**Project II: Machine Learning**

Pick a readily available dataset of your choice to build a regression or classification model. Write a report with the following details:

describe the problem with proper literature review (10 points)

provide a descriptive analysis of the data set including visualization (5 points)

build at least three machine learning models and compare the performances of these models (10 points)

Write a detailed description of the evaluation metrics used and the data-splitting techniques (10 points)

provide a thorough error analysis and propose recommendations for improving the models (5 points)

using the principles of responsible artificial intelligence discussed during the tutorials, discuss how your project meets responsible AI guidelines.

**Application of Machine learning for Predicting diabetics using related health conditions and lifestyle Factors**

# Introduction

Diabetes mellitus is a complicated condition involving metabolic dysfunction, marked by abnormally elevated levels of glucose in the blood. This elevation frequently results from either the body's inability to properly utilise insulin or an inadequate production of insulin (Blair 2016). Type 1 diabetes is an autoimmune disorder where the body's immune system attacks and destroys the insulin-producing beta cells in the pancreas, leading to a deficiency of insulin (Blair 2016). In contrast, type 2 diabetes primarily arises from issues regulating glucose levels, caused by dysfunction in both the insulin-producing pancreatic beta cells as well as resistance to the effects of insulin (Perveen et al., 2019). This condition can precipitate severe long-term complications such as cardiovascular issues, stroke, kidney failure, heart attacks, peripheral arterial disease, and nerve damage (Maniruzzaman et al., 2020). In 1980, approximately 122 million individuals globally were affected by diabetes, a number that surged to about 422 million by 2014. Projections suggest that by 2040 this figure will rise to about 642 million (Zimmet et al., 2016). Additionally, there were roughly 1.6 million diabetes-related deaths directly attributed to the condition.

The rapid growth of diabetes is a global concern, with a significant impact on health and healthcare costs (Galloway, 2022). The health condition is driven by a range of factors, including lifestyle changes, urbanization, and genetic predisposition (Galloway, 2022). Early detection is crucial in reducing the risk of complications and associated costs. Precision medicine is proposed as a potential solution to address the rising expenditures and inequalities in diabetes care (Galloway, 2022). However, the need for greater awareness and earlier detection is emphasized, particularly in Type 2 diabetes, to mitigate the accelerated complications associated with the disease (Kumar et al., 2019).

Data science solutions have catalysed a revolution in the healthcare sector, facilitating advancements in areas like drug discovery and understanding genetic diseases. Consequently, there remains substantial untapped potential in this domain that warrants further exploration for the betterment of society. Machine learning offers an array of algorithms capable of delivering heightened accuracy in predicting outcomes based on input data, bolstered by statistical analysis (Kumar et al., 2019). The main goal of machine learning algorithms is to make use of a given dataset to enhance our understanding of the data and discover useful knowledge. Supervised machine learning is characterized by the use of labelled data to train its algorithms and can be utilized for classification or regression tasks. The goal of classification is to assign each unknown instance to one of the possible classes or categories for prediction or diagnosis purposes (Jian et al., 2021). Machine learning is good at working with the nonlinear, nonnormal, correlation-structured, and complex nature of diabetic data (Maniruzzaman et al., 2020). Machine learning aids in accurately diagnosing diabetes, with identifying the most effective classifier being crucial for precise diabetes risk assessment.

In medicine, there is an accelerating focus on preventing diabetes. Leveraging machine learning models expedites the analysis of data, allowing analysts to scrutinize existing datasets to uncover patterns and trends associated with diabetes (Llaha & Rista, 2021). Through employing machine learning, physicians can predict illnesses with increased effectiveness, enabling them to better handle individuals deemed at high risk (Bisandu, 2019). The abundant data on diabetic diseases and the intricate interrelationships with lifestyle factors render medicine a suitable arena for the application of machine learning.

This project aims at using machine learning models for the diagnosis of diabetics in patients using health-related and lifestyle factors as predictors. The machine learning models are compared using multiple evaluation metrics and propose methods for their improvement.

