Classification and Prediction of Diabetes

WQD7001: PRINCIPLES OF DATA SCIENCE (GROUP 9)

Group Members	Student ID	Role
Lai Yuen Seng	22076493	Leader
Justin Ee Qing Sem	s2111464	Oracle
Leong Lip San	17217825	Detective
Chen You Hui	22113675	Maker
Wang Zi	22105299	Secretary

University of Malaya

Faculty of Computer Science & Information Technology

January 16, 2024

Contents

Chapter 1

Introduction and Project Details

1.1 Project Background

Diabetes is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both [2]. Diabetes is usually associated with long-term damage, dysfunction, and failure of different organs, such as the eyes, kidneys, nerves, heart, and blood vessels. Diabetes-related complications such as cardiovascular disease, kidney disease, nephropathy, blindness, and lower-extremity amputation are significant causes of increased morbidity and mortality among diabetic patients.

According to article [7], there were around 16.2 million diabetic individuals in the United States in 2005, and approximately 30% of the cases were undiagnosed. It is estimated that this number will grow to 48.3 million by 2050. These numbers will certainly create a substantial financial burden on the country's economy such as indirect costs of healthcare or unexpected disbursement due to reduced labour and productivity. Besides that, the framing of diabetes has moved from medical in 1993 to behavioural in 2001 then societal in 2013 which portrays modifiable risk factors for diabetes [8]. This has increased the importance and awareness of health monitoring nowadays.

Although medical science is rapidly growing, diabetes is still an incurable disease. Diabetes cannot be cured yet it is preventable based on the results of the controlled randomised trials stated in [9]. There is evidence which shows that intensive lifestyle interventions can effectively prevent or delay the onset of diabetes in high-risk individuals, by considering the main risk factors of diabetes.

In the early 2000s, there were several techniques for predicting diabetes, such as using statistical models and risk scores proposed in [13]. These models associate several attributes, such as age, BMI, waist circumference etc in scoring. Besides that, the emergence of Internet of Things (IoT) products such as wearable devices and sensors contributes valuable input of real-time health-related data. With relatively cheap computing power, we can adopt a machine learning approach to predict diabetes as an extension to the classical approaches [15].

As such, early diagnosis of diabetes using machine learning models can greatly benefit each individual such as early treatment to prevent complications like lower-extremity amputation, cardiovascular diseases etc and subsequently reduce the individual's financial burden and healthcare costs borne by the government.

1.2 Problem Statement

As mentioned in Section 1.1, prevention is better than cure [17]. However, the behaviour of diabetes factors changes over time and redefining the risk of factor requires to be updated time to time. Also, the rise of IoT technologies such as wearable health watch and body composition machine have provided convenience in measuring body and health data. We aim to identify the significant diabetes risk factors using statistical analysis for intensive lifestyle interventions, and develop machine learning models based on body data from wearable devices to predict whether an individual is at high risk for being diabetic.

Chapter 2

Project Objectives

The objectives of the project were outlined as follows:

1. To identify the significant risk factors or features that results in diabetes using statistical techniques.
2. To construct an effective binary classification model to predict whether an individual is at high risk for being diabetic.
3. To develop an easily interpretable and accessible IoT data product that can help non-technical individuals in assessing their diabetic risk.

Chapter 3

Data Interpretation (Statistical Analyses)

3.1 Review on Parametric and Non-Parametric Tests

In the context of statistical hypothesis testing, there are two broad categories employed, namely the *parametric* and *non-parametric* methods to draw inferences about populations based on sample data. The term 'parametric' refers to the parameters of the underlying statistical distribution. These two categories of tests differ fundamentally in their assumptions regarding the underlying distribution of the data, where it is important to ensure that the characteristics, applications, and considerations associated with the tests were well known before choosing the appropriate ones [21].

1. **Parametric Tests:**

   Parametric tests assume specific characteristics of the population distribution, such as normality and homogeneity of variance. Common examples include the Z-test, t-test and the Analysis of Variance (ANOVA). These tests usually have a higher statistical power when the assumptions are met, but can lead to unreliable conditions when these assumptions were violated.

