:

- **Decision Trees:** XGBoost uses decision trees as the weak models in the ensemble, and each tree is trained to predict the residual error of the previous trees.
- **Loss Function:** XGBoost supports several loss func- tions for different types of problems. The loss function is used to calculate the error of the model predictions and adjust the weights of the training samples.
- **Handling Missing Values:** XGBoost can handle missing values in the data by using a technique called ”sparsity-aware split finding”.

XGBoost is a powerful and flexible algorithm that can be used for a wide range of machine-learning tasks.

IV. MATERIALS AND METHODS OF DIABETES PREDICTION

The methodology starts with the problem statement followed by the preparation of a suitable data set. Once, data set is prepared, it has to undergo the pre-processing steps (normalization and removal of null values). In the next step, the informational index is separated into preparing and test sets. The training data set is used to train the models and the tests are run in the test set.

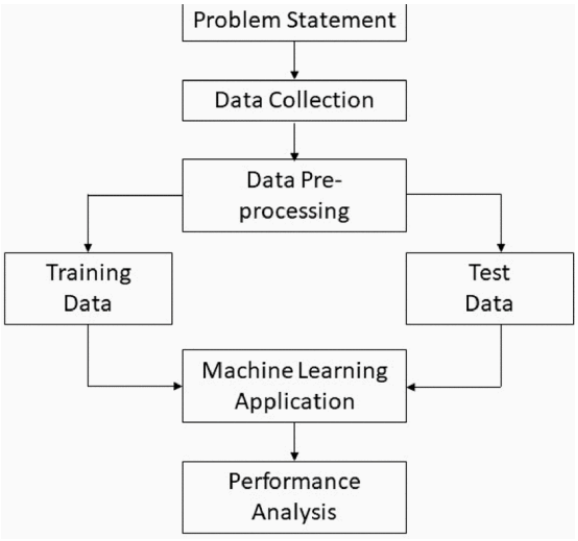


Figure 1: Methdology

A. Data set

The diabetes dataset was employed, determines whether a person has diabetes or not.

- **Dataset**: This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage [Fig2].

Pregnancies	Glucose	BloodPressur	SkinThicknes	Insulin	BMI	DiabetesPed	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1

Figure2: Dataset and EDA

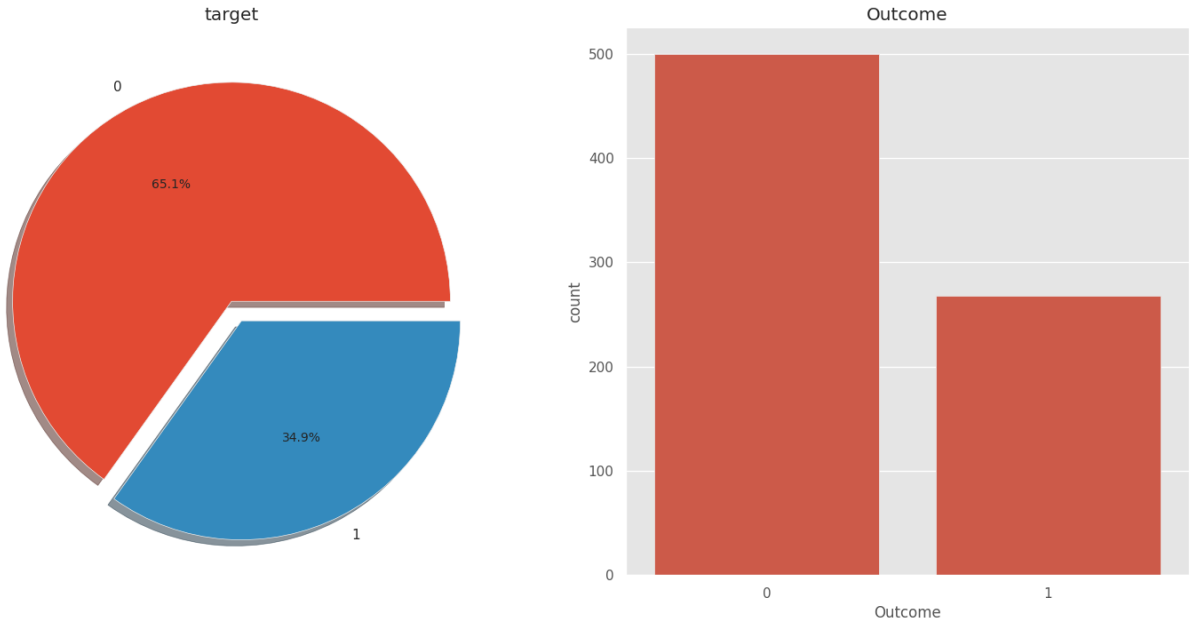


Figure 3: Dataset distribution 0 - Healthy, 1 - Diabetic

The dataset mentioned above has eight features which are defined in Table 1.

Features	Description
Pregnancies	Number of Pregnancies patients had earlier.
Glucose	Glucose level present in the patient.
Blood Pressure	Recorded blood pressure level at that particular time.
Skin Thickness	Skin thickness level of the patient.
Insulin	Amount of Insulin present in the body.
BMI	Body Mass Index of the individual.
Diabetes Pedigree Function	Family history of Diabetes disease.
Age	Age of an individual.

