UNIVERSITY OF PIRAEUS

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

DEPARTMENT OF DIGITAL SYSTEMS

POSTGRADUATE PROGRAM

“INFORMATION SYSTEMS AND SERVICES”

MSc Thesis

by

Makris Panagiotis

Submitted
in partial fulfillment of the requirements for the Master’s degree in the specialization «ΜΔΑ/ΠΠΣ/ΠΔ» of the
Postgraduate Program “Information Systems and Services”
at the
UNIVERSITY OF PIRAEUS
July 2024

Thesis Supervisor: Ilias Maglogiannis
Title: Analysis of medical data for diabetes for the creation of predictive indicators

Author Makris Panagiotis

Acknowledgments

I would like to express my deep gratitude and appreciation to all those who have contributed to the completion of this thesis. First and foremost, I am profoundly thankful to my thesis advisor ILIA MAGLOGIANNI, whose invaluable guidance, and insights played a pivotal role in shaping this research. I would also like to acknowledge the assistance of my colleagues and fellow researchers who provided a supportive and stimulating academic environment. Their discussions, shared resources, and camaraderie have greatly enriched my academic journey.

Finally, I want to express my appreciation to all the contributions made by previous researchers in this field, whose sources of data and reports made this research possible.

Table of Contents

Abstract

The advancement of technology and data science has provided significant assistance in the field of medicine, particularly in the timely and effective diagnosis and treatment of diseases. This thesis aims to study the diagnosis of diabetes using machine learning techniques. More specifically, it constitutes a comparative study of machine learning algorithms to find the one that offers the most accurate prediction.

For this purpose, a dataset from the patients of Sylhet Diabetes Hospital in Sylhet Bangladesh, was used. This dataset contains data and measurements that aid in the detection of diabetes at an early stage. Using the aforementioned dataset, modeling was conducted using a variety of suitable algorithms, and the calculation of appropriate evaluation metrics was performed to compare them, with the ultimate goal of identifying the optimal algorithm.

1.Introduction

Machine Learning (ML) is focused on the development of algorithms capable of learning from data, adapting to changes, and improving their performance over time. In an era where computers are expected to tackle increasingly complex challenges and become more integrated into our daily lives, ML has emerged as a critical technology. In the field of medicine, ML plays a pivotal role in disease recognition, health monitoring, and the recommendation of preventive measures. This is particularly important due to the challenges associated with manual disease diagnosis, which can range from minor ailments to severe conditions like cancer, often difficult to detect at early stages. Within the realm of diabetes research, ML and data mining techniques are employed to extract valuable insights from vast datasets. Diabetes carries significant social implications in addition to its individual health impacts. Given its substantial socioeconomic consequences, diabetes research generates a wealth of data that demands attention. The specific focus of this thesis lies in the application of ML methods to identify individuals with diabetes at an early stage. This falls under the classification problem category, where the objective is to categorize individuals as either diabetic or non-diabetic. The study involves the comparison of various algorithms used in the classification modeling process, ultimately presenting the most efficient approach.

1.1.Diabetes

As per information from the Center for Disease Control and Prevention in the United States, diabetes is a chronic condition that affects the body's ability to convert food into energy. When an individual consumes food, most of it is transformed into sugar, also known as glucose, and enters the bloodstream. When blood sugar levels rise, the pancreas releases insulin, which acts as a key to allow the blood sugar to enter the body's cells, where it can be used as an energy source. In cases of diabetes, the body either doesn't produce enough insulin or doesn't utilize it effectively. When there's insufficient insulin or when cells become unresponsive to insulin, elevated concentrations of blood sugar persist in the bloodstream.

Over time, this can lead to serious health complications such as heart disease, vision impairment, and kidney disease. It's important to note that while there is currently no cure for diabetes, several strategies can be beneficial. These include weight management, maintaining a healthy diet, and engaging in regular exercise. Additionally, taking prescribed medications as needed, undergoing diabetes self-management education and support, and attending healthcare appointments all contribute to mitigating the impact of diabetes on daily life.

1.1.1.Types

Type 1 diabetes, often referred to as insulin-dependent diabetes or juvenile diabetes, results from an autoimmune response in which the body's immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas. As a result, there is an absolute deficiency of insulin in the body. This type of diabetes is typically diagnosed in childhood or adolescence, although it can occur at any age. The exact cause is not fully understood, but genetics and environmental factors are believed to play a role. Some of the clinical characteristics are :

- Rapid onset of symptoms, including excessive thirst, frequent urination, unexplained weight loss, and fatigue.
- Daily insulin injections or use of an insulin pump is required for survival.
- Tight blood glucose monitoring is necessary to maintain blood sugar within a normal range.
- Susceptibility to diabetic ketoacidosis (DKA), a life-threatening condition caused by high blood sugar and a lack of insulin.