# Literature Review

Accurately categorizing and forecasting health outcomes based on patient data is essential for efficient disease management and prevention. Machine learning methods have become influential tools for extracting insights from intricate health datasets, enabling the identification of patterns, and predictions, and aiding clinical decision-making. Among these techniques, logistic regression, K-nearest neighbours (KNN), decision trees, and random forests, have shown significant potential in health data analysis. Each of these models offers unique strengths in tackling binary classification tasks, managing complex data sets, capturing nonlinear relationships, and enhancing prediction precision. This literature review investigates the applications, merits, and constraints of these machine learning approaches concerning diabetic outcome, classification and prediction, emphasizing their efficacy in identifying risk factors, diagnosing ailments, and guiding preventive measures.

Logistic regression is a widely used statistical model for binary classification in health data analysis, as it can predict the probability of a patient having a particular condition based on various predictor variables (Lydia, 2019; Cleophas, 2013). It does not assume a linear relationship between the predictors and the outcome, but instead uses the logistic sigmoid function to model the probability (Zaidi, 2023). This makes it a simple and common machine-learning algorithm for binary classification problems (Lydia, 2019). The logistic regression (LR) model predicts the likelihood of a binary outcome (such as being diabetic) based on certain predictor variables. It achieves this by establishing a relationship between these predictors and the log odds of the outcome. The logit transformation ensures that the estimated probabilities fall between 0 and 1, which is essential for binary outcomes. Mathematically, the logit of the outcome variable (Y) is represented as a linear combination of the predictor variables (X):

logit(P(Y = 1)) = log(P(Y = 1) / (1 - P(Y = 1))) = β0 + β1X1 + β2X2 + ... + βkXk

In this equation, P(Y = 1) denotes the probability of the event (e.g., being diabetic), β0 is the intercept term, and β1, β2, ..., βk are the regression coefficients corresponding to the predictor variables X1, X2, ..., Xk. These coefficients quantify the impact of each predictor on the log odds of the outcome.

The unknown regression coefficients (β's) are determined using the maximum likelihood estimation (MLE) method, which seeks the coefficients that maximize the likelihood of observing the given data. Once these coefficients are estimated, odds ratios (ORs) can be calculated by exponentiating the regression coefficients (e.g., OR = exp(βi)). The statistical significance of the regression coefficients (or ORs) can be evaluated using tests like the Z-test or the Wald test. These tests help identify which predictor variables are significantly associated with the outcome and thus should be included in the final model.

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning technique used for classification and regression purposes (Zhang, 2016). When applied to health data classification, KNN determines the class label of a new data point by identifying the K nearest points in the training dataset and assigning the label based on the majority vote among those neighbours (Galloway, 2022). The selection of K and the distance measurement method employed (such as Euclidean or Manhattan distance) can notably influence the model's effectiveness. The K-Nearest Neighbors (KNN) algorithm is a versatile learning technique commonly used for classification and regression tasks, known for its simplicity and effectiveness. In classification scenarios, like sorting health data, KNN predicts the class label of a new data point by finding its closest neighbours in the training set and determining the majority class among them. Unlike parametric methods, KNN does not assume any specific data distribution or learn a discriminative function; instead, it relies on the proximity of data points in the feature space to infer class membership. To use KNN, you first extract relevant features from the input data and normalize them if necessary to ensure fair distance calculations. Then, you choose a distance metric like Euclidean, Manhattan, or Minkowski to measure the similarity between the new data point and the training instances. The distance metric significantly influences KNN's performance. Once the distance metric is set, the algorithm calculates distances between the query point and all training instances, selecting the K nearest neighbours. The choice of K, typically an odd number to prevent ties, is crucial and affects model complexity: a small K may overfit, while a large K may underfit. After identifying the nearest neighbours, KNN assigns the class label to the query point based on the majority vote. If K is even, ties are resolved randomly or by considering the nearest neighbour among the tied instances. KNN's strengths lie in its simplicity, interpretability, and natural handling of multi-class problems. However, it has drawbacks, including sensitivity to the distance metric, computational demands with large datasets, and susceptibility to the curse of dimensionality, where performance suffers as the number of features increases.

Random forests are effective in handling high-dimensional health data. Capitaine (2019) and Zhe-xue (2013) both highlight the method's ability to manage complex data and improve prediction accuracy. Capitaine (2019) specifically discusses the application of random forests to high-dimensional longitudinal data, while Zhe-xue (2013) introduces an improved algorithm for classification problems with very high-dimensional data. Joly (2014) further enhances the method by proposing the use of random projections in the output space for high-dimensional multi-label classification, reducing learning time complexity without sacrificing accuracy. These studies demonstrate the robustness and versatility of random forests in handling the complexities of health datasets.