   Advantages:

   - Greater statistical power under when assumptions were valid.
   - More precise parameter estimates.

   Disadvantages:

   - Sensitive to distributional assumptions.
   - Less robust in the presence of outliers.

2. **Non-Parametric Tests:**

   Non-parametric tests, on the other hand, make *minimal* assumptions about the population distribution. They are often preferred when dealing with ordinal or non-normally distributed data. Common non-parametric tests

include the Kruskal-Wallis test and the Mann-Whitney U test. These tests are more robust to deviations from normality and outliers, making them suitable for smaller sample sizes or non-normally distributed data.

Advantages:

- Robust to violations of distributional assumptions.
- Applicable to a wide range of data types.

Disadvantages:

- Less statistical power with larger sample sizes.
- Provide less precise estimates compared to parametric tests.

3.1.1 Proportion Test

A proportion test is a parametric test used to determine whether there is a significant difference in proportions between groups or to determine if a sample proportion is significantly different from a hypothesized population proportion. These tests are particularly valuable when dealing with categorical data, where observations can be classified into different categories or groups [11].

Conditions and Assumptions:

- Large sample size for normal approximation.
- Samples are randomly and independently selected from the populations of interest.

In our project, we are interested with the difference in proportions of diabetic patients in the categorical variables ('gender' and 'location'), which consist of two levels.

Test Procedure: For each categorical variable $j \in \{$'gender', 'location'$\}$, let '1' and '2' denote the levels of each j, and that $p_{i,j}$ denote the proportion of diabetic patients in level i and category j. We then perform the following:

1. Formulate the null and alternative hypotheses as below

$$H_0 : p_{1,j} - p_{2,j} = 0$$
$$H_1 : p_{1,j} - p_{2,j} \neq 0$$

2. Compute the sample proportions $\hat{p}_{1,j}$ and $\hat{p}_{2,j}$
3. Compute the test statistics Z with the formula below, where n denotes the sample size, x denote the number of successes (diabetic patients):

$$Z = \frac{(\hat{p}_{1,j} - \hat{p}_{2,j})}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

4. Compute the P-value with reference to the observed statistic and the standard normal distribution to make conclusions.

3.1.2 χ^2 -Test for Association

The χ^2 test for association is a non-parametric test to determine if there is a significant association between two categorical variables. It assesses whether the observed distribution of frequencies in a contingency table differs from the distribution that would be expected if the two variables were independent [11].

Conditions and Assumptions:

- Expected frequencies in each cell is at least 5.
- Random selection of data and mutually exclusive categories.

We are interested whether 'frame' is associated to diabetic status. Hence, a χ^2 test of association is appropriate.

Test Procedure:

1. Formulate the null and alternative hypotheses as below

$$H_0 : \text{There is no association between 'frame' and diabetic status}$$
$$H_1 : \text{There is an association between 'frame' and diabetic status}$$

2. Construct a contingency table, where rows represent one variable, columns represent the other variable, and each cell contains the count of observations in the corresponding combination.
3. Compute the expected frequencies in each cell using the formula below, and ensure that they are at least 5.

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

4. Compute the test statistics $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.
5. Compute the degrees of freedom, $df = (r-1)(c-1)$, where r represents the number of rows and c represents the number of columns.
6. Compute the P-value as $\mathbb{P}(\chi^2_{df} \geq \chi^2)$.

3.1.3 The Z -Test

The Z-test is a parametric test used to determine whether there is a significant difference between sample and population means or between the means of two independent samples [11].

Conditions and Assumptions:

- Population is normally distributed with known variance.
- Population is normally/non-normally distributed, with unknown population variance and a large sample size. This result is due to the Central Limit Theorem which ensures the normality of the sampling distribution for $\bar{X}$.

In our project, we are interested with the difference in central tendency of diabetic and non-diabetic. Hence, we perform a *two-sample* Z-test, where the data indeed satisfy the assumptions.

