MACHINE LEARNING ANALYSIS OF HEALTH AND LIFESTYLE FACTORS IN UNDERSTANDING DIABETES

Original Research Article

Abstract

Diabetes poses a significant global health challenge, with approximately 537 million adults living with the condition in 2021, a number expected to rise to 783 million by 2045. To enhance predictive accuracy and gain deeper insights into the factors contributing to diabetes, this study employed machine learning algorithms to predict diabetes risk factors using a dataset encompassing health and lifestyle variables. Six supervised machine learning algorithms, including Gradient Boosting, Logistic Regression, and Random Forest, among others, were assessed for their effectiveness in classifying diabetes status into two categories: diabetes and no diabetes. The study found that Gradient Boosting achieved the highest overall accuracy at 85%, demonstrating the best recall for diabetic cases at 57%. Meanwhile, Logistic Regression excelled in precision for non-diabetic cases at 94%. Key risk factors identified include general health status, blood pressure, body mass index, cholesterol levels, and age. Notably, the study uncovered that higher income and education levels were associated with increased diabetes risk, contradicting some existing literature and indicating the potential impact of lifestyle factors.

Keywords: insulin, Machine Learning Analysis in Health, DIABETES, lifestyle

1. Introduction

The pancreas in the human body produces a hormone called insulin when it senses a rise in blood sugar (Huising, 2020). Insulin helps the body cells convert the blood sugar to energy and stores the unused blood sugar as glycogen in the liver and muscles (Andima et al., 2022). By doing this, the blood sugar level is reduced and maintained between 70 and 100 mg/dL (Kapoor et al., 2020). Hence, when there is an impairment in insulin regulation, the blood glucose rises significantly, leading to the situation referred to as diabetes. Diabetes can be of two types: Type 1 and Type 2 (Fan et al., 2024). Type 1 diabetes is an autoimmune condition typically diagnosed in childhood or adolescence, destroying insulin-producing beta cells in the pancreas. Conversely, Type 2 diabetes, which accounts for about 90-95% of diabetes cases, is largely influenced by lifestyle choices and is often diagnosed in adulthood. Diabetes has emerged as a significant global health challenge. Statistics by the International Diabetes Federation (IDF) show that approximately 537 million adults were living with diabetes globally in 2021, with an estimated value of 783 million by 2045 (Hossain et al., 2024). This escalating prevalence emphasises the critical need to understand the disease and its underlying risk factors.

Management of diabetes starts with a proper understanding of the health and lifestyle of the individuals, which includes diet, physical activity, and mental health. Salvia and Quatromoni (2023) argued that a balanced diet rich in whole grains, lean proteins, healthy fats, and ample fruits and vegetables can significantly improve glycemic control. Research by Dominguez et al. (2021) indicates that dietary patterns such as the Mediterranean diet, which emphasises whole foods and healthy fats, can lower blood glucose levels and improve overall metabolic health. Furthermore, regular exercise enhances insulin sensitivity and aids in weight control (Sgro et al., 2021). The American Diabetes Association recommends at least 150 minutes of moderate-intensity aerobic activity per week for adults with diabetes (Silva et al., 2020). Diabetic conditions can lead to psychological stress, anxiety, and depression, all of which can negatively impact blood sugar control. Research has shown that individuals who receive mental health support alongside diabetes management exhibit better adherence to treatment plans and improved glycemic control (Oyedeji et al., 2022; Heilbrun & Drossos, 2020).

The integration of machine learning (ML) into health analytics represents a transformative shift in the way healthcare data is utilised (Almushayqih et al., 2024; Zhang et al., 2022). Mathematical modelling using the deterministic approach has been found to have the limitation of not being able to predict accurately unless it is fitted with existing data (Oke, 2017; Bada et al., 2021). With the exponential growth of health-related data, traditional analytical methods often fall short in addressing the complexities of diseases such as diabetes. Machine learning algorithms are capable of identifying patterns, predicting outcomes, and personalising treatment approaches based on individual patient data. Rajula et al. (2020) demonstrated that ML models can outperform traditional statistical methods in accuracy and predictive power, making them invaluable tools in preventive healthcare. Machine learning has been applied to risk assessment, diagnosis and treatment of diabetes.

Predictive models in diabetes have focused on isolated risk factors, such as obesity (Wang et al., 2020) or sedentary behaviour (Li et al., 2022), without considering the multifaceted interactions between various health and lifestyle elements. Li et al. (2022) and many other studies utilise linear regression models that fail to capture the non-linear relationships inherent in these factors. Meanwhile, even though obesity is a well-established risk factor, its impact may vary significantly based on physical activity levels, dietary habits, and psychological well-being. This necessitates a more integrative approach to understanding diabetes risk. Another critical gap is the reliance on homogeneous datasets that lack demographic diversity, primarily concentrating on specific populations. This homogeneity can lead to models that are not generalisable, particularly for underrepresented groups. A comprehensive analysis of how lifestyle factors affect diabetes risk across different demographics, such as age, gender, and socioeconomic status, remains underexplored.

This study aims to address these specific gaps by employing machine learning algorithms that integrate a wide range of health and lifestyle factors. By utilising a dataset from over 250,000 records from various demographic backgrounds, the study analyses interactions among variables, such as the combined effects of BMI, physical activity, and dietary intake on diabetes status. A key novelty of this research lies in its focus on identifying features that significantly contribute to diabetes risk, as well as those that can potentially reduce it. This study aims to answer the following questions;

1. Can machine learning algorithms effectively classify diabetes status, and which algorithms perform best in this context?