The Random Forest algorithm operates by first creating multiple bootstrap samples from the original training dataset. Each bootstrap sample is created by randomly sampling instances from the training data with replacement, resulting in new datasets of the same size as the original training set. However, some instances may be repeated in each bootstrap sample, while others may be left out. For each bootstrap sample, the algorithm constructs an unpruned decision tree. During the tree-growing process, at each node, a random subset of features is selected from the total set of features. The feature that provides the best split is chosen from this random subset, rather than considering all features. This randomization process helps to reduce the correlation between the individual trees in the ensemble, promoting diversity and preventing overfitting. The process of creating bootstrap samples and growing decision trees is repeated multiple times, typically hundreds or thousands of times, resulting in an ensemble of decision trees, collectively referred to as a "forest." When a new, unlabeled instance needs to be classified or a prediction needs to be made, the instance is passed through each of the decision trees in the forest. For classification tasks, each tree casts a vote for the predicted class, and the class with the majority votes across all trees is assigned as the final prediction. For regression tasks, the predictions from all trees are averaged to obtain the final predicted value.

The Random Forest algorithm has several advantages, including its ability to handle high-dimensional data, resistance to overfitting due to the ensemble approach and feature randomization, and the ability to capture non-linear relationships between features and the target variable. One of the notable advantages of Random Forest is its ability to provide feature importance measures (Robnik-Šikonja, 2004). This is achieved by evaluating the decrease in node impurity (for classification) or the decrease in mean squared error (for regression) when splitting on a particular feature, averaged over all trees in the forest. Features with higher importance scores are considered more relevant for the prediction task. Additionally, it can handle missing values and outliers effectively, and it provides an estimate of feature importance, which can be valuable for feature selection or understanding the underlying relationships in the data. However, it is essential to note that Random Forest is a complex algorithm with several hyperparameters that need to be tuned, such as the number of trees, the maximum depth of each tree, the number of features to consider at each split, and the splitting criterion. Proper hyperparameter tuning and cross-validation techniques are crucial to ensure optimal performance and avoid overfitting or underfitting (Segal, 2004).

Decision trees are a widely used machine learning method for both classification and regression tasks. They are structured like trees, with each internal node representing a decision or test on a feature, branches indicating outcomes, and leaf nodes denoting class labels for classification or numerical values for regression. In classification, decision trees recursively divide the input space into smaller regions based on feature values. Starting at the root node, the algorithm selects the best feature and split point to separate classes in the training data, typically based on metrics like information gain, Gini impurity, or entropy reduction (Tangirala, 2020). This process continues until a stopping criterion, like reaching a maximum depth or achieving purity within nodes, is met. To classify new instances, the constructed tree is traversed from the root to a leaf node, following decisions based on feature values. The label associated with the reached leaf node is then assigned as the predicted class. Decision trees offer interpretability and suitability for tasks requiring transparency, such as medical diagnosis. They handle various data types and automatically select discriminative features while being robust to irrelevant ones. However, decision trees can overfit, especially with deep trees or noisy data (Tangirala, 2020). Techniques like pruning and ensemble methods such as Random Forests can address overfitting. Choosing the right splitting criterion, like information gain or Gini impurity, significantly influences the tree's structure and performance.

These machine learning algorithms have been used to predict diabetics using lifestyle, environmental and health-related factors. In most cases, the goal is to search for the best model that would give the best prediction. Machine learning models are compared using evaluation metrics such as accuracy, precision, Area under the receiver operating characteristic (AROC) curve, F-measure, and recall. Below are some notable literature on this subject.

Hang et al. (2019) presented predictive models using Gradient Boosting Machine and Logistic Regression techniques, random forest and decision tree models to predict the probability of patients having DM based on their demographic information and laboratory results from their visits to medical facilities. The predictors of interest encompassed sex, age, BMI, triglycerides, fasting blood sugar, systolic blood pressure, high-density lipoprotein, and low-density lipoprotein. The Area Under the Receiver Operating Characteristic Curve (AROC) for the proposed GBM model stands at 84.7%, with a sensitivity of 71.6%, while the AROC for the proposed Logistic Regression model is 84.0%, with a sensitivity of 73.4%. The GBM and Logistic Regression models outperform the Random Forest and Decision Tree models. The authors noted that fasting blood glucose, BMI, high-density lipoprotein, and triglycerides emerged as the most significant predictors in these models.