Table 1: LIST OF FEATURES PRESENT IN THE DATASET

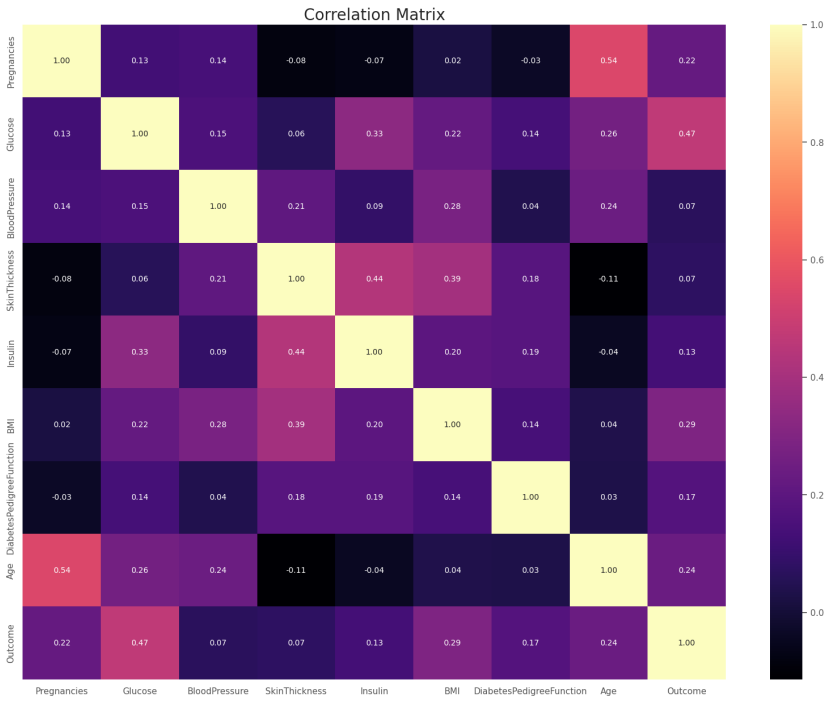


Figure 4: Correlation matrix of the dataset

The detailed information about [Fig 4] each attribute or features is discussed below.

- **Pregnancies:** Those who develop gestational diabetes are at higher risk of developing type 2 diabetes later in life. The subjects with more number of pregnancies have a higher risk of developing diabetes.

- **Glucose:** The subjects were given an oral glucose test, whereby, they were administered glucose and a reading of their plasma glucose concentration was taken after 2 hours. The subjects with higher levels of glucose concentration after 2 hours have a higher risk of developing diabetes.

- **Blood pressure:** Having blood pressure over 140/90 mmHg of Mercury are linked to having increased risk of developing diabetes. Although, certain sub- jects having diastolic blood pressure ¿70 mmHg may develop diabetes.

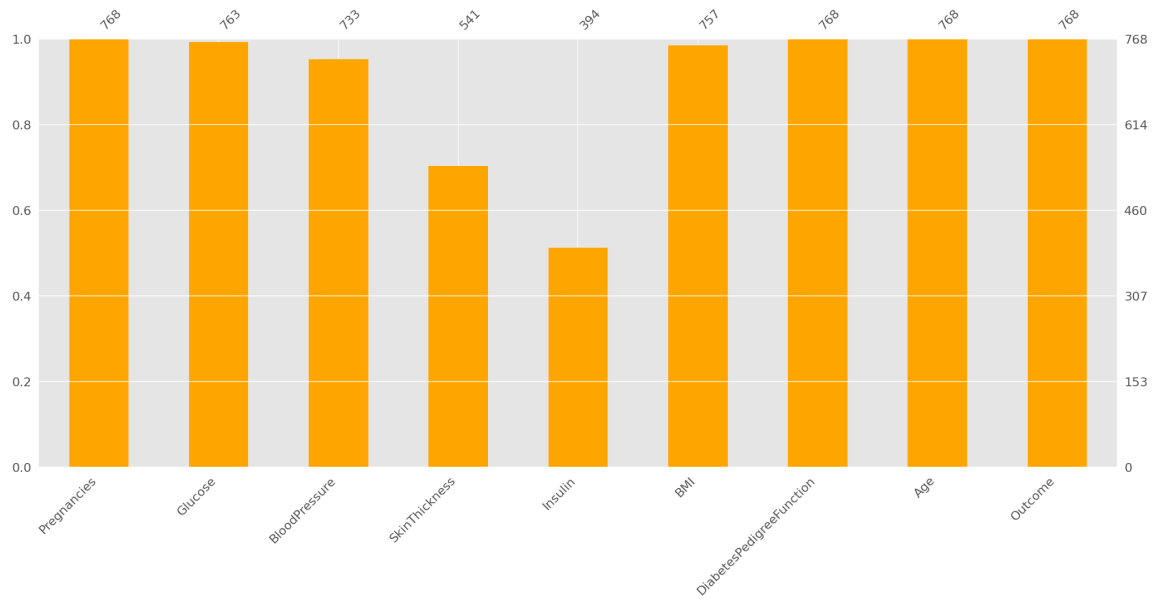
- **Skin Thickness:** Skin thickness is primarily deter- mined by collagen content and is increased in the case of insulin dependent diabetic patients. The subjects' tricep skin fold were measured and results showed that having a skin thickness of 30mm of greater are at a higher risk.

- **Insulin:** Normal insulin levels after 2 hours of glu- cose administration is 16-166 mlU/L. Subjects having lower or higher levels than said value are at a higher risk.

- **Body Mass Index (BMI):** Subjects having a BMI over 25 have a relatively high risk in having diabetes.

- **Diabetes Pedigree Function:** The diabetes pedigree function provides "a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject." The higher the DPF, the more likely it is for a subject to be diabetic.

- **Age:** Diabetes is prevalent in any age group, but is commonly found in middle aged adults (45 onwards). Taking that into consideration, subjects with in the higher age group have a higher expectancy of diabetes.



B. Data Processing

To conduct exploratory statistical analysis and train models effectively, data processing is crucial. The extent of feature analysis and the quality of predictive results depend on the level of data processing during both the training and testing phases.

1. Finding Missing Values from the Dataset:

- After checking the dataset, no missing values appeared

2. Making Categorical Variables in the Numeric Format:

- The categorical variables have been converted into a numerical representation to make the data suitable for analysis using machine learning or statistical models. Additionally, that can help to uncover patterns and relationships in the data that may not be immediately apparent when the data is in its original categorical form.

3. Feature Scaling:

- Some of the data have been normalized to ensure that the values of different variables are on a similar scale so that no variable dominates over the others. Especially for some machine learning algorithms that use distance-based methods to calculate

similarities between data points. This processing was applied to the dataset for a number of categorical variables (Age, Glucose, blood pressure, SkinThickness, and Insulin).