2. How does demographic diversity impact the relationship between health and lifestyle factors and diabetes status?

3. Which health and lifestyle factors are most significantly associated with diabetes status?

4. How do interactions between different health and lifestyle factors influence diabetes risk?

2. Methodology

This section begins with the collection and preparation of the dataset, followed by exploratory data analysis (EDA) to understand the distribution of variables and identify any underlying patterns. Once the dataset is prepared, six machine learning algorithms are applied to classify diabetes status based on health and lifestyle factors. The selected algorithms include Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting. Each of these algorithms is chosen for its ability to handle binary classification and to provide insights into the importance of different features.

2.1 Dataset Description

The dataset used in this study is derived from a comprehensive health and lifestyle survey from 253,680 patients, providing a diverse representation of individuals across various demographic backgrounds. This dataset includes multiple features that capture critical aspects related to diabetes, including demographic information, lifestyle choices, and health indicators. The dataset was obtained from the UCI repository titled “CDC Diabetes Health Indicators” (https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators). The dataset consists of 21 features and one target variable and is described as follows;

Target Variable:

The target variable is diabetes. The dataset classifies individuals into three categories based on their diabetes status: 0 = no diabetes, 1 = Prediabetes, and 2 = Diabetes.

Features of the Dataset

1. **HighBP**: This is a binary variable representing whether the individual has high blood pressure: 0 = no, 1 = yes.

2. **HighChol**: This is a binary variable representing whether the individual has high cholesterol: 0 = no, 1 = yes.

3. **CholCheck**: Indicates if the individual has had a cholesterol check in the last five years. It is a binary variable: 0 = no, 1 = yes.

4. **BMI**: The Body Mass Index (BMI) is a variable obtained using the formula

$$BMI = \frac{\text{weight}}{(\text{height})^2}(kg/m^2)$$

The result is approximated to the nearest whole number to keep it as an integer variable.

5. **Smoker**: This variable answers the question of whether the individual has smoked at least 100 cigarettes in their lifetime. It is a binary variable: 0 = no, 1 = yes.

6. **Stroke**: Indicates whether the individual has ever been told they had a stroke (0 = no, 1 = yes).

7. **HeartDiseaseorAttack**: Indicates the presence of coronary heart disease or myocardial infarction (0 = no, 1 = yes).

8. **PhysActivity**: Indicates the participation in physical activity in the past 30 days (0 = no, 1 = yes).

9. **Fruits**: Indicates daily consumption of fruit (0 = no, 1 = yes).

10. **Veggies**: Indicates daily consumption of vegetables (0 = no, 1 = yes).

11. **HvyAlcoholConsump**: Indicates heavy alcohol consumption based on defined thresholds (0 = no, 1 = yes).

12. **AnyHealthcare**: Indicates whether the individual has any form of health care coverage (0 = no, 1 = yes).

13. **NoDocbcCost**: Indicates whether cost prevented the individual from seeing a doctor in the past year (0 = no, 1 = yes).

14. **GenHlth**: Self-reported general health on a scale of 1 to 5 (1 = excellent, 5 = poor).

15. **MentHlth**: Number of days in the past 30 that mental health was not good (scale 1-30).

16. **PhysHlth**: Number of days in the past 30 that physical health was not good (scale 1-30).

17. **DiffWalk**: Indicates serious difficulty in walking or climbing stairs (0 = no, 1 = yes).

18. **Sex**: Gender of the individual (0 = female, 1 = male).

19. **Age**: Age category based on a 13-level scale (1 = 18-24 years old; 2 = 25-29 years old; 3 = 30-34 years old; 4 = 35-39 years old; 5 = 40-44 years old; 6 = 45-49 years old; 7 = 50-54 years old; 8 = 55-59 years old; 9 = 60-64 years old; 10 = 65-69 years old; 11 = 70-74 years old; 12 = 75-79 years old; 13 = 80 or older).

20. **Education**: Education level based on a 6-point scale (1 = Never attended school; 2 = Some primary education; 3 = Completed primary education; 4 = Some secondary education; 5 = Completed secondary education; 6 = College graduate).

21. **Income**: Income level based on an 8-point scale (1 = Less than $10,000; 2 = $10,000 - $24,999; 3 = $25,000 - $39,999; 4 = $40,000 - $49,999; 5 = $50,000 - $59,999; 6 = $60,000 - $74,999; 7 = $75,000 - $99,999; 8 = $100,000 or more).

2.2 Data Preprocessing

This section outlines the steps taken to clean and prepare the dataset. To ensure data integrity, the dataset was checked for missing values (using the code snippet in Figure 1), and no missing data was found (as shown in Figure 2).

```python
# 1. Data Preprocessing
# Check for missing values
print(data.isnull().sum())
```

Figure 1: Python code for checking for missing values

```
Diabetes                0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64
```

Figure 2: Output for missing values

To convert all variables to a similar scale, three features –BMI, MentHlth and PhysHlth– are standardised using the code snippet in Figure 3. The standardisation is carried out using the formula

$$z = \frac{x - \mu}{\sigma},$$

where x is the original feature value, μ is the mean of the feature, and σ is the standard deviation of the feature.

```python
numerical_features = ['BMI', 'MentHlth', 'PhysHlth']
scaler = StandardScaler()
data[numerical_features] = scaler.fit_transform(data[numerical_features])
```