Type 2 diabetes, also known as non-insulin-dependent diabetes, primarily results from insulin resistance, where the body's cells do not respond effectively to insulin. Over time, the pancreas may also fail to produce enough insulin, leading to relative insulin deficiency. This type of diabetes is typically diagnosed in adults, but it can also affect children and adolescents. It is strongly associated with lifestyle factors, including obesity, physical inactivity, and poor dietary choices. Genetics also play a role in susceptibility.
Some of the clinical characteristics are :

- Gradual onset of symptoms, including increased thirst, frequent urination, blurred vision, and fatigue.
- Initially managed with lifestyle changes, such as weight loss, healthy eating, and increased physical activity.
- Medications and, in some cases, insulin therapy may be required as the disease progresses.
- Higher risk of complications such as heart disease, stroke, and kidney problems.

Gestational diabetes develops during pregnancy when the body cannot produce enough insulin to meet increased requirements. It is believed to result from hormonal changes during pregnancy that impair insulin function. This type of diabetes typically occurs during the second or third trimester of pregnancy. Hormonal changes during pregnancy, genetics, and pre-pregnancy obesity can increase the risk of developing gestational diabetes.Some of the clinical characteristics are :

- Often asymptomatic and diagnosed through routine prenatal screening.
- Gestational diabetes usually resolves after childbirth, but it increases the risk of developing Type 2 diabetes later in life.
- Management involves dietary modifications, exercise, and sometimes insulin therapy to ensure a healthy pregnancy and birth.

1.1.2.Statistics

The Centers for Disease Control and Prevention (CDC) regularly publishes the National Diabetes Statistics Report, providing comprehensive insights into diabetes and prediabetes prevalence, incidence, risk factors, complications, mortality rates, and associated costs. This report serves as an update to the 2017 National Diabetes Statistics Report, catering primarily to the scientific community. The following key findings are derived from this report:

- Diabetes Affliction: In the United States, a staggering 34.2 million individuals are afflicted with diabetes, which corresponds to 10.5% of the entire US population. This figure includes 26.9 million diagnosed cases (26.8 million of whom are adults) and an additional 7.3 million individuals, constituting 21.4% of the population, falling into the undiagnosed category.

- Prediabetes Prevalence: Alarmingly, 88 million adults aged 18 or older, representing 34.5% of the total adult US population, have prediabetes. Furthermore, an estimated 24.2 million individuals aged 65 or older are living with prediabetes.

Within the European Union, the prevalence of diabetes has exhibited significant growth over the years. In 2019, approximately 32.3 million adults were diagnosed with diabetes, a considerable increase from an estimated 16.8 million individuals in the year 2000. Furthermore, an estimated 24.2 million people in Europe were projected to have diabetes in 2019 but remained untreated. Notably, since the turn of the century, both male and female diabetes diagnoses have surged, with male cases more than doubling, reaching 16.7 million in 2019, while female diagnoses increased by over 50%, from 9.5 million in 2000 to 15.6 million in 2019. Diabetes prevalence also exhibits age-related trends in the EU. A significant number of individuals aged 60-79, totaling 19.3 million, grapple with diabetes, while 11.3 million individuals aged 40-59 are affected, and a relatively smaller cohort of 1.7 million people aged 20-39 contend with the condition. Notably, while men face a higher likelihood of diabetes during middle age (between 40 and 59 years old), women tend to develop diabetes after the age of 70, primarily due to their longer life expectancy. Across EU countries in 2019, the average diabetes prevalence among adults (adjusted for age) stood at 6.2%. However, prevalence rates varied, with countries such as Cyprus, Portugal, and Germany exceeding 9%, while Ireland and Lithuania recorded rates below 4%. It's important to note that diabetes prevalence has stabilized in several European nations, particularly in the Nordic region. Conversely, in Southern, Central, and Eastern European countries, modest increases in diabetes prevalence have been observed. Contributing factors to these trends include rising rates of obesity, reduced physical activity, and their association with population aging. This comprehensive analysis of diabetes statistics in both the United States and the European Union underscores the magnitude of the diabetes epidemic, highlighting the pressing need for effective preventive measures and improved management strategies.

2.Research Methodology

The primary goal of this thesis is to develop predictive models for early-stage diabetes risk prediction using a dataset of medical signs and symptoms. The methodology employed in this research consists of several stages, including data preprocessing, feature selection, model training, evaluation, and comparison of different machine learning algorithms. This section provides a detailed description of each stage.

2.1.Data Collection and Preprocessing

The dataset utilized for this study is the "Early Stage Diabetes Risk Prediction Dataset," which contains medical signs and symptoms data of newly diabetic or potentially diabetic patients. The dataset includes various features such as age, gender, and symptoms like polyuria, polydipsia, sudden weight loss, weakness, and more. Upon analyzing the distribution of class values, it was evident that there was some imbalance, albeit not significantly skewed in either direction. This suggests that the dataset can be considered fairly balanced as shown in Figure 1.

Data preprocessing involved several key steps:

- *Label Encoding*: Categorical variables such as gender and symptoms were encoded using LabelEncoder to convert them into numerical values suitable for machine learning algorithms.
- *Normalization*: Features were normalized to ensure they have a mean of 0 and a standard deviation of 1. This step was crucial for algorithms sensitive to feature scales.
- *Train-Test Split*: The dataset was split into training and testing sets with an 80-20 ratio to evaluate model performance on unseen data.