4. Segmentation of the Dataset to Tow Sub-Datasets:

- Along with the feature, the dataset also has 2 label (0- No and 1- Yes) which is the outcome of the diabetes disease.

5. Training and Validation Datasets:

- To effectively train machine learning models, it is neces- sary to partition the data into training and testing sets. To accomplish this, the datasets were split into 80% for training and 20% for testing purposes. by using The train test split function, which enables the evaluation of a model's per- formance on previously unseen data. By randomly splitting a dataset into a training subset and a testing subset, the function allows the model to be trained on a subset of the data and tested on a completely independent subset. This helps to prevent overfitting. After the split, The model was then examined by subsets testing to ensure that it was accurate. The testing accuracy represents the procedure's overall testing accuracy as an average.

Dataset	Total	Percentage
Training	614	80%
Testing	154	20%

Table 2: SPLITTING THE DATASET OF MODEL

C. Setting Classification Metrics

To classify disease and get a prediction result, we need to set a few metrics which will help us in predicting the Diabetes disease. Since we are using scikit-learn (Sklearn) machine learning library [8] for our experiment, we have used confusion matrix as the classification measure met- rics. All the used metrics, i.e. *Precision, Recall, F1-Score and Accuracy* in our analysis, are listed below.

- **Precision (P)** is defined as the number of true posi- tives (Tp) over the number of true positives plus the number of false positives (Fp). Mathematically,

$$\mathrm{P} = \frac{\mathrm{T}p}{\mathrm{T}p + \mathrm{F}p}$$

- **Recall (R)** is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

$$R = \frac{\mathrm{T}p}{\mathrm{T}p + \mathrm{F}n}$$

- **F1-Score (F1)** is defined as the harmonic mean of precision and recall.

$$F1 = 2 * \frac{P * R}{P + R}$$

- **Accuracy (A)** is defined as follows.

$$A = \frac{Tp \ + \ Tn}{Tp \ + Tn \ + Fp \ + Fn}$$

D. Applying Machine Learning Algorithms

For our experiment, we will perform 6 supervised machine algorithms on the pre-processed dataset. The algorithms which we used are as follows-

1) Random Forest Classifier (RF)
2) Gradient Boosting Classifier
3) Decision Tree Classifier (DT)
4) Logistic Regression (LG)
5) Support Vector Machine (SVM)
6) XgBoost(Extreme Gradient Boosting)

E. System Design

1. Responsive Web Interface

- Developed using HTML, CSS , and JavaScript the website allows users to input health data, view predictions, and interact with additional tools.
- A BMI calculator (Figure 12) provides instant feedback on weight status based on user inputs.

2. Page for Diabetes Prediction

(Figure 6)

- A dedicated web page was created to handle diabetes predictions.
- The page interacts with the backend to dynamically load the trained machine learning model and process user inputs for real-time inference.
- This page also displays the prediction result, along with relevant metrics, ensuring user-friendly navigation and clarity.

3. Backend Integration

- Implemented with Flask, the backend handles requests, processes data, and serves predictions via RESTful APIs.
- MongoDB Atlas is used as the database to securely store user data. Two key collections are used:
  1. **Prediction Data**: Stores prediction-related information, including user inputs and the corresponding prediction result. (Figure 7)
  2. **User Email Data**: Stores user email addresses for personalization and account management. (Figure 8)
- The trained machine learning model, written in Python, is deployed to Heroku for real-time predictions.

4. Custom Domain

- The web application is hosted on **Heroku** and linked to a custom domain, **www.haifa.engineer**, for improved branding and accessibility.
- The deployment ensures scalability and seamless access for users worldwide.

5. Health Record Assistant by IBM watsonx:

(Figure 9)

- This allows the Retrieval Augmented Generation (RAG) to retrieves and summarizes reports in simple formats for patients and medical practitioners to take action, alongside a chatbot.
- Integrated to provide users with credible medical information.
- The assistant can also fetches data from trusted sources, answering patient queries about symptoms, treatments, and lifestyle modifications.