Test Procedure:

Denote '1' for the non-diabetic group and '2' for the diabetic group. For each continuous attribute j, we perform the following:

1. Formulate the null and alternative hypotheses as below

$$H_0 : \mu_{1,j} - \mu_{2,j} = 0$$
$$H_1 : \mu_{1,j} - \mu_{2,j} \neq 0$$

2. Compute the sample statistics such as $\bar{X_{1,j}}$, $\bar{X_{2,j}}$, $S^2_{1,j}$ and $S^2_{2,j}$, where $\bar{X}$ denote the sample mean and S^2 denotes the unbiased estimator of the population variance.
3. Compute the test statistics Z with the formula below, where n denotes the sample size:

$$Z = \frac{\bar{X}_{1,j} - \bar{X}_{2,j}}{\sqrt{\frac{S^2_{1,j}}{n_{1,j}} + \frac{S^2_{2,j}}{n_{2,j}}}}$$

4. Compute the P-value with reference to the observed statistic and the standard normal distribution to make conclusions.

3.1.4 The Mann–Whitney U Test

The Mann Whitney U Test is a non-parametric test to assess whether there is a significant difference between two independent and randomly sampled groups. The test compares the distribution of ranks between two groups to determine if one group tends to have higher values than the other [16].

Conditions and Assumptions:

- The sample drawn from the population is random.
- Independence within the samples and mutual independence.
- At least an ordinal measurement scale.

Instead of testing on the 'mean', the Mann-Whitney U Test determines whether there is a significant difference in the distribution.

Test Procedure:

For each continuous attribute j, we perform the following:

1. Formulate the null and alternative hypotheses as below

   H_0 : There is no difference in distribution between diabetic and non-diabetic patients for attribute j.

   H_1 : There is a difference in distribution between diabetic and non-diabetic patients for attribute j.

2. Combine data from both groups. Rank the data from smallest to largest without considering the group they belong to. In case of ties, assign the average rank to tied observations.
3. Calculate the sum of ranks, R_1 and R_2 for each group separately, and choose the test statistics as $U = \min\{U_1, U_2\}$, where $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ and $U_1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$.
4. Compute the P -value with reference to the observed statistic and the Mann-Whitney table to make conclusions. For larger sample sizes, a normal approximation can be used.

3.2 Review on Effect Sizes

Even though the statistical tests in Chapter 3 might detect a significant difference in the means, distributions and proportions, they do not quantify the magnitude of an observed effect or relationship in a study [6]. Hence, effect sizes are particularly important in addition to p-values because they offer a more comprehensive understanding of the results. Some common effect size measures include η^2 in ANOVA and R^2 in regression. In our context, we will focus on Cohen's d, Cohen's h and Cramér's V.

3.2.1 Cohen's d

Cohen's d [6] is defined as the difference between two means divided by a standard deviation for the data:

$$\boxed{d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}, \quad \text{where } s_p \text{ denotes the pooled standard deviation.}}$$

According to [6], Table 3.1 contains descriptors for magnitudes of d. A negative effect size indicates that the mean difference between two groups is in the opposite direction of the comparison.

Table 3.1: Interpretation of Magnitude of Cohen's d

Effect size	d
Small	0.20
Medium	0.50
Large	0.80

3.2.2 Cohen's h

Cohen's h [6] is a measure of distance between two proportions or probabilities:

$$\boxed{h = 2\arcsin(p_1) - 2\arcsin(p_2)}$$

According to [6], Table 3.2 contains descriptors for magnitudes of h. A negative effect size indicates that the proportion difference between two groups is in the opposite direction of the comparison.

Table 3.2: Interpretation of Magnitude of Cohen's h

Effect size	h
Small	0.20
Medium	0.50
Large	0.80

3.2.3 Cramér's V

Cramér's V [6] is a measure of association between two categorical variables. Cramér's V varies from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when each variable is completely determined by the other.

Cramér's V is computed by taking the square root of the χ^2 statistic divided by the sample size and the minimum dimension minus 1, where the notations used are identical to the previous sections:

$$\boxed{V = \sqrt{\frac{\frac{\chi^2}{n}}{\min\{r-1, c-1\}}}}$$

According to [6], Table 3.3 contains descriptors for magnitudes of V.