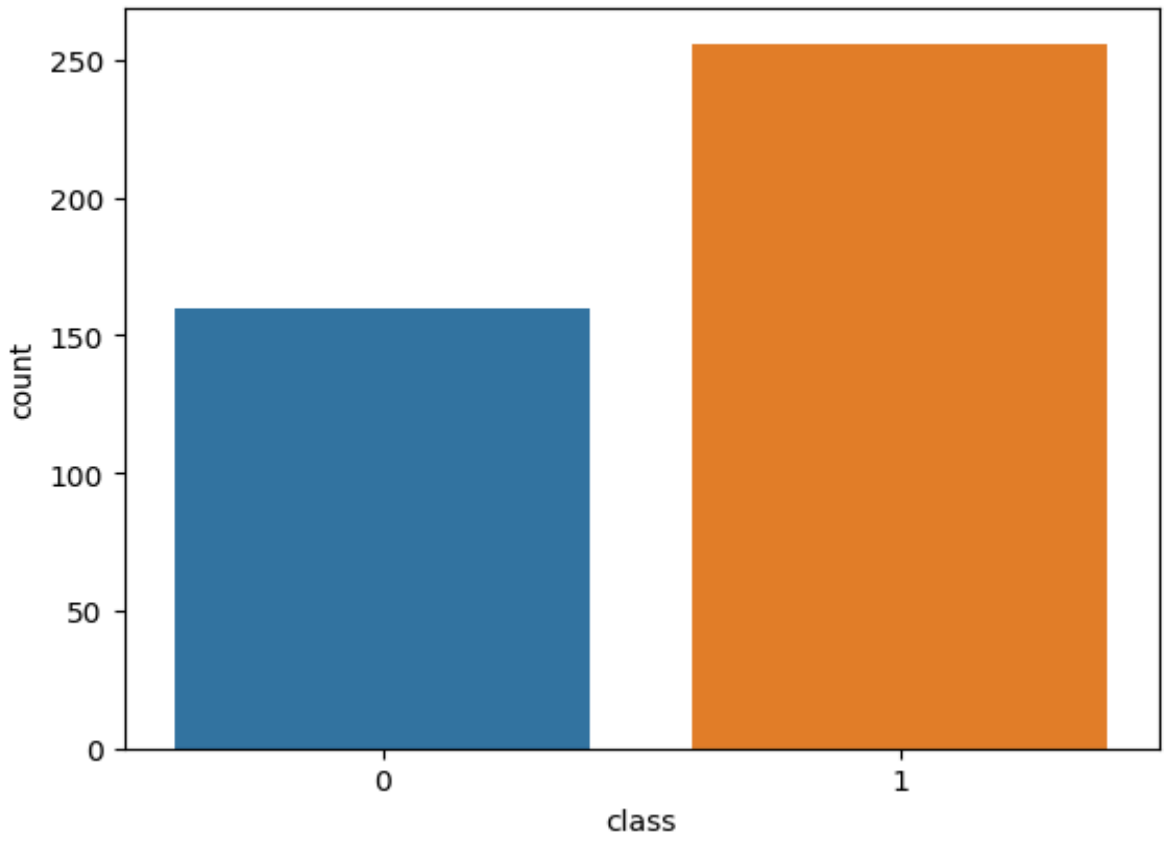

Figure 1. Class Distribution

2.2.Feature Selection with LASSO

Feature selection was conducted using the Least Absolute Shrinkage and Selection Operator (LASSO) regression, which helps in identifying the most significant features by imposing a penalty on the absolute size of the coefficients. Two LASSO models with different alpha values (0.01 and 0.1) were trained to explore the impact of different levels of regularization:

- *Alpha = 0.01*: This value resulted in the selection of all initial features.
- *Alpha = 0.1*: This value identified a subset of the most important features.

The selected features were then used to create separate datasets for subsequent model training and evaluation.

2.3.Model Training and Hyperparameter Tuning

Three machine learning algorithms were employed to develop predictive models:

- *Logistic Regression*
- *Support Vector Classifier (SVC)*
- *Random Forest Classifier*
- *Decision Tree Classifier*

Grid Search with Cross-Validation (CV) was used to identify the optimal hyperparameters for each model. The hyperparameter tuning process aimed to maximize the F1 score, balancing precision and recall.

2.4.Model Evaluation

The performance of each model was evaluated using several metrics:

- *Accuracy*: Measures the overall correctness of the model.
- *Cross-Validated Accuracy*: Provides an estimate of model performance on unseen data.
- *ROC AUC Score*: Assesses the model's ability to discriminate between positive and negative classes.
- *Precision, Recall, and F1-Score*: Evaluate the model's performance in correctly predicting positive and negative classes.

2.5.Comparison of Models

A comprehensive comparison of models was conducted based on their performance on both the entire feature set and the selected important features. This comparison aimed to identify the most suitable model for practical applications in medical settings, such as hospitals and private clinics.

This methodology ensures a thorough and rigorous approach to developing predictive models for early-stage diabetes risk prediction. By leveraging feature selection techniques

and comparing multiple machine learning algorithms, the study aims to provide valuable insights and practical solutions for medical practitioners.

2.6.Machine Learning

In the landscape of medical data analysis, the application of machine learning methodologies is instrumental in uncovering patterns, making predictions, and enhancing clinical decision-making. The next section offers a consolidated overview of the core machine learning techniques that served as the bedrock for extracting meaningful insights from the diabetes dataset.