Figure 5: Website Home Page

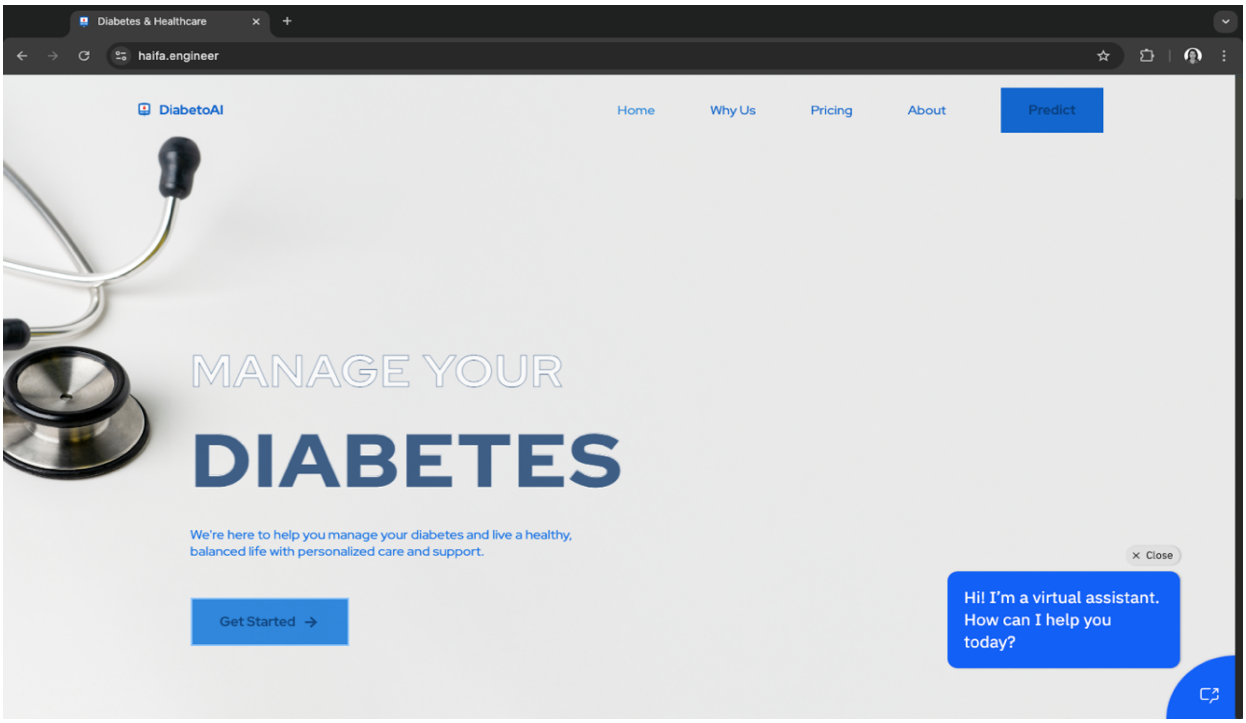


Figure 6: Diabetes Prediction Page

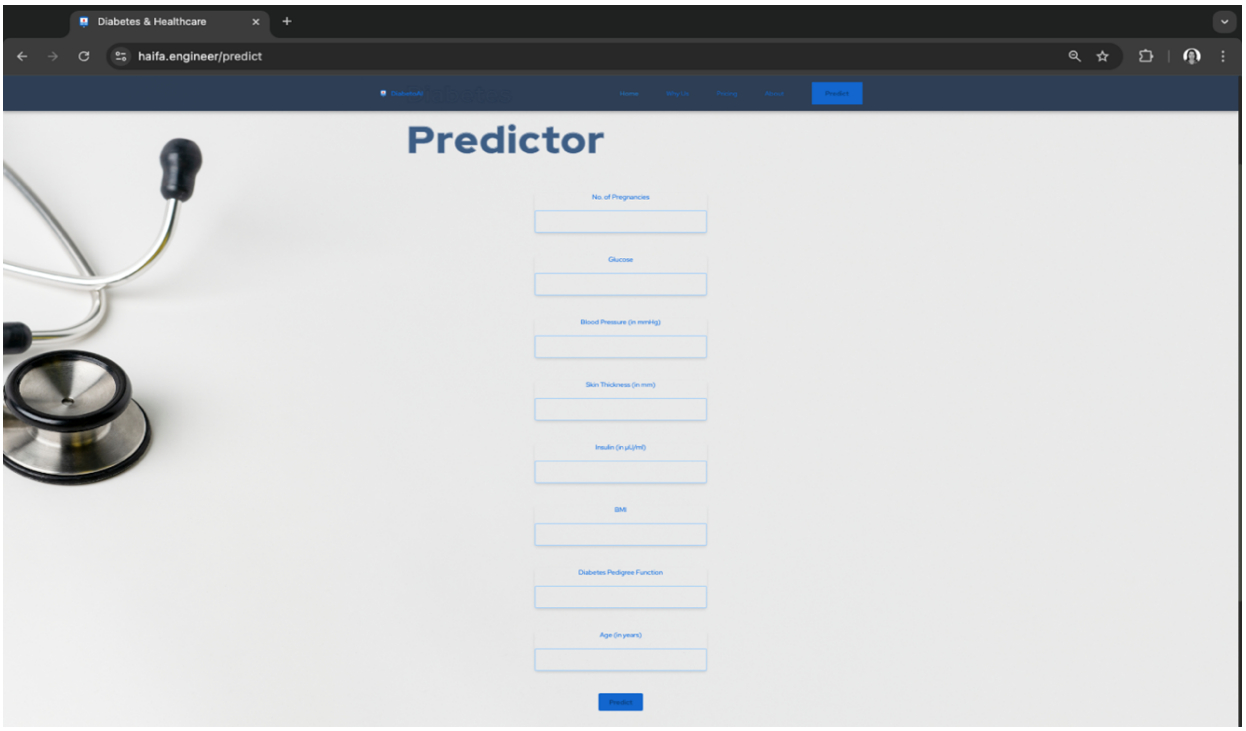


Figure 7: Prediction Data in MongoDB Atlas