Table 3.3: Interpretation of Magnitude of Cramér's V

Effect size	V
Small	0.10
Medium	0.30
Large	0.50

3.3 Statistical Results on the Dataset

As mentioned in the previous project, there is a plausible outlier in our dataset with notably high 'stab.glu' level. Upon further investigation, we decided to include this observation. The reasoning behind is that even though a high 'stab.glu' level for a reasonable 'time.ppn' could indicate diabetes, but our subsequent classification problems are based on 'A1c' values.

3.3.1 Statistical Hypothesis Testing

We conducted various tests on proportions, means, distributions and association between diabetic and non-diabetic patients. All tests were conducted at a 5% level of significance. The summary of results can be found in Appendix A.

On categorical variables with two levels, we conclude that the proportion of diabetic patients stratified (both independently) by 'gender' and 'location' do not significantly differ.

On categorical variables with three levels ('frame') and two levels ('diabetic'), we have significant evidence to conclude that there is an association between diabetic patients and 'frame'.

Moving on to continuous variables, the means and distributions of all variables stratified by diabetic status are significantly different as suggested by both the Z-test and Mann-Whitney U test, except for 'bp.1d', 'time.ppn' and 'height'.

From [3], it is noted that body frame index is a stronger predictor for diabetes as compared to body mass index (BMI). As such, we do not merge attributes, such as combining height and weight to obtain BMI.

To conclude, we observe that stabilised glucose, age, the two cholesterol types and their ratio, body measurements and frame type, first systolic blood pressure are significant diabetes risk factors.

3.3.2 Effect Sizes

Of all the non significant results in Section 3.3.1, we observe that their effect sizes are small, which suggest that they might not have much practical significance.

For 'frame', an effect size of 0.161 suggests that 'frame' has a small to medium effect on diabetic status.

Lastly, all significant continuous variables have at least a small to medium effct size, with 'stab.glu' having the highest value as expected. It is worth noting the negative effect size for 'hdl' indicates that diabetic patients tend to have a lower 'hdl' (good cholesterol) level.

For subsequent Machine Learning training, we will be using significant features as suggested by the statistical tests their effect sizes.

Chapter 4

Modelling with Machine Learning

4.1 Popular Machine Learning Techniques

Here are some brief introduction of some Machine Learning models that we will consider in classifying individuals into diabetic and non-diabetic class.

1. **Naïve Bayes**

   The Naïve Bayes classifier [14] falls under the category of Bayesian learning that emphasizes on the Bayes' Theorem. The algorithm calculates the probability of each feature and the probability of each class based on the training data. Given a new set of features, Naïve Bayes calculates the conditional probability of each class and selects the class with the highest probability as the prediction.

   Despite its simplicity, it is computationally efficient and its performance has been shown to be comparable to that of neural network in some domains.

2. **Support Vector Machine**

   Support Vector Machine (SVM) [1] is a set of powerful supervised learning techniques used for classification and prediction problems.

   The basic concept of SVM for a classifier is to construct a maximum-margin separating hyperplane in some transformed feature space. Instead of having to specify the exact transformation, SVM uses the principle of kernel substitution (kernel trick) to turn them into a general non-linear model.

3. **Logistic Regression**

   (Binary) Logistic Regression [19] analyzes the relationship between multiple exploratory variables and a (two-level) categorical response variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve.

   With the estimated probability, one can decide a decision boundary (typically 0.5). If the predicted probability is above the threshold, the instance is classified as the positive class and vice versa.

4. **Neural Networks**

   Neural Networks [1] are mathematical representations that attempt to mimic the functionality of a human brain. The benefits of Neural Network include the ability and flexibility to model most non-linear association between input variables and target variables. Several architectures of Neural Networks have been proposed, but we will be focusing on Multilayer Perceptron (MLP).

   MLP consists of an input layer (neurons for all input variables), hidden layers (any number of neurons) and an output layer. Each neuron processes its inputs and emit its output to neurons in the subsequent layer. Each connection between neurons is assigned a weight during training. The output of hidden neuron is computed by applying an activation function like the logistic and hyperbolic tangent function to the weighted inputs and its bias term.

All 4 models will be trained with our dataset, and the best model will be selected for deployment based on certain evaluation metrics.