2.6.1.Logistic Regression

2.6.1.1.Overview

Logistic regression is a valuable modeling technique employed in medical data analysis to establish a statistical relationship between multiple independent variables and a binary dependent variable. This method is particularly suited for categorizing binary outcomes, such as the presence or absence of a disease, or the occurrence of an event.

2.6.1.2.Binary Classification

In the context of medical data analysis, logistic regression is frequently utilized to address binary classification tasks. For instance, it can be applied to predict whether an individual is likely to have a particular disease (positive class) or not (negative class) based on various clinical and demographic factors.

2.6.1.3.The Logistic Function

At the core of logistic regression is the logistic function, often referred to as the sigmoid function. This function is instrumental in transforming an input value, which can range between 0 and 1, into an output value that also falls within the 0 to 1 range. The logistic function, denoted as f(z), reflects the likelihood or probability of an event or disease occurrence. The logistic function is defined as follows:

- z represents an index that combines various risk factors
- -e is the base of the natural logarithm.

2.6.1.4.Probability Estimation

By virtue of the logistic function, the output f(z) yields probabilities ranging between 0 and 1. Consequently, logistic regression is adept at estimating the likelihood or risk of an event or disease occurring. This feature is pivotal in medical applications where understanding the probability of a medical condition is of paramount importance.

2.6.1.5.Model Interpretability

One noteworthy advantage of logistic regression is its interpretability. The model's coefficients offer valuable insights into the impact of each predictor variable on the probability of the outcome. Positive coefficients indicate a positive association, while negative coefficients suggest a negative association.

2.6.1.6.Clinical Applications

Logistic regression finds widespread use in the medical field, including disease prediction, risk assessment, and outcome modeling. For instance, it can be applied to predict the likelihood of an individual developing diabetes based on factors such as age, BMI, family history, and blood glucose levels.

2.6.1.7.Evaluation and Validation

To assess the performance of a logistic regression model in medical data analysis, various evaluation metrics are employed, including accuracy, sensitivity, specificity, ROC-AUC, and precision-recall curves. These metrics aid in gauging the model's ability to make accurate predictions. Logistic regression is a versatile and interpretable statistical technique used in medical data analysis to predict binary outcomes, including the likelihood of disease occurrence. Its ability to provide probability estimates makes it a valuable tool in healthcare applications, offering insights that can aid in clinical decision-making and patient care.

2.6.2.Support Vector Classification (SVC)

2.6.2.1.Overview

Support Vector Classification (SVC) is a robust machine learning technique frequently employed in medical data analysis for binary and multi-class classification tasks. This section

aims to provide a comprehensive understanding of SVC, its applications, and its significance in healthcare analytics.

2.6.2.2.Classification with SVC

Support Vector Classification is a supervised learning algorithm designed to classify data into distinct categories. In medical data analysis, SVC is particularly valuable for predicting binary outcomes, such as disease presence or absence, based on a set of features or attributes.

2.6.2.3.Separation with Hyperplanes

SVC operates by finding the optimal hyperplane(s) that maximally separate data points belonging to different classes. These hyperplanes are strategically positioned to create a clear boundary, known as the margin, between data points. This margin maximization enhances the algorithm's robustness and ability to generalize well to unseen data.

2.6.2.4.Kernel Functions

SVC can handle complex and nonlinear relationships between input features and the target variable through the use of kernel functions. Kernel functions allow SVC to transform the input data into a higher-dimensional space, where linear separation becomes feasible. Common kernel functions include the linear, polynomial, radial basis function (RBF), and sigmoid kernels.

2.6.2.5.Margin and Support Vectors

In the context of SVC, the margin represents the distance between the decision boundary (hyperplane) and the nearest data points from each class. These nearest data points are termed "support vectors." SVC strives to maximize this margin while minimizing classification errors.

2.6.2.6.Robustness and Generalization

SVC is known for its robustness, which means it can effectively handle noisy or overlapping data points. It achieves this by focusing on the data points that are most challenging to classify, the support vectors, which often convey the most critical information.

2.6.2.7.Applications in Healthcare

Support Vector Classification finds widespread application in healthcare analytics, including disease diagnosis, patient risk assessment, and outcome prediction. For example, SVC can be used to predict whether a patient is at high or low risk of developing a specific medical condition based on their medical history, genetics, and lifestyle factors.

2.6.2.8.Evaluation and Validation

The performance of an SVC model in medical data analysis is assessed using various evaluation metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve (ROC-AUC). These metrics help gauge the model's effectiveness in correctly classifying patients and are essential for assessing its clinical utility. Support Vector Classification is a powerful and versatile machine learning method used in medical data analysis to classify patients into distinct categories, such as disease presence or absence. Its ability to handle complex relationships, robustness to noisy data, and capacity for margin optimization make it an invaluable tool for healthcare applications.

2.6.3.Decision Tree

2.6.3.1.Overview

Decision Trees are a fundamental machine learning technique that plays a pivotal role in medical data analysis. In this section, we delve into the principles of decision trees, their relevance in healthcare analytics, and their significance in extracting actionable insights from medical datasets.