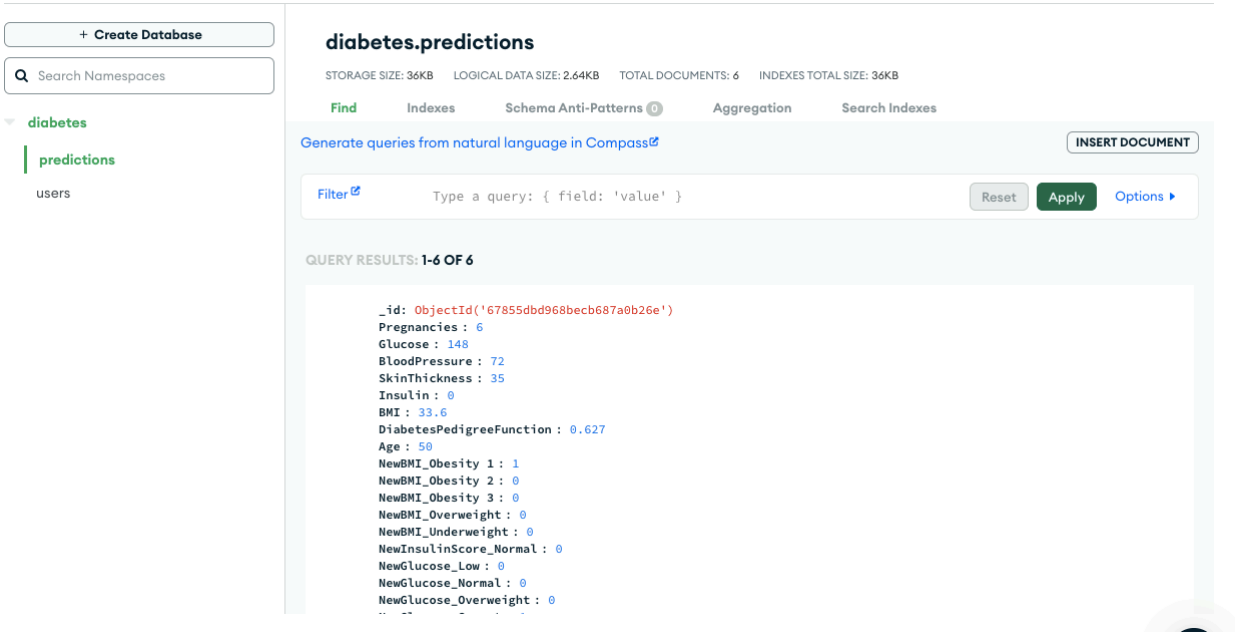


Figure 8:User Email Data in MongoDB Atlas

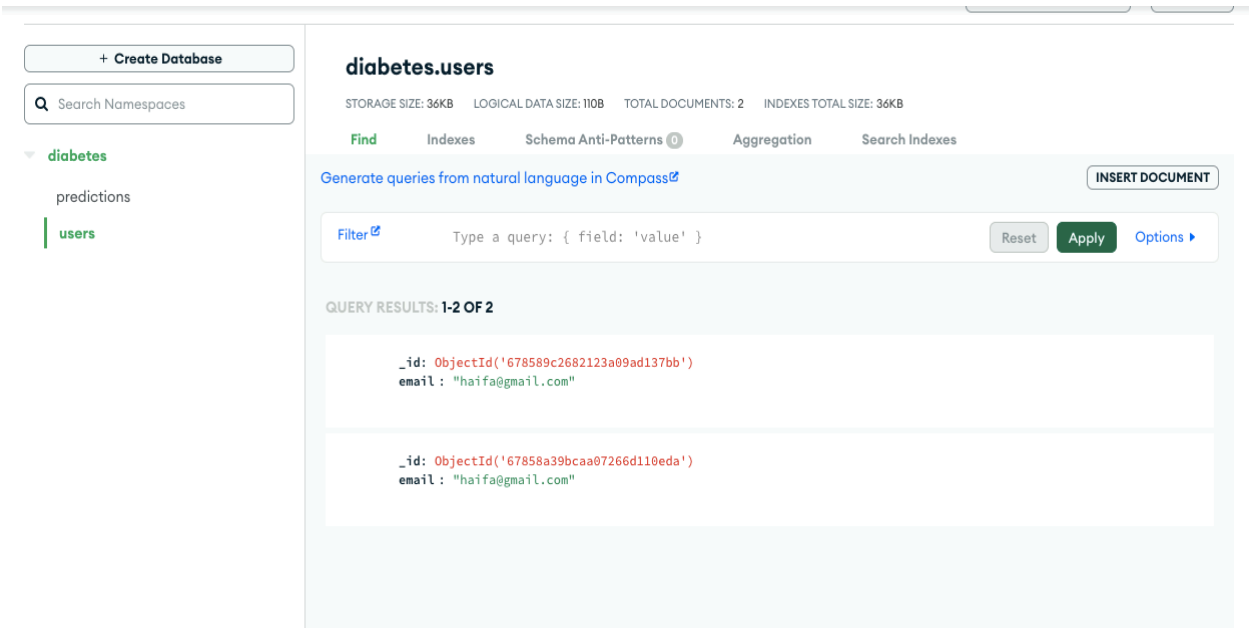


Figure 9: IBM Watsonx Health Record Assistant

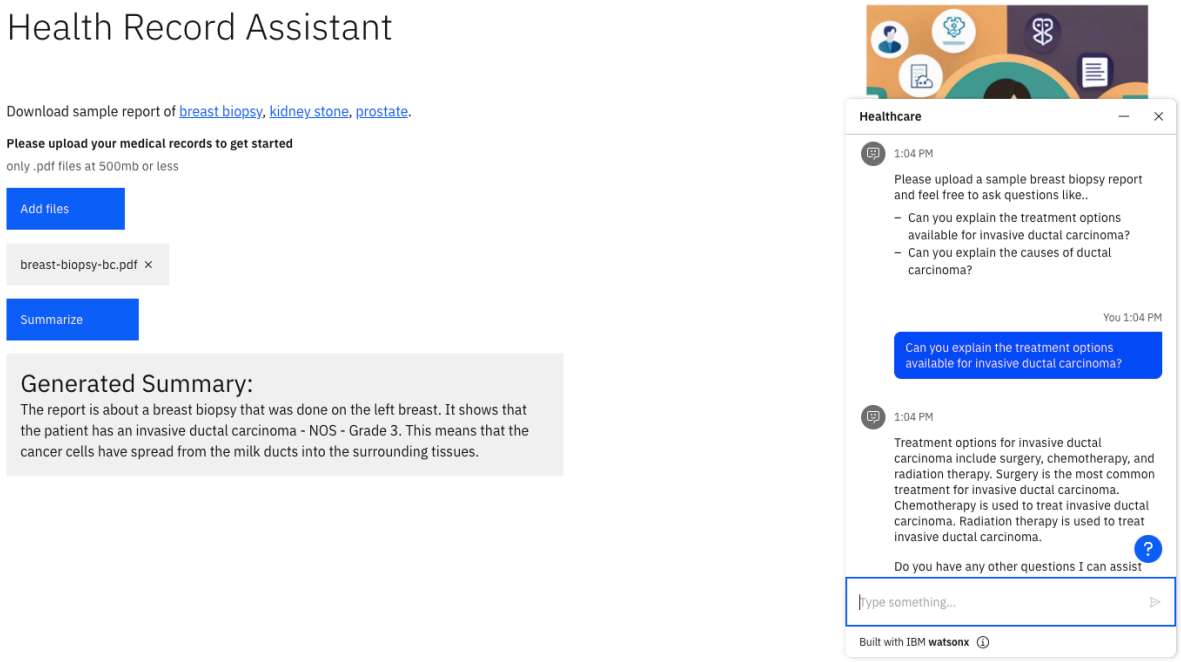