4.2 Evaluation Metrics

We outline several performance metrics that will be used to gauge the performance of our machine learning model. These metrics were based on the *confusion matrix*.

Confusion matrix is a popular metric used for classification problems. It is used for binary classifications, and can be extended to multi-class classifications. Table 4.1 shows the structure of a confusion matrix for a binary classification problem, and it can be used to derive four metrics with different emphasis [12].

Table 4.1: A Confusion Matrix

		Predicted Condition	
		Negative	Positive
Actual Condition	Negative	True Negative **(TN)**	False Positive **(FP)**
	Positive	False Negative **(FN)**	True Positive **(TP)**

1. **Accuracy**

   One of the most popular metrics for classification is accuracy, which is calculated using the formula below. However, it can be misleading when dealing with imbalanced dataset, known as the *accuracy paradox* [22].

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

2. **Precision**

   Precision shows how accurate the model is for predicting positive values. Hence, it measures the accuracy of a predicted positive outcome. The formula for precision is given below.

$$Precision = \frac{TP}{FP + TP}$$

3. **Recall**

   Recall on the other hand measures the strength of a model to predict positive outcomes. It is defined as:

$$Recall = \frac{TP}{FN + TP}$$

4. **F1-score**

   The F1-score is a special case of the F measure, which is defined by the weighted harmonic mean between precision and recall. For the classification of positive instances, it helps to understand the trade-off between correctness and coverage.

$$F_\beta = (1 + \beta^2)\frac{precision * recall}{(\beta^2 * precision) + recall}$$

   The term β can be varied to put emphasis in either precision or recall. Commonly, β is taken to be 1, which give rise to the F1-score.

The selection of appropriate metrics depends on (but not limited to) the classification problem, domain of interest and the class balance of the dataset.

4.3 Dealing with Imbalanced Dataset

Often in real life situations, data collected from various sources such as the internet, websites and databases suffer data quality issues . These issues introduce biases in various stages of the data science lifecycle. Aside from common issues such as missing and inconsistent entries, data imbalance is also one of the issues, mainly encountered for classification problems [12].

Data imbalance occurs when a dataset has unequal class distribution. In such scenarios, the class which has majority instances is considered as a majority class or a negative class, and the underrepresented class is viewed as a minority class or a positive class.

As an example, we can consider the classification model which predicts fraudulent transactions based on a dataset. In reality, almost all transactions are valid, with a very small proportion of fraudulent transactions. The classification model Will then treat fraudulent transactions as non-fraudulent transactions because of the unequal proportion of class distribution in the data. Such models usually yield high accuracy, but bad performance in predicting fraud.

Hence, data imbalance should be dealt, and this can be achieved using data level methods. Common techniques include oversampling and undersampling [12]. Oversampling aims to increase the count of minority class instances to match it with the count of majority class instances. Undersampling works the other way round, as it downsizes the the majority class instances. to match with the minority class instances.

In our case, the minority class will be the diabetic instances, and we will be oversample it to match with the non-diabetic instances. There are several oversampling techniques, such as the Random Oversampling [12], Synthetic Minority Oversampling Technique (SMOTE) [5] and Adaptive Synthetic Sampling (ADASYN) [10]. We will be utilising the Random Oversampling on the training data, where instances from the minority class are selected at random with replacement which results in a balanced class distribution for our problem.

4.4 Model Fitting and Evaluation

Before we perform Machine Learning modelling, we first did some modifications on the dataset. This includes the removal of non-significant attributes, modifying strings to integers for categorical attributes such as ‘frame’, and randomly split the data into training and testing set with a proportion of 0.8 and 0.2 respectively. Lastly, all models were trained at a random state of 333, and comparison will be made using evaluation metrics.

Specifically, the we utilize ‘liblinear’ as our optimisation solver for Logistic Regression. Also, our neural network consists of 2 hidden layers, each with 128 and 64 neurons respectively, ‘adam’ as our optimisation solver, $\tanh$ as our activation function and a maximum iteration of 10000.

In the sense of medical classification, we are more concerned of false negative instead of false positive. The reason is that a false negative (diabetic patient, but classified as a non-diabetic patient) could delay treatment, leading to more serious complications. Hence, we will be emphasizing on recall while comparing models.