Figure 3: Standardisation of the features

Machine learning algorithms require numerical inputs for learning. Hence, we transform the categorical variables into numeric variables by encoding them using the code snippet in Figure 4. The categorical variables are Sex, Age, Education, and Income.

```python
# Encode categorical features
categorical_features = ['Sex', 'Age', 'Education', 'Income']
label_encoders = {}
for col in categorical_features:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le
```

Figure 4: Encoding Categorical Features

The target variable was originally classified into three classes (0-no diabetes, 1-prediabetes, 2-diabetes), but it is safe to consider both the diabetic and the prediabetic cases as diabetic together. Using the code snippet in Figure 5, the prediabetes class and the diabetic class are combined into a single diabetic class due to the imbalance in the dataset. This means the target variable is now a binary variable that can take only 0 or 1.

```python
data['Diabetes'] = data['Diabetes'].apply(lambda x: 1 if x > 0 else 0)
data.to_csv('diabetes_health_indicators_binary.csv', index=False)
```

Figure 5: Convert prediabetic cases to diabetic

Finally, the dataset is split into Features and Target variables using the code snippet in Figure 6.

```
# Split the data into features and target variable
X = data.drop('Diabetes', axis=1)
y = data['Diabetes']
```

Figure 6: Data Split

2.3 Data Distribution

The dataset consists of 22 features and 1 target variable of 253,680 patients' records with no missing value. There are 213703 non-diabetic patients and 39977 diabetic patients. The summary statistics are shown in Table 1, and the mean for diabetics is 0.15788, which contributes to 15.79% of the dataset. Hence, 15.79% of the dataset are diabetic, while the remaining 84.21% represent non-diabetic cases.

Table 1: Summary Statistics

	Count	mean	min	25%	50%	75%	max
Diabetes	253680	0.157588	0	0	0	0	1
HighBP	253680	0.429001	0	0	0	1	1
HighChol	253680	0.424121	0	0	0	1	1
CholCheck	253680	0.96267	0	1	1	1	1
BMI	253680	28.38236	12	24	27	31	98
Smoker	253680	0.443169	0	0	0	1	1
Stroke	253680	0.040571	0	0	0	0	1
HeartDiseaseorAttack	253680	0.094186	0	0	0	0	1
PhysActivity	253680	0.756544	0	1	1	1	1
Fruits	253680	0.634256	0	0	1	1	1
Veggies	253680	0.81142	0	1	1	1	1
HvyAlcoholConsump	253680	0.056197	0	0	0	0	1
AnyHealthcare	253680	0.951053	0	1	1	1	1
NoDocbcCost	253680	0.084177	0	0	0	0	1
GenHlth	253680	2.511392	1	2	2	3	5
MentHlth	253680	3.184772	0	0	0	2	30
PhysHlth	253680	4.242081	0	0	0	3	30
DiffWalk	253680	0.168224	0	0	0	0	1
Sex	253680	0.440342	0	0	0	1	1
Age	253680	8.032119	1	6	8	10	13
Education	253680	5.050434	1	4	5	6	6
Income	253680	6.053875	1	5	7	8	8

The pie chart in Figure 7 shows the age distribution for the diabetic class. The pie chart shows that 12.0% is from Group 8, 16.1% is from Group 9, 18.1% is from Group 10, 14.1% is from Group 11 and 9.6% is from Group 12. This accounts for 70.2% of the diabetic patients. Hence, diabetes is more frequent among people from the age of 55 and older. Furthermore, Figure 8 shows that out of the diabetic population, 31.1% have some secondary education, 29.2% completed secondary education, and 29.7% have completed college. This contributes to a total of 90% of diabetic patients who have been to secondary school or higher.

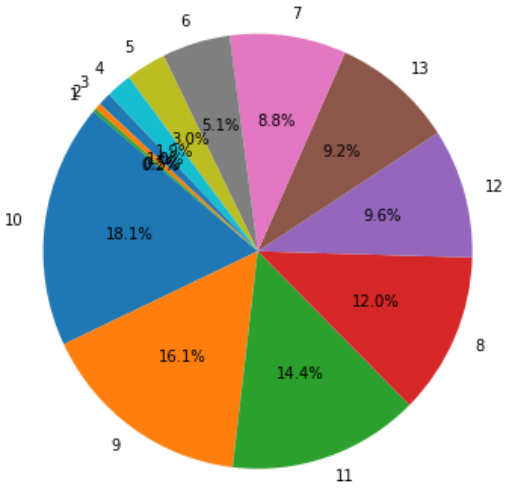


Figure 7: Age Distribution for Diabetic Patients

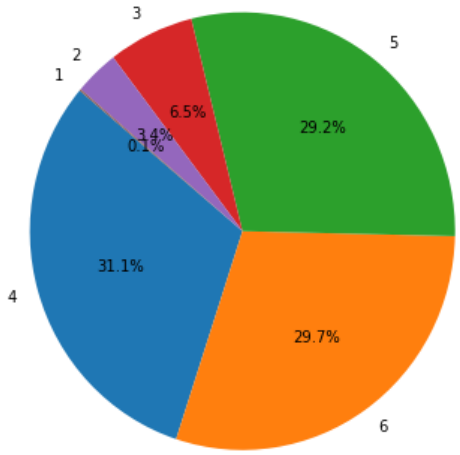


Figure 8: Education Level Distribution for Diabetic Patients

The histogram and pie chart of Figure 9 shows that the number of diabetic patients increases with increasing income. By summing the contributions in the pie chart, 74.6% of diabetic patients earn at least $40,000. Finally, Figure 10 shows the distribution of the binary gender among the diabetic class. It shows that 52.6% of the diabetic class are female, while 47.4% are male.