Perveen et al. (2019) pioneered the application of a Hidden Markov Model (HMM) to assess the effectiveness of the Framingham Diabetes Risk Scoring Model (FDRSM). Their objective was to forecast the 8-year probability of an individual developing type 2 diabetes, leveraging longitudinal Electronic Medical Record (EMR) data from 172,168 primary care patients. In our study, focusing on a subset of 911 individuals with complete risk factor information and follow-up data, the Area Under the Receiver Operating Characteristic Curve (AROC) reached 86.9%.

Monalisa et al. (2021) assessed the predictive performance of various machine learning models - including logistic regression, K-nearest neighbours, support vector machine, and gradient boosting classifier for predicting diabetes diagnoses. Their study employed a dataset sourced from Kaggle, consisting of 768 observations and 9 attributes. Their results indicated that the gradient boosting classifier demonstrated strong performance on this dataset, as evidenced by its accuracy and recall metrics. Moreover, the K-nearest neighbours algorithm proved effective particularly for larger datasets, offering faster processing times. Additionally, the support vector machine model adeptly handled the diverse range of features present in the dataset. Similarly, Llaha & Rista (2021) aimed to classify individuals diagnosed with diabetes using algorithms such as Logistic Regression (LR), Random Forest, Support Vector Machine (SVM), K-nearest neighbours (KNN), Gradient Boosting (GB), and Decision Tree (DT). They utilized the same dataset as Monalisa et al. (2021). Their experiment revealed that the K-nearest neighbours algorithm outperformed other classification algorithms, achieving an accuracy of 85%. The authors suggested that comparing multiple machine learning algorithms would yield optimal results for predicting and diagnosing diabetes.

Emon et al. (2021) aimed to investigate the relationships between different symptoms and diseases that contribute to diabetes, to aid early diagnosis. They employed eleven different machine learning classification algorithms: Logistic Regression (LR), Gaussian Process (GP), Adaptive Boosting (AdaBoost), Decision Tree (DT), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Bernoulli Naive Bayes (BNB), Bagging Classifier (BC), Random Forest (RF), and Quadratic Discriminant Analysis (QDA). Among these classifiers, the Random Forest (RF) algorithm demonstrated the best performance, achieving an accuracy of 98%. Furthermore, the RF classifier also had the highest Area Under the Curve (AUC) value, which is a measure of the classifier's ability to distinguish between positive and negative instances. The approach of leveraging multiple machine learning classifiers to uncover the relationships between various symptoms and diabetes aligns with recent research trends in this domain.

Daanouni et al. (2019) proposed a support diagnostic system that compared four prediction algorithm models for predicting diabetes under two different scenarios. Two data sets were used: The first dataset was obtained from a hospital in Frankfurt, Germany, while the second dataset was the well-known Pima Indian dataset. Both datasets consisted of two parts - data from healthy patients and data from diabetic patients. The first dataset contained 2000 instances, and the second dataset had 768 patient records with 8 features/attributes and one output label indicating whether the patient was diabetic (1) or not diabetic (0). The researchers found that the pre-processing phase was relevant for disease prediction, and deep neural networks (DeepNNs) demonstrated their potential to achieve higher accuracy. The KNN and DeepNN algorithms attained the highest accuracies of 97.53% and 96.35%, respectively. In this case, KNN provided the best result because the classes were binomial and quite separable. Additionally, KNN generates a highly convoluted decision boundary as it is driven by the raw training data itself.

# 3. Materials and Methods

## 3.1. Data collection:

The data utilized in this study was sourced from Kaggle, comprising a consolidated dataset derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015. BRFSS is an annual health-related telephone survey administered by the Center for Disease Control (CDC), gathering information from Americans regarding health-related risk behaviours, chronic health conditions, and preventative service usage. Conducted since 1984, each year's survey captures a broad spectrum of health-related data. Specifically for this project, a CSV file of the 2015 dataset available on Kaggle was used. . Featuring 21 variables, this dataset is balanced in terms of class distribution., with half of the respondents reporting no diabetes and the other half reporting either prediabetes or diabetes. The target variable, Diabetes\_binary, comprises two classes: 0 indicating the absence of diabetes and 1 indicating prediabetes or diabetes

## 3.2. Data cleaning and variable selections

The data consisting of 70,692 rows and 21 attributes was clean, with no missen values or outliers. The predictor variable, which was of a binary class, was balanced (50-50%). However, 9 predictor variables were selected for developing the machine learning models. These variables include Age, sex, high alcohol consumption, education, Body mass index (BMI), high cholesterol level, fruit consumption, and healthcare coverage.