4.4.1 Without Oversampling

We first train the Machine Learning models without oversampling the minority class instances.

From Table B.1 , based on the four evaluation metrics, Naïve Bayes performs the best in terms of accuracy, recall and F1 score, while support vector machine has the highest precision. Neural network is the worst model as compared to the rest.

We can observe relatively high accuracy for all four models. However, there is a sharp drop in recall, especially for support vector machine, logistic regression and neural network. This suggests that all four models are biased toward the majority class.

4.4.2 With Oversampling

From Table B.2 ,after oversampling the minority class, Naïve Bayes still performs the best in terms of accuracy, recall and F1 score, while neural network performs the best in terms of precision.

As expected, oversampling the minority class instances did improve the recall rate to a great extent, at the cost of a slight reduction in accuracy and precision and accuracy.

To conclude, Naïve Bayes is the best model among our proposed models, and will be used for deployment.

Chapter 5

Plan for Reproducible Research

To achieve the goals of reproducible research, which emphasizes transparency and openness in the research process, software and methods availability and promote the 'Transfer of Knowledge', we ensured that our project was easily comprehensible and interpretable from the start of the project.

The detailed steps and information in our research process are compiled in the format of jupyter notebook as they provide an effective platform for conducting and sharing reproducible research. The notebook can be accessed from our repository on GitHub, `https://github.com/justin-sem/GNBM_diabetes`, which is an open-source software development and hosting platform which ensures that our research process can be replicable without much effort and that the study findings are verifiable. In addition, the trained Naïve Bayes model (.sav) is also available on the repository to increase the degrees of reproducibility.

We hope that our planning not only facilitates reproducibility and enables discoverability, but also serves as a catalyst for further research endeavours, particularly in the area of diabetes detection. Through our commitment to openness and clarity, we aspire to foster a collaborative environment where insights gained from reproducible research can drive meaningful progress in addressing the challenges posed by diabetes and related health concerns.

Chapter 6

Deployment

Though our original data product is to deploy on IoT devices like smartwatches that are capable on detecting the required variables [4, 18] without relying on lab based 'A1c' values. However, due to the limitations in facilities and project scope, we do not accomplish this at the moment.

Alternatively, we deployed our selected machine learning model on Streamlit, which is accessible via the link: `https://gnbmdiabetesdeploy-pnniqp24kyakur6c7f7hz5.streamlit.app/`. By filling in the required values, the model will classify the diabetic status of an individual.

However, one must be aware of the limitations of such classification model, and it is not intended to replace lab-based medical tests nor formal diagnosis by medical professionals.

Chapter 7

Summary

7.1 Insights

The statistical analysis identified stabilised glucose levels, cholesterol types and their ratios, age, and body measurements and first systolic blood pressure as key diabetes risk factors. This aligns with existing medical knowledge and validates their inclusion in predictive models.

One of the key insights from this research is that the developed Naive Bayes model could help healthcare professionals and patients quickly assess the risk of diabetes based on basic health measurements that can be easily collected during a clinical visit or even using wearable devices. Patients could also utilize the model themselves by inputting their measurements periodically to monitor their risk levels over time. This would allow them to take preventive actions if their risk starts becoming elevated. Thus, the model can serve as a convenient first-line screening tool for this chronic condition.

Oversampling the minority class is another crucial strategy in this study that is highlighted to address the imbalance in the datasets between patients with and without diabetes. By oversampling, the models were protected against developing a bias in favor of the majority class, which would have reduced their sensitivity. Recall was increased to 92.3% thanks to the better balance, which did not materially compromise specificity or accuracy. This highlights how crucial it is to address class disparities in machine learning systems in order to prevent prejudice towards disadvantaged groups

7.2 Limitations

Several limitations remain which point to opportunities for further research and improvement. The limitations can be broadly categorized into three areas:

1. **Data-related limitations**

   The dataset used in this research was relatively small and restricted to patients from only two counties in Virginia, USA. This restricts the created model's capacity to generalize to bigger and more varied populations. Extending the data breadth by include patients from various geographic locations, age groups, ethnicities, and so on would increase model robustness and external validity.