Figure 10: Diabetes healthcare Services

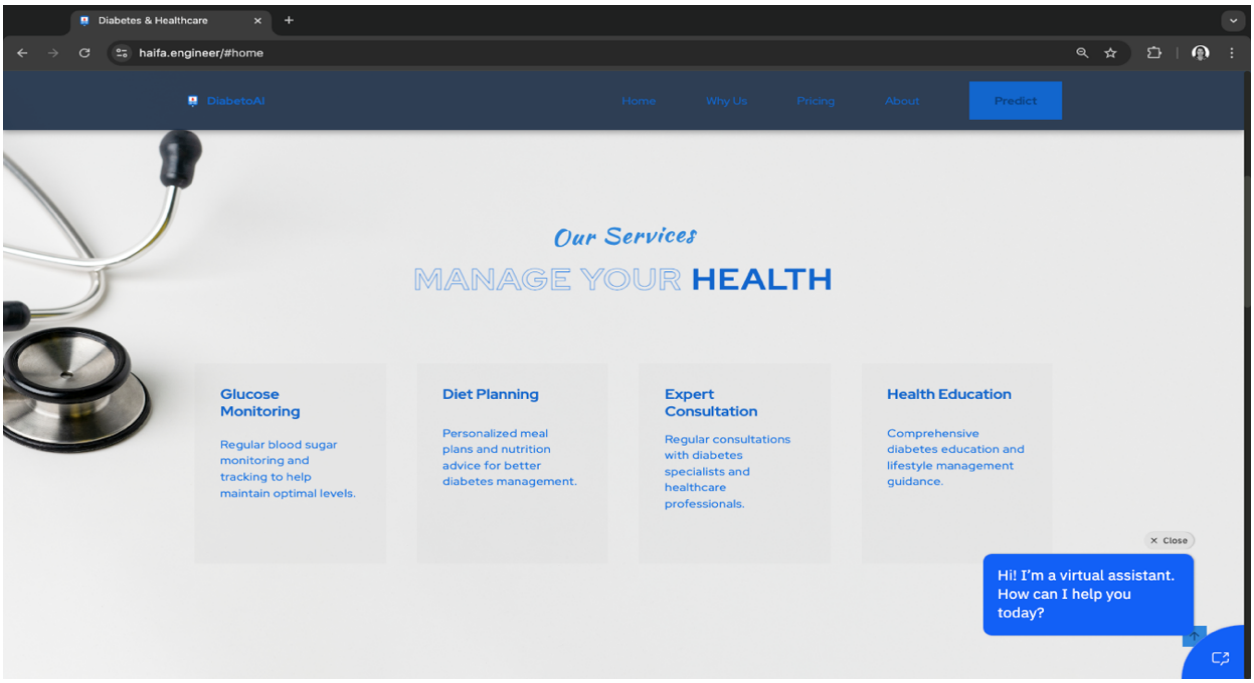


Figure 11: Pricing Plan Page

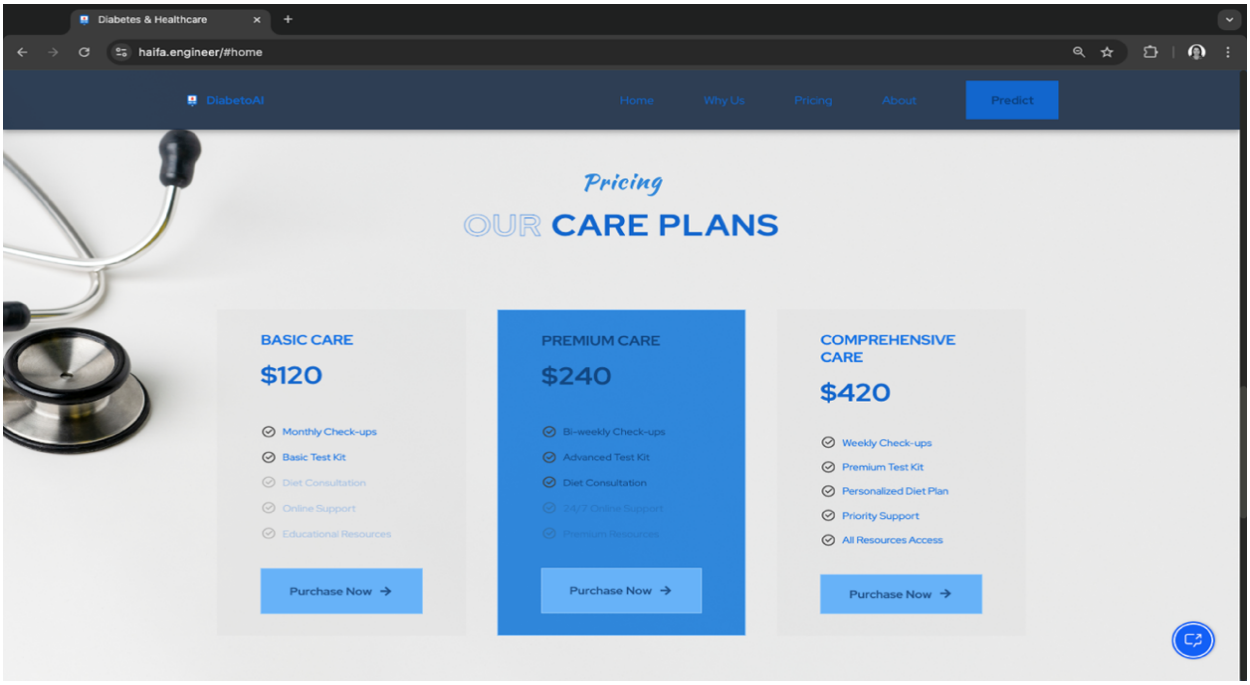
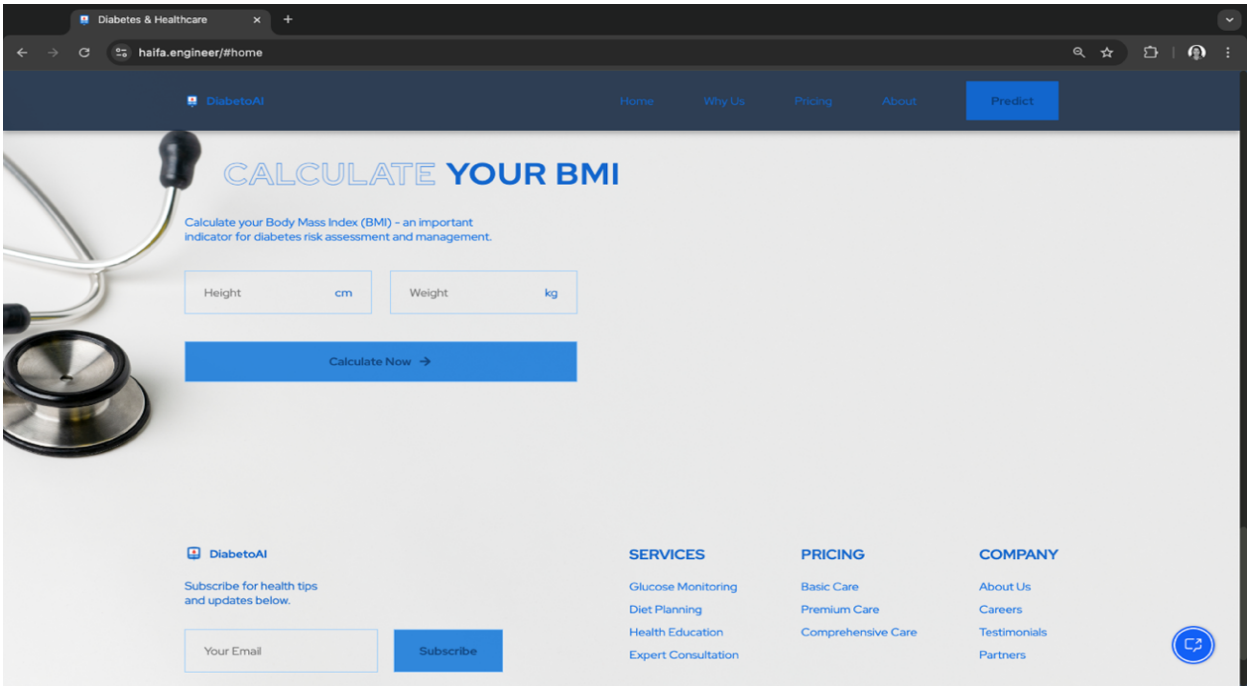