2.6.3.2.Tree-Based Learning

Decision Trees are a class of supervised learning algorithms used for both classification and regression tasks. They are particularly valuable in healthcare analytics for classifying patients into distinct categories, predicting outcomes, and aiding clinical decision-making.

2.6.3.3.Hierarchical Structure

At the core of a Decision Tree is a hierarchical structure resembling an inverted tree. The

tree is composed of nodes, where each internal node represents a decision based on a feature, and each leaf node corresponds to a class label or a numerical prediction. The tree is constructed through a recursive process that selects the most informative features for decision-making.

2.6.3.4.Splitting Criteria

Decision Trees employ various criteria to determine how to split the data at each node. Common splitting criteria include Gini impurity, entropy, and mean squared error. The goal is to create splits that maximize the homogeneity of classes within each branch.

2.6.3.5.Interpretability

One of the notable advantages of Decision Trees is their interpretability. Healthcare practitioners can easily trace the decision-making process from the root node to a leaf node, gaining insights into the factors that influence patient outcomes. This transparency is crucial for clinical validation and trust in the model.

2.6.3.6.Prone to Overfitting

While Decision Trees are interpretable and can capture complex relationships, they are also susceptible to overfitting when the tree becomes overly complex and tailored to the training data. Pruning techniques, which remove branches that do not contribute significantly to predictive performance, are often used to mitigate overfitting.

2.6.3.7.Applications in Healthcare

Decision Trees find extensive application in healthcare analytics. They can be used to predict disease risk, stratify patients based on their health status, recommend treatment options, and identify significant clinical markers. For instance, Decision Trees can assist in determining whether a patient is at high or low risk of developing a particular medical condition.

2.6.3.8.Evaluation and Validation

Evaluating the performance of a Decision Tree model in medical data analysis is essential. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the

receiver operating characteristic curve (ROC-AUC). These metrics provide insights into the model's accuracy and its ability to correctly classify patients. Decision Trees are a foundational machine learning technique used extensively in medical data analysis. Their hierarchical and interpretable structure, along with their ability to capture complex relationships, makes them valuable tools for extracting insights from healthcare data.

2.6.4.Random Forest

2.6.4.1.Overview

Random Forest is a prominent ensemble learning technique widely applied in medical data analysis for classification and regression tasks. This section elucidates the core principles of Random Forest, its utility in healthcare analytics, and its significance in deriving valuable insights from medical datasets.

2.6.4.2.Ensemble Learning

Random Forest is a member of the ensemble learning family, which leverages the collective wisdom of multiple individual models to improve predictive accuracy and robustness. In medical data analysis, Random Forest excels in classifying patients into binary or multiclass categories based on their health status or outcomes.

2.6.4.3.Decision Trees as Building Blocks

At the heart of Random Forest are decision trees, which serve as fundamental building blocks. A Random Forest comprises a multitude of decision trees, each constructed using a random subset of the training data and a subset of the input features. This randomness helps combat overfitting and enhances model generalization.

2.6.4.4.Bootstrapping and Aggregation

Random Forest employs bootstrapping, a resampling technique that creates multiple datasets by randomly selecting observations with replacement from the original dataset. Each decision tree is trained on one of these bootstrapped datasets. After training, the predictions of individual trees are aggregated, typically through a majority vote (for classification) or averaging (for regression), to arrive at the final prediction.

2.6.4.5.Feature Importance

One of the notable advantages of Random Forest is its ability to quantify feature importance. By assessing how much each input feature contributes to the model's predictive performance, healthcare practitioners can gain insights into the clinical and demographic factors that most strongly influence patient outcomes.

2.6.4.6.Robustness and Generalization

Random Forest is known for its robustness against noisy data and outliers. The ensemble nature of the model, coupled with its use of diverse subsets of data and features, helps mitigate the impact of individual anomalies, making it well-suited for analyzing real-world medical datasets.

2.6.4.7.Applications in Healthcare

Random Forest is widely applied in healthcare analytics for tasks such as disease diagnosis, patient risk stratification, and drug response prediction. For instance, it can be used to predict the likelihood of a patient developing a specific medical condition based on their medical history, genetic markers, and lifestyle factors.

2.6.4.8.Evaluation and Validation

Assessing the performance of a Random Forest model in medical data analysis involves using various evaluation metrics. These metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). Evaluating these metrics helps gauge the model's effectiveness in making accurate predictions and informs clinical decision-making. Random Forest is a powerful ensemble learning method applied extensively in medical data analysis. Its ability to combine multiple decision trees, handle noisy data, quantify feature importance, and provide robust predictions makes it an invaluable tool for healthcare applications.

3.Results

This section presents the results of our predictive models for early-stage diabetes risk prediction. We have evaluated four machine learning models: Logistic Regression, Support Vector Classifier (SVC), Random Forest, and Decision Tree. Each model was assessed using two different feature sets: all features and the selected important features identified through LASSO regression. The results are displayed in the following tables, followed by a comprehensive comparison of all models.