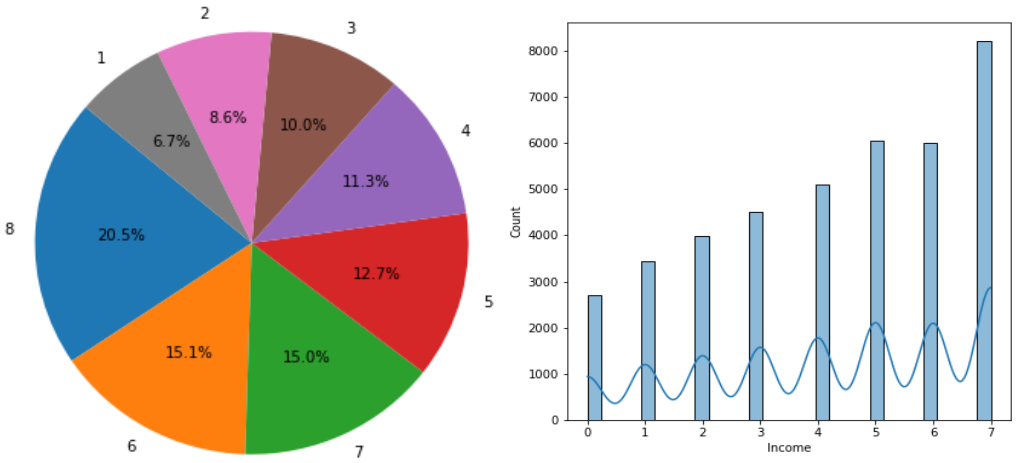


Figure 9: Income Distribution for Diabetic Patients

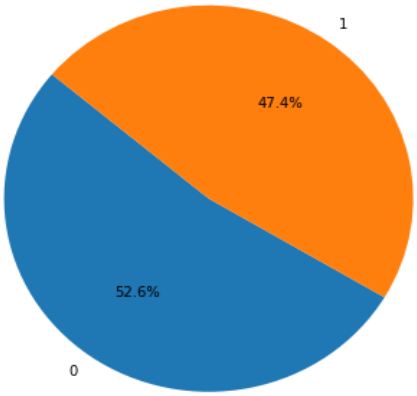


Figure 10: Gender Distribution among Diabetic Patients

The correlation matrix in Figure 11 shows the correlation between the 22 features. The figure shows a strong positive correlation between the number of days the patient's physical health was not good and the patient's self-

reported health conditions.

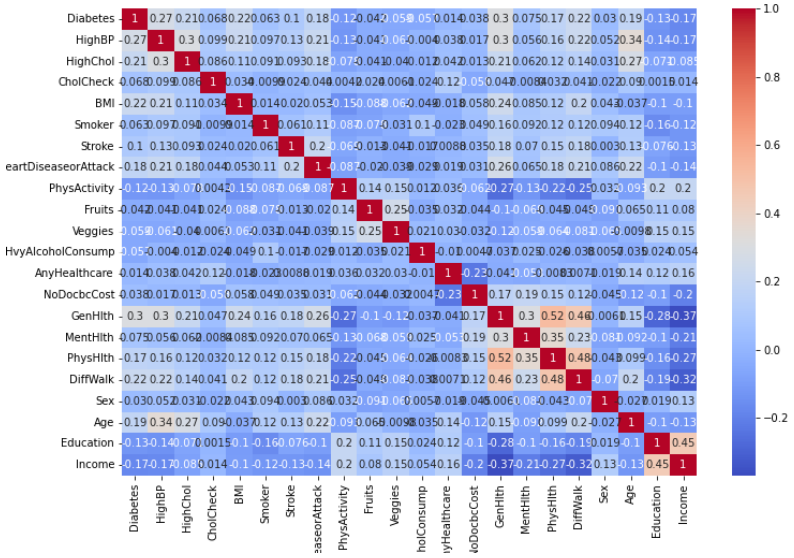


Figure 11: Correlation Matrix

2.4 Machine Learning Techniques

Machine learning algorithms are used to train a dataset so that the computer can learn from the dataset and formulate a model that can be used to predict the outcome of new inputs. In this study, six machine learning algorithms are adopted to train the dataset. The algorithms are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBoost), *k*-Nearest Neighbours (*k*NN) and Naive Bayes (NB) algorithms were chosen because they can handle binary classification. When compared to other algorithms, such as Support Vector Machines (SVM), Neural Networks, and AdaBoost, the chosen methods offer distinct benefits that make them more appropriate for our analysis. LR is a simple binary classification which is effective and easily interpretable. Unlike the Support Vector Method (SVM), which can be computationally intensive, especially with large datasets, LR provides probabilistic outputs and is less resource-demanding, making it a practical choice for this study. DT algorithm was chosen for its intuitive and easy-to-visualize nature, which aids in understanding the decision-making process. While Neural Networks can model complex relationships in the data, they often act as "black boxes" with limited interpretability. Meanwhile, DT provides clear, understandable rules, making them valuable for this analysis. RF is an ensemble method that has the ability to reduce overfitting by averaging the results of multiple decision trees, resulting in improved accuracy and robustness. Single classifiers like SVM or a single Neural Network may overfit the data. GBoost was chosen for its high predictive accuracy and ability to handle complex relationships in the data. Compared to AdaBoost, another boosting technique, GBoost, often yields better performance by minimizing loss functions and more effectively addressing bias and variance. *k*-Nearest Neighbours (*k*NN) was included due to its non-parametric nature, making no assumptions about the data distribution. It is simple to implement and effective, especially in smaller datasets where relationships among data points are significant. Unlike SVM or Neural Networks, which require extensive parameter tuning and longer training times, *k*NN provides quick, reliable results. NB algorithm was chosen for its strong independent assumptions between features and its ability to perform well in real-world scenarios. While algorithms like Neural Networks or SVM may require extensive computational resources and complex tuning, NB is computationally efficient and robust. In addition to the strengths of the chosen algorithms, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting have the “Feature Importance” attribute that can be used to estimate the contribution of each feature to the chance of a patient being diabetic.