Figure 12: BMI calculator

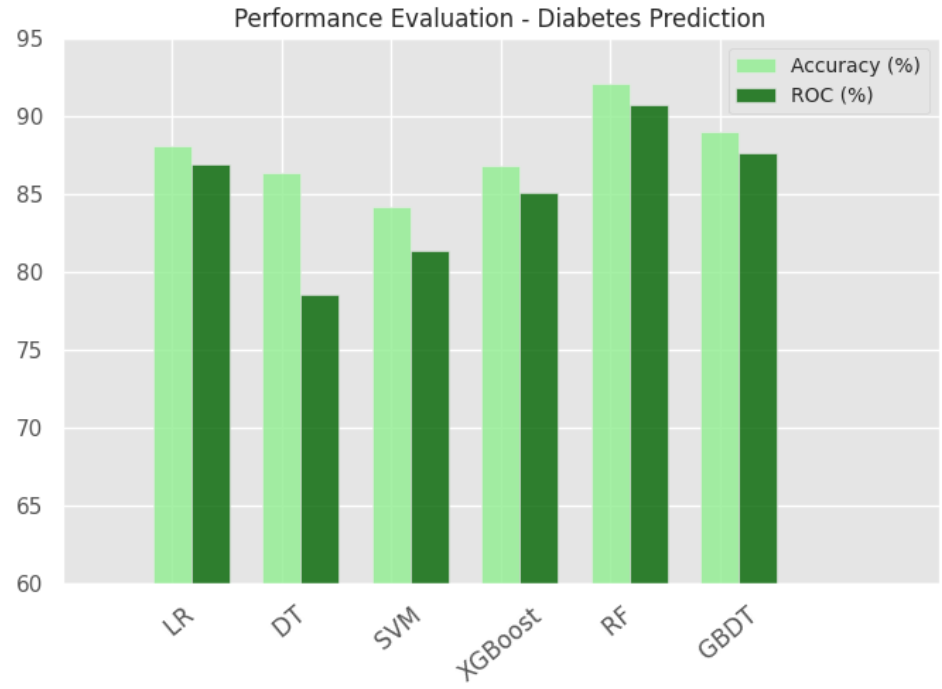


V. EXPERIMENTAL RESULTS

To perform our experiment, we split the dataset into training and testing set with the ratio of 80:20 as in (Table2) , respectively. We used python 3.12 version for applying all of our machine learning classifiers.

Classifier	Precision (0/1)	Recall (0/1)	F1-Score (0/1)	Accuracy	Macro Avg F1	Weighted Avg F1
Logistic Regression	0.91 / 0.84	0.91 / 0.83	0.91 / 0.83	0.88	0.87	0.88
Decision Tree Classifier	0.90 / 0.80	0.89 / 0.81	0.89 / 0.81	0.86	0.85	0.86
Random Forest Classifier	0.94 / 0.88	0.93 / 0.90	0.94 / 0.89	0.92	0.91	0.92
GBDT	0.91 / 0.86	0.93 / 0.83	0.92 / 0.84	0.89	0.88	0.89
XGBoost	0.89 / 0.83	0.91 / 0.79	0.90 / 0.81	0.87	0.85	0.87
SVM	0.85 / 0.82	0.91 / 0.72	0.88 / 0.76	0.84	0.82	0.84

Table 3: PERFORMANCE METRICS OF DIFFERENT CLASSIFIERS MODELS



The evaluation of various classifiers reveals that the **Random Forest Classifier** achieved the highest accuracy at 92%, making it the most reliable model for the given dataset. **Gradient Boosting Decision Trees (GBDT)** and **Logistic Regression** followed closely with accuracies of 89% and 88%, respectively, showcasing their effectiveness in balancing performance across classes. While **XGBoost** and **Decision Tree Classifier** demonstrated moderate accuracy scores of 87% and 86%, the **Support Vector Machine (SVM)**, with an accuracy of 84%, lagged slightly behind. Despite its lower overall accuracy, SVM may still be useful in scenarios requiring simpler margin-based classification, though it struggles with minority class predictions. These results highlight Random Forest's robust performance and the competitive nature of GBDT and Logistic Regression as alternative models.