3.1.Logistic Regression

The Logistic Regression model demonstrated high accuracy on both the training and test datasets. The cross-validated accuracy further confirmed the model's good generalization ability. The excellent ROC AUC score of 0.98 ,as shown in Table 1, indicates a strong discriminative power between the positive and negative classes. The combination of high accuracy, precision, and recall values suggests that the model is well-balanced and effective in classifying instances correctly. When using the selected important features, the Logistic Regression model maintained a high level of performance. Although the train accuracy was slightly lower at 88.70%, the test accuracy remained high at 92.31%. The cross-validated accuracy of 87.74% indicates good generalization with a reduced feature set. The ROC AUC score of 0.95 also highlights the model's capability to discriminate between classes effectively. The model showed excellent precision and recall, particularly for the positive class, indicating its strong predictive power even with fewer features. The results demonstrate that the Logistic Regression model is highly effective for predicting early-stage diabetes risk. Using all features provides the best overall performance, with high accuracy and an excellent ROC AUC score. However, the use of important features identified through LASSO regression ,as shown in Table 2, also results in a robust model with slightly lower, but still very high, performance metrics.

Table 1. Logistic Regression Accuracy

	Train Accuracy (%)	**Test Accuracy(%)**	**Cross Validated Accuracy(%)**	**ROC AUC score**
All Features	93.03	93.27	90.86	0.98
Important Features	88.70	92.31	87.74	0.95

Table 2. Logistic Regression Report

	Accuracy	**Recall-0**	**Recall-1**	**Precision-0**	**Precision-1**	**F1-0**	**F1-1**
All Features	0.93	0.85	0.97	0.93	0.93	0.89	0.95
Important Features	0.92	0.76	1.00	1.00	0.90	0.86	0.95

3.2.Support Vector Classification (SVC)

The SVC model demonstrated high accuracy on the training set and good accuracy on the test set as shown in Table 3. The cross-validated accuracy confirmed the model's ability to generalize well. The ROC AUC score of 0.97 indicates a strong discriminative power between the positive and negative classes. However, the slight drop in test accuracy suggests potential overfitting to the training data, which might need further investigation. When using the selected important features, the SVC model maintained a high level of performance. Although the train accuracy was slightly lower at 88.70%, the test accuracy remained high at 92.31%. The cross-validated accuracy of 88.70% indicates good generalization with a reduced feature set. The ROC AUC score of 0.96 also highlights the model's capability to discriminate between classes effectively. The model showed excellent precision and recall, particularly for the positive class as shown in Table 4, indicating its strong predictive power even with fewer features.The results demonstrate that the SVC model is highly effective for predicting early-stage diabetes risk. Using all features provides slightly better performance in terms of training accuracy and cross-validated accuracy. However, using the important features identified through LASSO regression also results in a robust model with high test accuracy and a high ROC AUC score, while offering the benefit of model simplification.

Table 3. SVC Accuracy

	Train Accuracy (%)	**Test Accuracy(%)**	**Cross Validated Accuracy(%)**	**ROC AUC score**
All Features	93.51	89.42	93.25	0.97
Important Features	88.70	92.31	88.70	0.96

Table 4. SVC Report

	Accuracy	**Recall-0**	**Recall-1**	**Precision-0**	**Precision-1**	**F1-0**	**F1-1**
All Features	0.89	0.85	0.92	0.82	0.93	0.84	0.92
Important Features	0.92	0.76	1.00	1.00	0.90	0.86	0.95

3.3.Random Forest

The Random Forest model demonstrated excellent accuracy on both the training set and the test set. The cross-validated accuracy further confirmed the model's ability to generalize well. The ROC AUC score of 0.99, as shown in Table 5, indicates a strong discriminative power between the positive and negative classes. The combination of high accuracy, precision, and recall values suggests that the model is well-balanced and highly effective in classifying instances correctly. However, the very high train accuracy suggests potential overfitting, which may need to be addressed. When using the selected important features, the Random Forest model maintained a high level of performance. Although the train accuracy was lower at 88.70%, the test accuracy remained high at 92.31%. The cross-validated accuracy of 86.54% indicates good generalization with a reduced feature set. The ROC AUC score of 0.95 highlights the model's capability to discriminate between classes effectively. The model showed excellent precision and recall, particularly for the positive class as shown in Table 6, indicating its strong predictive power even with fewer features. The results demonstrate that the Random Forest model is highly effective for predicting early-stage diabetes risk. Using all features provides the best overall performance, with very high accuracy and an excellent ROC AUC score. However, the use of important features identified through LASSO regression also results in a robust model with high test accuracy and a high ROC AUC score, while offering the benefit of model simplification.

Table 5. Random Forest Accuracy

	Train Accuracy (%)	**Test Accuracy(%)**	**Cross Validated Accuracy(%)**	**ROC AUC score**
All Features	98.56	94.23	94.00	0.99
Important Features	88.70	92.31	86.54	0.95

Table 6. Random Forest Report

	Accuracy	**Recall-0**	**Recall-1**	**Precision-0**	**Precision-1**	**F1-0**	**F1-1**
All Features	0.94	0.97	0.93	0.86	0.99	0.91	0.96
Important Features	0.92	0.76	1.00	1.00	0.90	0.86	0.95