The performance of the algorithms is measured using the confusion matrix, precision, recall, F1-score, accuracy and cross-validation technique (Almushaygih et al., 2024). Figure 12 shows the confusion matrix, with the items explained below;

True positive: the number of non-diabetic classifications that are actually non-diabetic.

False positive: the number of non-diabetic classifications that are not actually diabetic.

False negative: the number of diabetic classifications that are not actually diabetic.

True negative: the number of diabetic classifications that are actually diabetic.

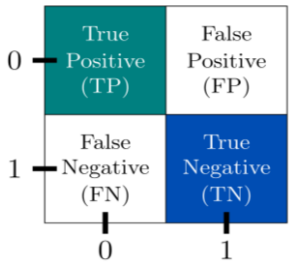


Figure 12: Confusion Matrix General Form

Given a positive classification, the precision measures the accuracy of the classification and is defined as;

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

The recall measures the ability of the model to capture all positive instances and is defined as the ratio of the true positive to the total positive.

$$\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}.$$

As a balance between the recall and the precision, the F1 score is defined as the harmonic mean of the precision and recall as

$$\text{F1 Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The accuracy is defined as

$$\text{accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

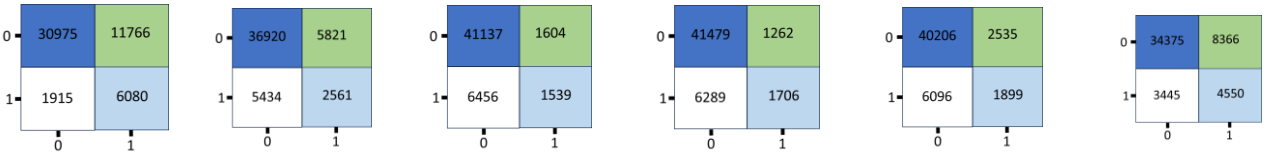
and it measures the overall correctness of the algorithm.

One of the challenges with machine learning algorithms is that it is not always very certain whether the accuracy, precision or recall is reliable. Cross-validation is a technique that helps to assess the performance and generalisation of predictive models (Szeghalmy & Fazekas, 2023). In the cross-validation technique, the dataset is divided into several subsets, commonly referred to as folds (Allgaier & Pryss, 2024). In a k-fold cross-validation scenario; the data is split into k equally sized folds. The model is then trained on $k - 1$ of these folds and validated on the remaining one. This process is repeated k times, with each fold serving as the validation set exactly once. The performance metrics from each iteration are then averaged, resulting in a robust estimate of the model's performance.

3. Results

3.1 Analysis of Results

By following the work of Almushaygih et al. (2024), the dataset is divided it into two subsets; 80 percent for training and 20 percent for testing. A total of 50736 examples were involved in the testing, out of which 42741 are nondiabetic and 7995 are diabetic. Figure 13 shows the confusion matrices for all the six algorithms. Gradient Boosting has the highest True Positive and the highest False Positive, while Logistic Regression has the lowest True Positive and the lowest False Positive.



(a) Logistic Regression (b) Decision Tree (c) Random Forest (d) Gradient Boosting (e) *k*-Nearest Neighbors (f) Naive Bayes

Figure 13: Confusion Matrices for the Six Algorithms

The performance metrics used to measure the performance of the algorithms are the Precision, Recall, F1-score and Accuracy. The results are arranged in order to increase accuracy in Table 2. For the classification of the nondiabetic, the Logistic Regression has the highest Precision (94%), while Gradient Boosting has the highest Recall (97%) and F1-score (92%). For the classification of diabetic, Gradient Boosting has the highest Recall (57%), while Logistic Regression has the highest Recall (76%) and F1-score (47%). In all, the accuracy of the Gradient Boosting (85%) is the highest among the algorithms. It is also important to note that the performance of Random Forest and Gradient Boosting is close. Hence, the models from Logistic Regression will be used to model the chance of a patient being non-diabetic. At the same time, the Gradient Boosting and Random Forest will be compared to model the chance of a patient being diabetic.

Table 2: Performance Metrics for the 6 Algorithms

		Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Logistic Regression	0	94	72	82	73
	1	34	76	47	
Naive Bayes	0	91	80	85	77
	1	35	57	44	
Decision Tree	0	87	86	87	78
	1	31	32	31	
k-Nearest Neighbors	0	87	94	90	83
	1	43	24	31	
Random Forest	0	86	96	91	84
	1	49	19	28	
Gradient Boosting	0	87	97	92	85
	1	57	21	31	

Using cross-validation of 5 folds, the accuracies of the models are validated, and the results are shown in Table 3. The results validate the results in Table 2, and thus, the Gradient Boosting and Random Forest models perform better than the remaining four models. Hence, the models from the Random Forest and Gradient Boosting are used for the Feature Importance to estimate the chance of diabetic condition. In contrast, the Logistic Regression model is used to estimate the chance of non-diabetic conditions.