VI. WEBSITE PREDICTION RESULT

After processing the input data, the prediction results are displayed directly on the page. These results include:

- **Binary classification**: For instance, "Diabetes Detected: Yes" or "No."

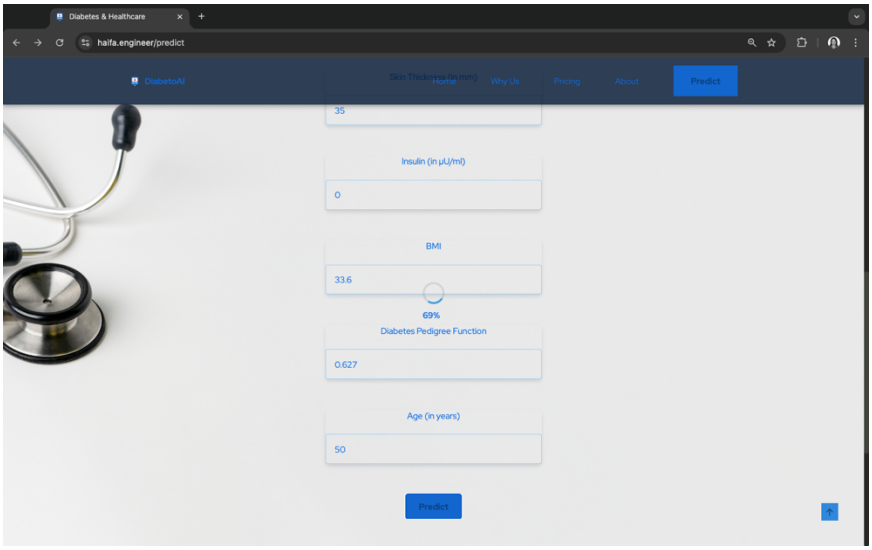
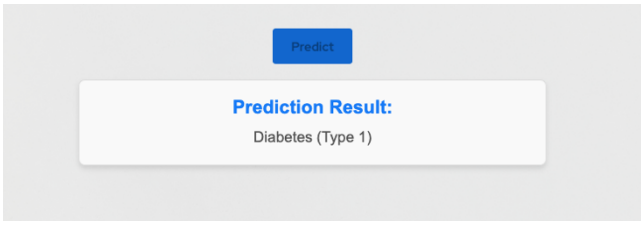


Figure 13: Prediction Result

Example Output:

- **Prediction Result**: "Diabetes (Type 1)"



This result format ensures that users can interpret the output easily and take informed action. **Figure 13** illustrates the layout and functionality of the Diabetes Prediction page, highlighting the seamless integration of the backend and model inference system.

VII. CONCLUSION

One of the critical challenges in predictive analytics is ensuring the accuracy and reliability of models in identifying specific outcomes. In this study, we systematically evaluated multiple classifiers to design a robust model for predictive tasks. Using a balanced dataset, we achieved the highest accuracy of 92% with the **Random Forest Classifier**, demonstrating its superior performance. **Gradient Boosting Decision Trees (GBDT)** and **Logistic Regression** also showed competitive results, with accuracies of 89% and 88%, respectively, indicating their effectiveness for similar applications. While **XGBoost** and **Decision Tree Classifier** achieved moderate accuracies of 87% and 86%, the **Support Vector Machine (SVM)** recorded the lowest accuracy at 84%, highlighting its limitations in this context. These results affirm the adequacy of the models in handling the dataset and predicting outcomes with high precision and recall.

In the future, we aim to enhance our approach by collecting more domain-specific datasets, collaborating with industry experts, and exploring advanced Machine Learning and Deep Learning techniques. By integrating these improvements, we hope to develop systems capable of achieving even higher accuracy and addressing more complex predictive challenges effectively.

ACKNOWLEDGEMENTS

This project was supported by IBM for Startups as part of their initiative to sponsor healthcare startups focused on leveraging cutting-edge technologies for improving health outcomes. IBM's sponsorship provided valuable resources, tools, and expertise that contributed significantly to the development and deployment of the diabetes prediction system.

REFERENCES

1) World Health Organization.(April 2023). diabetes[Online] Available: https://www.who.int/ar/news-room/fact-sheets/detail/diabetes.

2) Kumar, R., Saha, P., Kumar, Y., Sahana, S., Dubey, A., & Prakash, O. (2020). A Review on Diabetes Mellitus: Type1 & Type2. World Journal of Pharmacy and Pharmaceutical Sciences, 9(10), 838-850.

3) A.MirandS.N.Dhage,"Diabetesdiseasepredictionusingmachine learning on big data of healthcare," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–6.

4) Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918308548

5) Sivaranjani,S.,Ananya,S.,Aravinth,J.,&Karthika,R.(2021,March). Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 141-146). IEEE.

6) Han, Y. M., Yang, H., Huang, Q. L., Sun, Z. J., Li, M. L., Zhang, J. B., ... & Lin, H. (2022). Risk prediction of diabetes and pre-diabetes based on physical examination data. Math Biosci Eng, 19, 3597-608.

7) Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

8) Ali,Z.A.,Abduljabbar,Z.H.,Taher,H.A.,Sallow, A.B.,Almufti,S. M. (2023). Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. Academic Journal of Nawroz University, 12(2), 320-334.

9) Chaki, J., Ganesh, S. T., Cidham, S. K., Theertan, S. A. (2022). Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University-Computer and Information Sciences, 34(6), 3204- 3225.