3.4.Decision Tree

The Decision Tree model demonstrated high accuracy on both the training set and the test set. The cross-validated accuracy further confirmed the model's ability to generalize well. The ROC AUC score of 0.99 ,as shown in Table 7, indicates a strong discriminative power between the positive and negative classes. The combination of high accuracy, precision, and recall values suggests that the model is well-balanced and highly effective in classifying instances correctly. When using the selected important features, the Decision Tree model maintained a high level of performance. Although the train accuracy was lower at 88.70%, the test accuracy remained high at 92.31%. The cross-validated accuracy of 88.22% indicates good generalization with a reduced feature set. The ROC AUC score of 0.96 highlights the model's capability to discriminate between classes effectively. The model showed excellent precision and recall, particularly for the positive class as shown in Table 8, indicating its strong predictive power even with fewer features. The results demonstrate that the Decision Tree model is highly effective for predicting early-stage diabetes risk. Using all features provides the best overall performance, with high accuracy and an excellent ROC AUC score. However, the use of important features identified through LASSO regression also results in a robust model with high test accuracy and a high ROC AUC score, while offering the benefit of model simplification.

Table 7. Decision Tree Accuracy

	Train Accuracy (%)	**Test Accuracy(%)**	**Cross Validated Accuracy(%)**	**ROC AUC score**
All Features	96.39	96.15	93.99	0.99
Important Features	88.70	92.31	88.22	0.96

Table 8. Decision Tree Report

	Accuracy	**Recall-0**	**Recall-1**	**Precision-0**	**Precision-1**	**F1-0**	**F1-1**
All Features	0.96	0.91	0.99	0.97	0.96	0.94	0.97
Important Features	0.92	0.76	1.00	1.00	0.90	0.86	0.95

3.5.Comparison

The following section presents the performance metrics for the different models evaluated in this study. Each model was assessed using two feature sets: all features and the selected important features identified through LASSO regression. The goal is to determine the most effective model for early-stage diabetes risk prediction by comparing their performance through the metrics shown in Table 9 - Table 12.

Table 9. Models Accuracy For All Features

All Features	**Train Accuracy (%)**	**Test Accuracy(%)**	**Cross Validated Accuracy(%)**	**ROC AUC score**
Logistic Regression	93.03	93.27	90.86	0.98
SVC	93.51	89.42	93.25	0.97
Random Forest	98.56	94.23	94.00	0.99
Decision Tree	96.39	96.15	93.99	0.99

Table 10. Models Accuracy For Important Features

Important Features	**Train Accuracy (%)**	**Test Accuracy(%)**	**Cross Validated Accuracy(%)**	**ROC AUC score**
Logistic Regression	88.70	92.31	87.74	0.95
SVC	88.70	92.31	88.70	0.96
Random Forest	88.70	92.31	86.54	0.95
Decision Tree	88.70	92.31	88.22	0.96

Table 11. Report For All Features

All Features	Accuracy	Recall-0	Recall-1	Precision-0	Precision-1	F1-0	F1-1
Logistic Regression	0.92	0.85	0.97	0.93	0.93	0.89	0.95
SVC	0.89	0.85	0.92	0.82	0.93	0.84	0.92
Random Forest	0.94	0.97	0.93	0.86	0.99	0.91	0.96
Decision Tree	0.96	0.91	0.99	0.97	0.96	0.94	0.97

Table 12. Report For Important Features

Important Features	Accuracy	Recall-0	Recall-1	Precision-0	Precision-1	F1-0	F1-1
Logistic Regression	0.92	0.76	1.00	1.00	0.90	0.86	0.95
SVC	0.92	0.76	1.00	1.00	0.90	0.86	0.95
Random Forest	0.92	0.76	1.00	1.00	0.90	0.86	0.95
Decision Tree	0.92	0.76	1.00	1.00	0.90	0.86	0.95

4.Conclusions

The final choice of the optimal model for early-stage diabetes risk prediction hinges on the balance between performance and simplicity, a crucial consideration in real-world applications. Based on our comprehensive evaluation, we can draw nuanced conclusions regarding the suitability of each model for practical use.

4.1.Performance vs. Simplicity

- *High Performance*: Models such as Random Forest and Decision Tree, when utilizing all available features, demonstrate the highest levels of accuracy and predictive power. These models are characterized by their exceptional ability to discern between classes, as evidenced by their superior ROC AUC scores and high cross-validated accuracy. Such performance metrics suggest that these models are particularly adept at handling complex, high-dimensional data, making them suitable for scenarios where the utmost precision in prediction is paramount.
- *Simplicity*: On the other hand, Logistic Regression and SVC models, when constrained to significant features identified through LASSO regression, offer a compelling balance of performance and simplicity. These models, while slightly trailing in raw predictive power compared to their more complex counterparts, provide robust accuracy and generalization capabilities. Their reduced complexity translates to faster training times and more interpretable results, which are highly valuable in clinical settings where model transparency and ease of use are critical.

4.2.Practical Applications in Healthcare

- *Logistic Regression with Significant Features*: This model stands out for its ease of interpretation, making it particularly valuable in a clinical environment. Healthcare professionals can readily understand the influence of individual predictors on the model's output, facilitating informed decision-making. The model's simplicity also ensures quick training and deployment, crucial in time-sensitive medical contexts.
- *Random Forest with All Features*: While more complex, this model's superior accuracy and generalization make it an excellent choice for environments where predictive performance is non-negotiable. In settings like hospitals and private clinics, where the stakes of diagnostic accuracy are high, the use of a Random Forest model ensures that predictions are as precise as possible. The trade-off in interpretability can be mitigated by employing techniques such as feature importance analysis, which can help elucidate the model's decision-making process.