   Furthermore, the number of input variables was limited. Other key risk variables, including as nutrition, exercise levels, family history, and so on, might assist improve the model's predictive ability. A broader set of inputs would allow for the capture of non-linear correlations and interactions between variables.

2. **Method-related limitations**

   The interpretation of the developed machine learning model was limited to basic measures like feature importance. More advanced model interpretation approaches, such as SHAP [20], might aid in better explaining the rationale behind the predictions to end users.

   The model was evaluated using a small held-out test set that was divided 80-20 from the original dataset. More stringent validation on bigger test sets utilizing k-fold cross validation would assist assure stable performance across different samples. The hyperparameter tuning experiments for the models studied in this article were likewise restricted. Further optimization of model hyperparameters could potentially help boost the performance even more.

3. **Generalization-related limitations**

   No information was available in the dataset on the type of diabetes i.e. Type 1 or Type 2. Developing separate models customized for each population could help improve the specificity of predictions.

7.3 Conclusion

This project contributed to the creation of an automated data-driven tool for diabetes prediction, intended for applications on IoT devices that are accessible to all. Furthermore, we demonstrated the practicality and performance of machine learning in preventive healthcare. Important diabetes risk factors were identified by statistical tests and analyses, and class imbalance was rectified by oversampling for machine learning modeling. An effective Naïve Bayes prediction model was created and made available online.

However, limitations around data constraints, model validation, and interpretation draw attention to the need for more study. This proof-of-concept model will be developed further, enhanced methods, and extensive testing will contribute to its development into a therapeutically valuable decision assistance tool. The knowledge gained from this research adds to the expanding possibilities of AI in customized medicine and predictive analytics. No details regarding the types of diabetes (Type 1 vs Type 2). Separate models for each type could improve specificity.

Appendix A

Statistical Results

A.1 Hypotheses Testing and Effect Sizes

The following tests were conducted at a significance level of $\alpha = 5\%$. Statistically significant results were highlighted.

Table A.1: Diabetic Proportion Test on 'gender' and 'location' with Cohen's h

Categorical Variable	***P-Value***	***Cohen's*** h
gender	0.581	-0.056
location	0.717	0.038

Table A.2: χ^2 Test of Association between 'frame' and Diabetic Status with Cramer's V

Categorical Variable	***P-Value***	***Cramer's*** V
frame	0.006	0.161

Table A.3: Z-Test and Mann-Whitney U Test on Continuous Variables with Cohen's d

Continuous Variable	Z***-Test*** P***-Value***	***Mann-Whitney*** U***-Test*** P***-Value***	***Cohen's*** d
stab.glu	0.000	0.000	1.699
age	0.000	0.000	0.965
ratio	0.000	0.000	0.647
waist	0.000	0.000	0.624
bp.1s	0.000	0.000	0.619
chol	0.003	0.002	0.528
weight	0.000	0.000	0.402
hip	0.003	0.002	0.390
bp.1d	0.322	0.350	0.134
time.ppn	0.471	0.433	0.102
height	0.732	0.720	0.046
hdl	0.020	0.001	-0.326

Appendix B

Machine Learning

B.1 Model Performance without Oversampling

B.1.1 Confusion Matrices

- **Naïve Bayes**

$$\begin{bmatrix} 61 & 4 \\ 2 & 11 \end{bmatrix}$$

- **Support Vector Machine**

$$\begin{bmatrix} 63 & 2 \\ 5 & 8 \end{bmatrix}$$

- **Logistic Regression**

$$\begin{bmatrix} 63 & 2 \\ 6 & 7 \end{bmatrix}$$

- **Neural Networks**

$$\begin{bmatrix} 62 & 3 \\ 6 & 7 \end{bmatrix}$$

B.1.2 Evaluation Metrics

Table B.1: Results Summary Without Oversampling

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Gaussian Naive Bayes	92.31	73.33	84.62	78.57
Support Vector Machine	91.03	80.00	61.54	69.57
Logistic Regression	89.74	77.78	53.85	63.64
Neural Network	88.46	70.00	53.85	60.87