The data age and BMI variable were Z-score transformed before creating the machine learning models.

All Machine learning models were made in R version 4.03.

## 3.3. Data Splitting Techniques:

The data was randomly split into training (70%) and testing sets (30%). Each machine learning model would be trained on the ‘training set’, and tested on the ‘testing set’.

## 3.4. Model Evaluation Metrics:

The machine learning models were evaluated based on accuracy, precision, recall and F-measure.

### 3.4.1 Accuracy:

Accuracy is the most intuitive metric in model evaluation, representing the ratio of correct predictions to the total number of predictions. It is calculated by dividing the sum of true positives (TP) and true negatives (TN) by the total number of instances, as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where FP and FN are false positives and false negatives, respectively.

Accuracy is a good measure when the dataset is balanced, meaning the number of instances in each class is approximately equal. Such is the case in this work. However, it can be misleading, especially when dealing with imbalanced datasets, as it does not distinguish between the types of errors made by the model (Akosa, 2017).

### 3.4.2 Precision:

Precision measures the fraction of true positives among the instances classified as positive by the model. It is calculated as:

Precision = TP / (TP + FP)

Precision is particularly useful when the cost of false positives is high, such as in diabetic detection systems, where incorrectly classifying a legitimate instance as positive can have severe consequences.

### 3.4.3. Recall (Sensitivity or True Positive Rate):

Recall, also known as sensitivity or true positive rate, measures the fraction of actual positives that the model correctly identifies. It is calculated as:

Recall = TP / (TP + FN)

Recall is important when the cost of false negatives is high, such as in a diabetic diagnosis, where missing a positive instance can have severe consequences.

### 3.4.4. F-measure:

The F-measure, also known as the F-score or F1 score, is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it useful when both metrics are equally important. The F-measure is calculated as:

F-measure = 2 × (Precision × Recall) / (Precision + Recall)

The F-measure ranges from 0 to 1, with 1 being the best score.

These metrics are commonly used in binary classification problems, where instances are classified into two classes (positive or negative).

# 4. Results

## 4.1 Descriptive analysis of the data set including visualization

The distribution of the factors used in this project are presented in figures (1 – 10), respectively. The survey encompassed 70,692 participants, with 9 predictor variables utilized to elucidate or predict diabetic status. Among the participants, 39,832 exhibited high blood pressure, of which 26,604 were diabetic. Additionally, 23,686 tested positive for diabetes of 37,163 participants with elevated cholesterol levels. Diabetic patients had higher BMI levels (31.94±7.36) compared to non-diabetic individuals (27.77±6.19). While the male population was smaller than the female population overall, a majority of participants with diabetes were females (18,411). A greater proportion of participants (43,249) reported consuming fruits at least once a week compared to those who did not (27,443). Furthermore, fewer educated individuals were afflicted with diabetes compared to their non-educated counterparts. Additionally, 95.50% of participants had adequate healthcare coverage, with 33,924 out of 67,508 having diabetes. Only 832 participants reported excessive alcohol consumption. For educational level, a slightly higher number of people tend to have diabetics at every level except at the highest educational level (6).

Figure 1: Distribution of Diabetics status among participants

Figure 2: Distribution of respondents with high and low blood pressure

Figure 3: Cholesterol level in respondents

Figure 4: Body mass index of participants. The Bar chart with error bars represents the mean and standard deviation.

Figure 5: Distribution of age of participants. The Bar chart with error bars represents the mean and standard deviation.

Figure 6: Sex distribution in participants

Figure 7: Distribution of the number of Active fruit eating participants

Figure 8: Participate with healthcare coverage

Figure 9: Distribution of heavy alcohol-consuming participants.

Figure 10: Educational level and diabetic status among participants.

## 4.2: Machine learning model for the prediction of diabetics

The effect of the selected factors on diabetic diagnostics as explained by logistic regression are presented in Table 1. All predictor variables (except for ‘Any health care’) were statistically significant (P<0.05) in predicting diabetics. From the result, BMI, High blood pressure, high alcohol consumption, and age were highly positively associated with the proclivity to be diabetic. Males also had a higher chance of being diabetics than females. Those with high heavy alcohol consumption are more likely to not have diabetes, and those who eat fruit have reduced chances of being diabetic.