Table 3: Cross-Validation Accuracy

Model	**Score**
Logistic Regression	72.91%
Naive Bayes	76.63%
Decision Tree	77.77%
k-Nearest Neighbors	83.07%
Random Forest	83.98%
Gradient Boosting	85.11%

Finally, the Feature Importance generated from the three important algorithms are shown in Figures 14 – 16. By checking the significance of the features in Figure 14, it shows that regular cholesterol checks, blood pressure, cholesterol level, general health condition and body mass index are important in determining that a patient is non-diabetic. However, Figures 15 and 16 show that General Health, Blood Pressure, Body Mass Index, Cholesterol Level, Age, Income and Education are risk factors in determining the chance of diabetes in a patient.

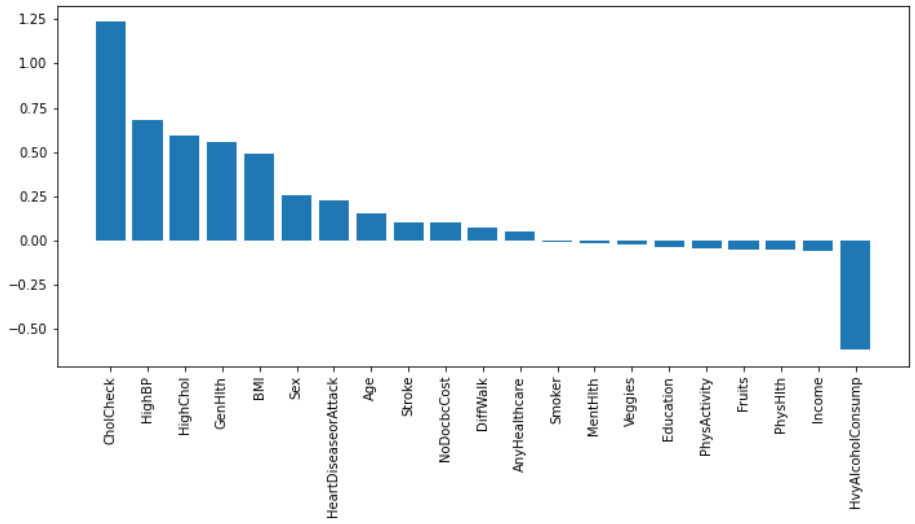


Figure 14: Feature Importance by Logistic Regression

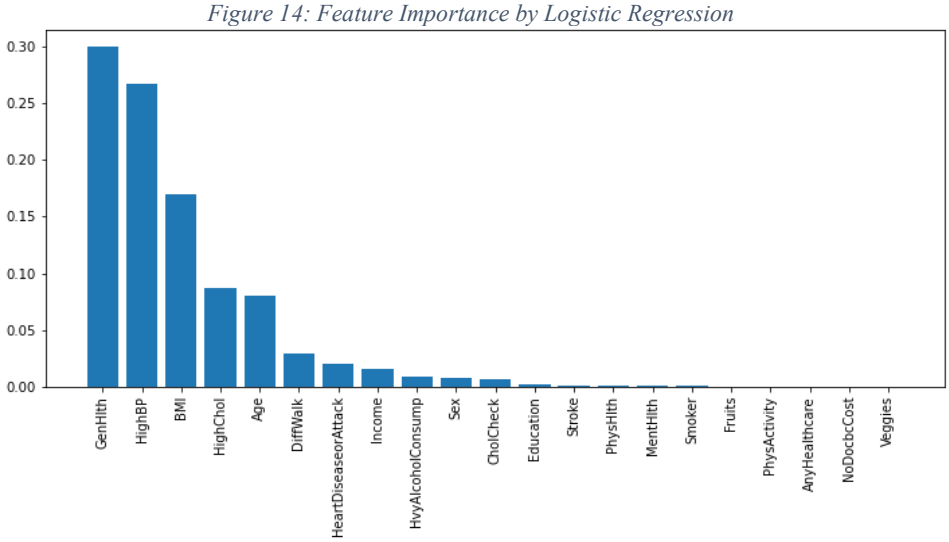


Figure 15: Feature Importance by Gradient Boosting

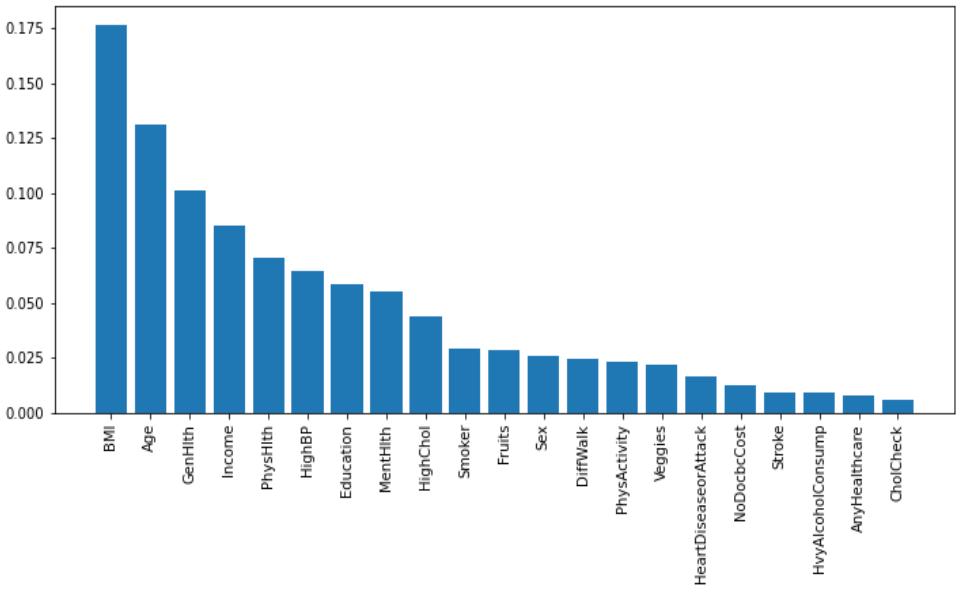


Figure 16: Feature Importance by Random Forest

3.2 Discussion

The performance of the machine learning models varied across different metrics, revealing the strengths and limitations of each approach. The Gradient Boosting model achieved the highest overall accuracy at 85%, with the best recall for diabetic cases at 57%. This model was particularly effective in identifying true diabetic cases, although it also had a higher false positive rate. This suggests that while gradient boosting is adept at detecting diabetes, it may require further refinement to reduce false positives. On the other hand, the Logistic Regression demonstrated high precision at 94% for non-diabetic cases and achieved reasonable overall accuracy at 73%. This model was effective in minimising false positives, making it a reliable choice for identifying non-diabetic individuals. However, its lower recall for diabetic cases indicates that it may miss some true positive cases, highlighting a trade-off between precision and recall. The Random Forest model performed closely to Gradient Boosting, with high overall accuracy at 84% and balanced performance across metrics. This suggests that random forest is a robust model for predicting diabetes status, offering a good balance between identifying true positive and true negative cases.

Based on the models, the risk factors were identified. General health status consistently emerged as a top predictor, indicating that overall well-being plays a crucial role in diabetes risk. General health status depends on several factors, such as blood pressure, body mass index, and cholesterol level. Perhaps this explains why high blood pressure, BMI, and high cholesterol levels also have a high contribution to increasing the chance of diabetes. According to past studies, High blood pressure (Do et al., 2023), BMI (Ebrahimpour et al., 2020), and cholesterol level (Song et al., 2021) are significant indicators of diabetes. Hence, the outcome of the models in this study agrees with the existing literature.

Age also emerged as a significant factor, with older age groups more likely to develop diabetes. This aligns with existing literature on age-related diabetes risk and suggests that age-specific strategies may be necessary for effective diabetes prevention (Bahour et al., 2022). It has been shown that the production of insulin declines as individuals grow older. Moreover, as age increases, there is an accumulation of lifestyle. This probably explains why age is a risk factor for diabetes.