B.2 Model Performance with Oversampling

B.2.1 Confusion Matrices

- **Naïve Bayes**

$$\begin{bmatrix} 60 & 5 \\ 1 & 12 \end{bmatrix}$$

- **Support Vector Machine**

$$\begin{bmatrix} 59 & 6 \\ 2 & 11 \end{bmatrix}$$

- **Logistic Regression**

$$\begin{bmatrix} 59 & 6 \\ 3 & 10 \end{bmatrix}$$

- **Neural Networks**

$$\begin{bmatrix} 62 & 3 \\ 4 & 9 \end{bmatrix}$$

B.2.2 Evaluation Metrics

Table B.2: Results Summary With Oversampling

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Gaussian Naive Bayes	92.31	70.59	92.31	80.00
Support Vector Machine	89.74	64.71	84.62	73.33
Logistic Regression	88.46	62.50	76.92	68.97
Neural Network	91.03	75.00	69.23	72.00

Bibliography

[1] Abraham Iorkaa, A., Barma, M., and Muazu, H. (2021). Machine Learning Techniques, methods and Algorithms: Conceptual and Practical Insights. *International Journal of Engineering Research and Applications*, 11:55–64.

[2] American Diabetes Association (2018). 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2018*. *Diabetes Care*, 41(Supplement 1):S13–S27.

[3] Bawadi, H., Abouwatfa, M., Alsaeed, S., Kerkadi, A., and Shi, Z. (2019). Body Shape Index Is a Stronger Predictor of Diabetes. *Nutrients*, 11(5):1018.

[4] Chang, T., Li, H., Zhang, N., Jiang, X., Yu, X., Yang, Q., Jin, Z., and Meng, H. (2022). Highly integrated watch for noninvasive continual glucose monitoring. *Microsystems & Nanoengineering*, 8.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[6] Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

[7] Deshpande, A. D., Harris-Hayes, M., and Schootman, M. (2008). Epidemiology of Diabetes and Diabetes-Related Complications. *Phys. Ther.*, 88(11):1254–1264.

[8] Foley, K., McNaughton, D., and Ward, P. (2020). Monitoring the 'diabetes epidemic': A framing analysis of United Kingdom print news 1993-2013. *PLoS One*, 15(1):e0225794.

[9] Galaviz, K. I., Narayan, K. M. V., Lobelo, F., and Weber, M. B. (2018). Lifestyle and the Prevention of Type 2 Diabetes: A Status Report. *Am. J. Lifestyle Med.*, 12(1):4–20.

[10] He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

[11] Kanji, G. (1999). *100 statistical tests: new edition*. Sage Publications Limited.

[12] Kulkarni, A., Chong, D., and Batarseh, F. A. (2020). 5 - foundations of data imbalance and solutions for a data democracy. In Batarseh, F. A. and Yang, R., editors, *Data Democracy*, pages 83–106. Academic Press.

[13] Lindström, J. and Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26(3):725–731.

[14] Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

[15] Mujumdar, A. and Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165:292–299. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[16] Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, 4.

[17] Naidoo, S. (2017). Prevention is better than cure! *Southern African Journal of Infectious Diseases*, 32:1.

[18] Ni, J., Hong, H., Zhang, Y., Tang, S., Han, Y., Fang, Z., Zhang, Y., Zhou, N., Wang, Q., Liu, Y., Li, Z., Wang, Y., and Dong, M. (2021). Development of a non-invasive method for skin cholesterol detection: pre-clinical assessment in atherosclerosis screening. *Biomed. Eng. Online*, 20(1):52.

[19] Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *J. Korean Acad. Nurs.*, 43(2):154–164.

[20] Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G., and Facchinetti, A. (2023). The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using shap. *Scientific Reports*, 13(1).

[21] Uchechi, H. and Akwiwu, E. (2019). Choice of Parametric and Nonparametric Statistical Procedures in Clinical and Biomedical Research. *Sokoto Journal of Medical Laboratory Science*, 4:5–15.

[22] Valverde-Albacete, F. J. and Peláez-Moreno, C. (2014). 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. *PloS one*, 9:e84217.