Table 1: Effect of selected factors on diabetics using logistic regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Predictors | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | 0.18611 | 0.0709 | 2.625 | 0.00867 | \*\* |
| BMI | 0.65148 | 0.01289 | 50.539 | < 2e-16 | \*\*\* |
| HighBP | 0.95359 | 0.02242 | 42.524 | < 2e-16 | \*\*\* |
| HighChol | 0.71362 | 0.02152 | 33.167 | < 2e-16 | \*\*\* |
| Age | 0.49213 | 0.01229 | 40.051 | < 2e-16 | \*\*\* |
| Fruits | -0.12252 | 0.02186 | -5.606 | 2.07E-08 | \*\*\* |
| Education | -0.22118 | 0.01058 | -20.9 | < 2e-16 | \*\*\* |
| Sex | 0.20091 | 0.02116 | 9.493 | < 2e-16 | \*\*\* |
| HvyAlcoholConsump | -0.88849 | 0.05652 | -15.721 | < 2e-16 | \*\*\* |
| AnyHealthcare | -0.01685 | 0.05218 | -0.323 | 0.74677 |  |

*P<.05 is statistically significant*

Figure 11: Evaluation metric comparison for the three machine learning classifier

Table 2: Comparison of model prediction using accuracy, precision, recall and F-measure

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluation Metrics | Random forest | Logistic Reg. | Decision Tree |
| Accuracy | 0.7204 | 0.7205 | 0.7186 |
| Precision | 0.6986 | 0.71 | 0.7022 |
| Recall | 0.7713 | 0.7468 | 0.7554 |
| F-Measure | 0.7331 | 0.7269 | 0.7278 |

# Discussion

The Random Forest Model had a commendable recall score (0.7713), indicating its effectiveness in identifying a significant portion of individuals with diabetes. This is good for timely detection and screening of diabetes, as overlooking true positive cases may result in delayed diagnosis and treatment, potentially exacerbating the condition and heightening the risk of complications such as cardiovascular diseases, neuropathy, and kidney issues. Early detection and intervention play a crucial role in effectively managing diabetes and staving off the onset of associated health concerns. However, the model's slightly lower precision score (0.6986) suggests a propensity for generating a greater number of false positive cases, wherein non-diabetic individuals are erroneously categorized as having diabetes. While false positives can prompt unnecessary further testing or treatments, their impact is generally less severe compared to missing true positive cases in diabetes diagnosis. Despite this, the Random Forest model's robust overall performance underscores its potential as a valuable tool for healthcare professionals and screening programs. Its capability to identify individuals at risk of diabetes early on enables timely implementation of lifestyle interventions, medication, and monitoring protocols.

The Logistic Regression model's elevated precision score (0.71) indicates its proficiency in accurately identifying individuals without diabetes (true negatives), potentially lessening the need for unnecessary follow-up tests or treatments, which can pose financial and logistical challenges for healthcare systems and patients. However, the diminished recall score (0.7468) suggests a higher likelihood of overlooking true diabetic cases compared to the Random Forest model. In diabetes diagnosis, missing true positive cases can result in the progression of the condition and heightened risks of complications due to delayed treatment. The Logistic Regression model's balance between precision and recall necessitates careful assessment based on the specific priorities and constraints of the healthcare environment. It might be suitable for situations where minimizing false positives holds greater significance, such as resource-constrained settings or when the expenses associated with follow-up testing or treatment are notably high.

Despite exhibiting the lowest overall performance among the three models, the Decision Tree model's potential lies in its interpretability and simplicity. Decision trees offer transparent and easily understandable models, which can prove beneficial in clinical decision support systems or when elucidating the decision-making process to healthcare professionals or patients. Nevertheless, the inferior performance metrics indicate that the Decision Tree model may not be the most precise option for diabetic diagnosis and prediction in this specific scenario.

In the context of diabetic diagnosis and prediction, selecting the most appropriate model requires a thorough evaluation of the trade-offs between accuracy, precision, recall, and interpretability. The Random Forest model emerges as a promising choice, striking a good balance across these metrics. However, the Logistic Regression model could also be a viable option if minimizing false positives takes precedence.