Higher-income levels were associated with higher diabetes risk, suggesting that socioeconomic factors play a critical role in health outcomes. As opposed to many perspectives on diabetes (such as the one held by Sieglie et al., 2020), this study shows that increasing income could contribute to increasing chance of diabetes. People with high incomes are able to afford more than those with lower incomes can afford, and as a result, high-income earners may also indulge in high-sugar food due to the sweet tastes and satisfaction it brings. Increasing education levels were linked to increased diabetes risk. This is also against the popular opinion that increasing education also increases awareness of diabetes (Sieglie et al., 2020). However, increasing education also promotes staying in the same position for a long time, either to sit down to read and write or to remain standing at the same spot in the laboratory. These activities do not encourage blood sugar usage; rather, they leave more sugar in the blood.

4. Recommendations

Based on the analysis, several recommendations can be made to improve diabetes prevention and management. Regular health check-ups, including cholesterol and blood pressure monitoring, should be encouraged, as these are significant predictors of diabetes risk. Promotion of healthy lifestyle choices and weight management strategies is essential, given the impact of BMI on diabetes risk. Age-specific diabetes prevention programs should be developed to recognise the increased risk in older age groups.

5. Conclusion

In this study, we evaluated the performance of six machine learning algorithms on a dataset containing health and lifestyle factors to predict diabetes status. The study identified the most effective models for classifying non-diabetic and diabetic cases. It ultimately provided insightful information on the features that influence the chance of diabetes in a patient. Gradient Boosting demonstrated the highest overall accuracy (85%) and the highest recall (97%) and F1-score (92%) for non-diabetic classification. It also showed the highest recall (57%) for diabetic classification. Logistic Regression excelled in precision (94%) for non-diabetic classification and had a notable F1-score (47%) for diabetic classification, making it suitable for modelling non-diabetic cases. Random Forest performed closely to Gradient Boosting, validating its robustness for diabetic predictions. Cross-validation results confirmed the superiority of Gradient Boosting (85.11%) and Random Forest (83.98%) over the other models, reinforcing their reliability. For non-diabetic predictions, important features included regular cholesterol checks, high blood pressure, cholesterol level, general health condition, and body mass index. For diabetic predictions, critical risk factors were general health, blood pressure, body mass index, cholesterol level, age, income, and

education.

Future work can focus on further enhancing these models, addressing potential class imbalances, and integrating additional health metrics to improve predictive accuracy. This approach offers a valuable framework for healthcare practitioners to identify at-risk individuals and tailor interventions accordingly, ultimately contributing to better diabetes management and prevention strategies.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References

Allgaier, J., & Pryss, R. (2024). Cross-Validation Visualized: A Narrative Guide to Advanced Methods. *Machine Learning and Knowledge Extraction*, *6*(2), 1378-1388.

