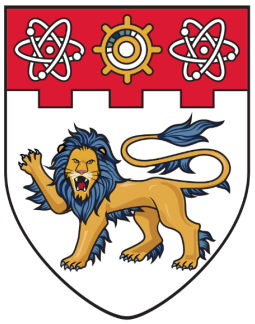


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

AY2025/26 SEMESTER 1

BC2406 ANALYTICS I: VISUAL & PREDICTIVE TECHNIQUES

Title: Final Project

Seminar: 5

Group: 3

Date of Submission: 1 November 2025

Full Name	Matriculation Number
ANASTASIA KOESOEMO	U2440944A
LIM LI XUAN VALENCIA	U2340445L
NGIAM KAI HER DARRYN	U2340038J
RAE NG	U2440617L
STEPHEN MICHAEL LEE	U2410465B

Table of Contents

Executive Summary

The Healthier SG initiative marked a shift in Singapore's healthcare philosophy, which moved the focus from reactive treatment to preventive care. This was aimed at delaying the onset of chronic diseases like Type 2 Diabetes. However, current preventive efforts are mostly "one size fits all", lacking the personalisation that is needed to effectively target the most at- risk individuals who require additional care, or simply the average patient who has pre-diabetes. This project hence focuses on a unique solution, DiaScope, to demonstrate and prove that predictive analytics can be integrated into the Healthier SG framework to make preventive care more efficient, personalised and effective.

In our approach, we analysed two datasets named "Early Symptoms" and "Health Indicators" using Logistic Regression and CART. These two models were chosen to reliably predict an individual's diabetes risk based on self reported symptoms as well as from their clinical results and routine checkups. Our analysis found that both models are effective and distinct, providing a unique edge. The CART model is highly accurate in predicting symptom-based data due to its unique splitting model, making it suitable for a public-facing self-assessment too. The Logistic Regression model on the other hand performed better dealing with complex clinical data, making it extremely competent in applications such as dashboards.

Therefore, in our solution DiaScope, we implemented a two part solution that creates an integrated ecosystem, allowing residents and their primary care doctors to be connected.

In DiaScope, individuals will have a smart health companion, in which DiaScope will transform the existing Healthy365 app into a proactive, personalised health coach. DiaScope will use the predictive models generated to send personalised, predictive-risk alerts to users and further provides targeted engagement. This includes personalised suggestions like a community cooking workshop for individuals who like to cook. Primary Care supporters like family doctors will get a comprehensive and smart dashboard, known as a Predictive Clinical Dashboard. This simple tool provides them with an overview of their patients ranked in Highest priority to Lowest priority, allowing care teams to focus their limited resources on the highest risk individuals. The dashboard also syncs with the patient's app, allowing continuous monitoring instead of relying solely on their health check ups.

Therefore, by introducing a personalised engagement tool that is built on accurate and robust analytical models, this will allow Singapore to align closer to the health goals it envisioned. The model should next be validated and trained using local data, making it more robust than it currently is.

1. Introduction

1.1 Background

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disease that occurs when the body is unable to use insulin effectively, resulting in high blood glucose levels over time. If left uncontrolled, T2DM can lead to serious complications such as cardiovascular disease, kidney failure, and blindness (World Health Organization, 2024).