In conclusion, the selection of the appropriate model for early-stage diabetes risk prediction must be informed by the specific needs and constraints of the deployment environment. For applications demanding the highest predictive accuracy and where computational resources permit, Random Forest with all features emerges as the preferred model. Conversely, in scenarios where model interpretability, speed, and ease of deployment are paramount, Logistic Regression with significant features presents a highly viable alternative. These findings underscore the importance of a tailored approach to model selection, ensuring that the chosen model aligns with both the technical requirements and practical considerations of its intended use.

5.Suggestions

Based on the findings of this study, several suggestions can be made to further enhance the effectiveness and applicability of machine learning models for early-stage diabetes risk prediction

5.1.Feature Engineering and Selection

- Continual improvement in feature engineering techniques could enhance model performance. Incorporating domain knowledge from medical professionals to create new features or refine existing ones could lead to more accurate predictions.
- Regularly update the set of significant features using techniques like LASSO regression to ensure that the model adapts to any changes in the data patterns over time.
- *Implementation Note*: The code has been developed to facilitate easy input of LASSO parameters through a table format. This modular design allows for straightforward adjustments and updates. However, the overall process is contingent on the computational power available, as the complexity of the computations can be significant.

5.2.Integration with Clinical Workflows

- Develop user-friendly interfaces and decision support systems that seamlessly integrate the predictive models into existing clinical workflows. This ensures that the models are accessible and usable by healthcare professionals with varying levels of technical expertise.
- Incorporate feedback mechanisms that allow clinicians to provide input on the model's predictions, enabling continuous improvement and adaptation of the model to clinical needs.

5.3.Regular Updates and Monitoring

- Continuously monitor the model's performance over time and update it as new data becomes available. This helps in maintaining the accuracy and relevance of the predictions.
- Establish a system for periodic retraining of the models to incorporate new data and adapt to any changes in the underlying population or clinical practices.

5.4.Ethical Considerations and Data Privacy

- Ensure that the models are developed and deployed in compliance with ethical standards and data privacy regulations. Protecting patient data and maintaining transparency in how the models use this data is crucial for gaining trust and acceptance in clinical settings.
- Conduct bias assessments to ensure that the models provide fair and unbiased predictions across different patient demographics.

By implementing these suggestions, the predictive models for early-stage diabetes risk can become more accurate, interpretable, and seamlessly integrated into clinical practice, ultimately contributing to better patient outcomes and more efficient healthcare delivery.

References

Aishwarya, R., et al. “A Method for Classification Using Machine Learning Technique for Diabetes.” *Research Gate*, June 2013, https://www.researchgate.net/publication/291313554_A_Method_for_Classification_Using_Machine_Learning_Technique_for_Diabetes.

“Analysis of Various Decision Tree Algorithms for Classification in Data Mining.” *International Journal of Computer Applications*, https://www.ijcaonline.org/archives/volume163/number8/gupta-2017-ijca-913660.pdf. Accessed 4 August 2024.

Brownlee, Jason. “4 Types of Classification Tasks in Machine Learning - MachineLearningMastery.com.” *Machine Learning Mastery*, 19 August 2020, https://machinelearningmastery.com/types-of-classification-in-machine-learning/. Accessed 4 August 2024.

“Diabetes Basics | Diabetes.” *CDC*, 15 May 2024, https://www.cdc.gov/diabetes/about/index.html. Accessed 4 August 2024.

“Diabetes prevalence | Health at a Glance: Europe 2020: State of Health in the EU Cycle.” *OECD iLibrary*, 19 November 2020, https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2020_83231356-en. Accessed 4 August 2024.

Kaur, Gaganjot, and Amit Chhabra. *Diabetes Diagnosis using Machine Learning*, https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/14226/Mamandra_me1933.pdf?sequence=3&isAllowed=y. Accessed 4 August 2024.

Kaur, Gaganjot, and Amit Chhabra. “Improved J48 Classification Algorithm for the Prediction of Diabetes.” *Research Gate*, July 2014, https://www.researchgate.net/publication/269669737_Improved_J48_Classification_Algorithm_for_the_Prediction_of_Diabetes.

Kumari, Vinita. “(PDF) Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach.” *ResearchGate*, 13 February 2019, https://www.researchgate.net/publication/329807137_Predictive_Modelling_and_Analytics_for_Diabetes_using_a_Machine_Learning_Approach. Accessed 4 August 2024.

“Machine Learning in Healthcare - Unlocking the Full Potential!” *DataFlair*, https://data-flair.training/blogs/machine-learning-in-healthcare/. Accessed 4 August 2024.

“Machine Learning Theory.” *CMU School of Computer Science*, https://www.cs.cmu.edu/~avrim/Talks/mlt.pdf. Accessed 4 August 2024.

Sanner, MF. “Python: a programming language for software integration and development.” *PubMed*, https://pubmed.ncbi.nlm.nih.gov/10660911/. Accessed 4 August 2024.