Almushayqih, J., Oke, A. S., & Juma, B. A. (2024). Analysis of patient data to explore cardiovascular risk factors. *Mathematical Modelling and Numerical Simulation with Applications*, *4*(2), 133-148.

Andima, R. N., Mutuku, W. N., Farai, N., Awuor, K., & Oke, A. S. (2022). Mathematical Modelling of Diabetes under a Constrained Hospitalisation Resources. *Open Access Library Journal*, *9*(10), 1-14.

Bada, O. I., Oke, A. S., Mutuku, W. N., & Aye, P. O. (2021). Analysis of the dynamics of SI-SI-SEIR avian influenza A (H7N9) epidemic model with re-infection. *Earthline Journal of Mathematical Sciences*, *5*(1), 43-73.

Bahour, N., Cortez, B., Pan, H., Shah, H., Doria, A., & Aguayo-Mazzucato, C. (2022). Diabetes mellitus correlates with increased biological age as indicated by clinical biomarkers. *Geroscience*, 1-13.

Do, D. V., Han, G., Abariga, S. A., Sleilati, G., Vedula, S. S., & Hawkins, B. S. (2023). Blood pressure control for diabetic retinopathy. *Cochrane Database of Systematic Reviews*, (3).

Dominguez, L. J., Di Bella, G., Veronese, N., & Barbagallo, M. (2021). Impact of Mediterranean diet on chronic non-communicable diseases and longevity. *Nutrients*, *13*(6), 2028.

Ebrahimpour, S., Zakeri, M., & Esmaeili, A. (2020). Crosstalk between obesity, diabetes, and alzheimer's disease: Introducing quercetin as an effective triple herbal medicine. *Ageing research reviews*, p. *62*, 101095.

Fan, G., Zhang, B., Wang, J., Wang, N., Qin, S., Zhao, W., & Zhang, J. (2024). Accurate construction of NIR probe for visualizing HClO fluctuations in type I, type II diabetes and diabetic liver disease assisted by

theoretical calculation. *Talanta*, *268*, 125298.

Heilbrun, A., & Drossos, T. (2020). Evidence for mental health contributions to medical care in diabetes management: economic and professional considerations. *Current diabetes reports*, *20*, 1-7.

Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Science Reports*, *7*(3), e2004.

Huising, M. O. (2020). Paracrine regulation of insulin secretion. *Diabetologia*, *63*, 2057-2063.

Kapoor, R., Timsina, L. R., Gupta, N., Kaur, H., Vidger, A. J., Pollander, A. M., Jacobi, J., Khare, S. & Rahman, O. (2020). Maintaining blood glucose levels in range (70–150 mg/dL) is difficult in COVID-19 compared to non-COVID-19 ICU patients—a retrospective analysis. *Journal of Clinical Medicine*, *9*(11), 3635.

Li, D. D., Yang, Y., Gao, Z. Y., Zhao, L. H., Yang, X., Xu, F., ... & Su, J. B. (2022). Sedentary lifestyle and body composition in type 2 diabetes. *Diabetology & Metabolic Syndrome*, *14*(1), 8.

Oke, A. S. (2017). Convergence of differential transform method for ordinary differential equations. *Journal of Advances in Mathematics and Computer Science*, *24*(6), 1-17.

Oyedeji, A. D., Ullah, I., Weich, S., Bentall, R., & Booth, A. (2022). Effectiveness of non-specialist delivered psychological interventions on glycemic control and mental health problems in individuals with type 2 diabetes: a systematic review and meta-analysis. *International Journal of Mental Health Systems*, *16*(1), 9.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, *56*(9), 455.

Salvia, M. G., & Quatromoni, P. A. (2023). Behavioral approaches to nutrition and eating patterns for managing type 2 diabetes: a review. *American Journal of Medicine Open*, *9*, 100034.

Seiglie, J. A., Marcus, M. E., Ebert, C., Prodromidis, N., Geldsetzer, P., Theilmann, M., ... & Manne-Goehler, J. (2020). Diabetes prevalence and its relationship with education, wealth, and BMI in 29 low-and middle-income countries. *Diabetes care*, *43*(4), 767-775.

Sgro, P., Emerenziani, G. P., Antinozzi, C., Sacchetti, M., & Di Luigi, L. (2021). Exercise as a drug for glucose management and prevention in type 2 diabetes mellitus. *Current opinion in pharmacology*, *59*, 95-102.

Silva, V. R., Belozo, F. L., Pereira, R. M., Katashima, C. K., Cordeiro, A. V., Alves, J. F., ... & De Moura, L. P. (2020). The effects of ninety minutes per week of moderate intensity aerobic exercise on metabolic health in individuals with Type 2 Diabetes: a pilot study. *Journal of Rehabilitation Therapy*, *2*(2).

Song, Y., Liu, J., Zhao, K., Gao, L., & Zhao, J. (2021). Cholesterol-induced toxicity: An integrated view of the role of cholesterol in multiple diseases. *Cell metabolism*, *33*(10), 1911-1925.

Szeghalmy, S., & Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors*, *23*(4), 2333.

Wang, Y., Min, J., Khuri, J., Xue, H., Xie, B., Kaminsky, L. A., & Cheskin, L. J. (2020). Effectiveness of mobile health interventions on diabetes and obesity treatment and management: systematic review of systematic reviews. *JMIR mHealth and uHealth*, *8*(4), e15400.

Zhang, A., Xing, L., Zou, J., & Wu, J. C. (2022). Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, *6*(12), 1330-1345.