The results obtained from the logistic regression analysis highlight several key factors that have a significant impact on the likelihood of being diagnosed with diabetes. These findings carry important implications for the diagnosis, prediction, and prevention of diabetes. The positive relationship between BMI and diabetes underscores the significance of maintaining a healthy body weight through lifestyle changes, such as adopting a balanced diet and engaging in regular physical activity. Early identification of individuals with high BMI and targeted interventions to promote weight management could play a crucial role in preventing or delaying the onset of diabetes. Similarly, the positive relationship between high blood pressure and diabetes emphasizes the importance of comprehensive screening and management for both conditions, as they often coexist and exacerbate each other's complications. The inverse association between fruit consumption and diabetes risk supports the protective effects of a diet rich in fruits and vegetables, which provide essential nutrients and antioxidants beneficial for glycemic control and overall metabolic health. This underscores the importance of promoting a healthy diet as part of a comprehensive diabetes prevention and management approach. Furthermore, the negative association between educational attainment and the probability of receiving a diabetes diagnosis underscores how socioeconomic and environmental factors may shape the onset and control of the condition. Lower educational attainment is frequently linked with restricted access to health knowledge, diminished understanding of health matters, and inadequate resources to embrace healthier lifestyle habits.

The association of older age with increased diabetes risk was not surprising. This may be due to factors such as decreased insulin sensitivity and lifestyle changes. This highlights the importance of regular screening and early detection efforts, particularly among older populations, to identify and manage diabetes before complications arise. The observation that males have a higher likelihood of diabetes diagnosis than females is in line with epidemiological evidence showing higher diabetes prevalence among men in certain populations. However, it is crucial to recognize that gender differences in diabetes risk can vary across different ethnic and socioeconomic groups, emphasizing the need for tailored prevention and management strategies.

# Recommendations

Feature engineering may improve the quality of the logistic regression model. This may be done by transforming existing features to better represent the underlying problem and improve the model's predictive power. This can include techniques such as feature selection, feature scaling, principal component analysis, and creating interaction terms or polynomial features. The selection of better predictor variables may increase the accuracy of the model.

Techniques such as L1 (Lasso) or L2 (Ridge) regularization aid in mitigating overfitting by appending a penalty term to the cost function, thereby diminishing the coefficients of less significant features toward zero (Li, 2017). Adjusting the strength of regularization can enhance the model's ability to generalize and select features effectively (Jung, 2010).

For Decision trees, The selection of the splitting criterion, such as Gini impurity or entropy, influences the structure and effectiveness of the decision tree. Different criteria may be better suited to varying dataset characteristics or problem types (Arekhi, 2012).

Tuning the number of features considered for splitting at each node in a random forest can improve the model's performance by reducing the correlation between trees and increasing the diversity of the ensemble. Random forest models may be improved by using multiple attribute evaluation measures and weighted voting. Experiment with hyperparameter tuning, such as adjusting the number of trees in the forest, the maximum depth of the trees, or the minimum number of samples required to split a node (Probst, 2019). This can help optimize the model's performance.

# Conclusion

This study assessed the efficacy of three machine learning models - Random Forest, Logistic Regression, and Decision Tree in diagnosing diabetes utilizing lifestyle and related health risk as predictors. The Random Forest model showcased the most robust overall performance, striking a commendable balance between precision and recall. The Logistic Regression model had the highest precision, demonstrating proficiency in accurately discerning non-diabetic individuals. While the Decision Tree model had lower performance metrics, its interpretability and simplicity could be invaluable in clinical decision support systems.

The findings from the logistic regression analysis underscore the significance of modifiable risk factors such as high BMI, elevated blood pressure, excessive alcohol consumption, and inadequate fruit intake in influencing diabetes risk. Additionally, non-modifiable factors like age and gender were observed to impact diagnosis likelihood. Addressing these factors via targeted interventions, lifestyle adjustments, and routine screening can contribute to effective diabetes prevention and management strategies. Moreover, the study underscores the potential influence of socioeconomic and environmental determinants, such as education level, on diabetes risk. Alleviating these factors through interdisciplinary collaborations, targeted health education initiatives, and policy measures can help mitigate health disparities and enhance overall population health outcomes.

Ultimately, the integration of these machine learning models into healthcare systems could serve as invaluable decision support tools, augmenting the expertise of healthcare professionals and facilitating early detection, prompt interventions, and effective diabetes management, thereby alleviating the burden of this chronic condition on individuals and healthcare systems.

# References

Daanouni, O., Cherradi, B., & Tmiri, A. (2019). Predicting diabetes diseases using mixed data and supervised machine learning algorithms. *Proceedings of the 4th International Conference on Smart City Applications*, 1–6. https://doi.org/10.1145/3368756.3369072

Emon, M. U., Keya, M. S., Kaiser, Md. S., Islam, Md. A., Tanha, T., & Zulfiker, Md. S. (2021). Primary Stage of Diabetes Prediction using Machine Learning Approaches. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 364–367. https://doi.org/10.1109/ICAIS50930.2021.9395968

Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A Machine Learning Approach to Predicting Diabetes Complications. *Healthcare*, *9*(12), 1712. https://doi.org/10.3390/healthcare9121712

Kumar, Y. J. N., Shalini, N. K., Abhilash, P. K., Sandeep, K., & Indira, D. (2019). *Prediction of Diabetes using Machine Learning*. *8*(7).

Llaha, O., & Rista, A. (2021). *Prediction and Detection of Diabetes using Machine Learning*.

Maniruzzaman, Md., Rahman, Md. J., Ahammed, B., & Abedin, Md. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1), 7. https://doi.org/10.1007/s13755-019-0095-z

NCD Risk Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 192 million participants. Lancet. 2016;387(10026):1377–96.

Zimmet, Paul, et al. "Diabetes mellitus statistics on prevalence and mortality: facts and fallacies." Nature Reviews Endocrinology 12.10 (2016): 616-622.

Bisandu, D. B., Datiri, D. D., Onokpasa, E., Thomas, G., Haruna, M. M., Aliyu, A., & Yakubu, J. Z. (2019). Diabetes prediction using data mining techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), 4(6), 103-111.

Blair, M. (2016). Diabetes mellitus review. Urologic nursing, 36(1).

Galloway, N., Blanchette, J. E., Faizullabhoy, A., Mobley-Bukstein, W., Rinker, J., & Isaacs, D. (2022). Exploring the Role of Precision Medicine in Diabetes Diagnoses. ADCES in Practice, 10(1), 26-32.

Cleophas, T. J., Zwinderman, A. H., Cleophas, T. J., & Zwinderman, A. H. (2013). Logistic regression for health profiling. Machine Learning in Medicine, 17-24.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).

Galloway, N., Blanchette, J. E., Faizullabhoy, A., Mobley-Bukstein, W., Rinker, J., & Isaacs, D. (2022). Exploring the Role of Precision Medicine in Diabetes Diagnoses. ADCES in Practice, 10(1), 26-32.

Capitaine, L., Genuer, R., & Thiébaut, R. (2021). Random forests for high-dimensional longitudinal data. Statistical methods in medical research, 30(1), 166-184.

董师师, & 黄哲学. (2013). A brief theoretical overview of random forests. *Journal of Integration Technology*, *2*(1), 1-7.

Joly, A., Geurts, P., & Wehenkel, L. (2014). Random forests with random projections of the output space for high dimensional multi-label classification. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14 (pp. 607-622). Springer Berlin Heidelberg.

Li, J., Zhou, H., Xie, P., & Zhang, Y. (2017, August). Improving the Generalization Performance of Multi-class SVM via Angular Regularization. In IJCAI (pp. 2131-2137).

Jung, Y. (2010). Regularization of Case Specific Parameters: A New Approach for Improving Robustness and/or Efficiency of Statistical Methods (Doctoral dissertation, The Ohio State University).

Arekhi, D. S. (2012). Assessment of the effectiveness of decision tree classification method for extracting landuses map by using sattellite data in Cham Gardalan Catchment Area. Geography and Territorial Spatial Arrangement, 2(4), 17-26.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301.

Akosa, J. (2017, April). Predictive accuracy: A misleading performance measure for highly imbalanced data. In Proceedings of the SAS global forum (Vol. 12, pp. 1-4). Cary, NC, USA: SAS Institute Inc..

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2), 612-619.

Robnik-Šikonja, M. (2004, September). Improving random forests. In European conference on machine learning (pp. 359-370). Berlin, Heidelberg: Springer Berlin Heidelberg.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.