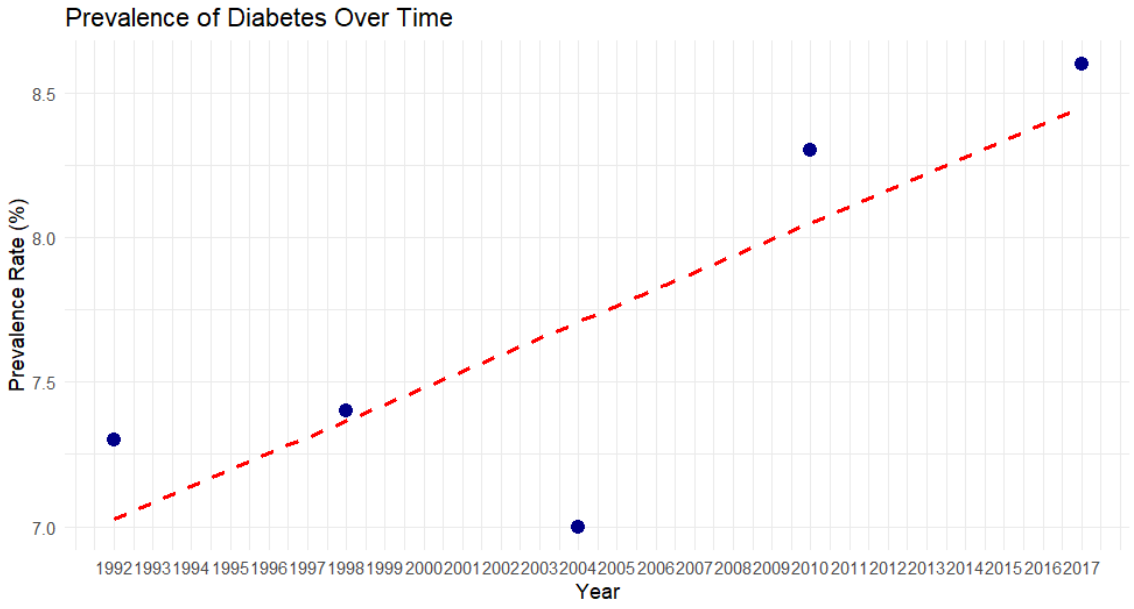


Figure 1: Prevalence of Diabetes among Singapore residents aged 18 to 69 years, 1992 – 2017. (Ministry of Health, 2024)

The prevalence of T2DM in Singapore has been increasing over time. By 2017, about 14% of Singaporeans, aged 18 to 69 years, were diagnosed with pre-diabetes (Ministry of Health, 2017) and 8.6% of Singaporeans diagnosed with T2DM (Ministry of Health, 2024). This trend highlights the urgency of preventive measures for T2DM.

In 2023, the Ministry of Health (MOH) launched the Healthier SG initiative to shift the healthcare system from solely treatment focused to preventive care. The programme aims to encourage residents to enrol with family doctors, adopt personalised health plans, and embrace healthier lifestyles through digital platforms like HealthHub and Healthy365. This initiative supports Singapore's long-term goal of reducing the burden of chronic diseases by promoting preventive care.

1.2 Problem Statement

Despite nationwide screening and health promotion programmes, the prevalence of T2DM remains high in Singapore. Current preventive efforts are one-sized-fits-all and not largely personalised. This makes it difficult to target individuals who are most at risk. Without targeted risk prediction, resources may be allocated inefficiently and high-risk individuals may not receive timely support.

Through this project, we aim to leverage on predictive modelling techniques such as regression and CART to innovate on further solutions. Thereby enhancing prevention, enabling doctor engagement and supporting Healthier SG's goal towards proactive community based care.

1.3 Current Situation

Singapore's 3 healthcare clusters, SingHealth, National Healthcare Group and National University Health System, are currently implementing Healthier SG mainly through primary-care networks and community programmes. Residents are encouraged to undergo regular screenings for blood pressure, cholesterol and glucose. Meanwhile public campaigns are promoting exercise and healthy eating. However, these initiatives still rely on population averages instead of predictive models that use individual data to assess risk. By leveraging large-scale health datasets, we can build predictive tools to identify at-risk groups and support targeted preventive action under Healthier SG.

1.4 Benefits of Preventive Health

Diabetes risk factors can be classified into three main categories (World Health Organization, 2024):

- Lifestyle factors: Physical inactivity, unhealthy diets and smoking
- Clinical factors: Obesity, high blood pressure and high cholesterol
- Demographic factors: Older age and gender

Through early identification of such factors using predictive analytics, interventions can be implemented to reduce the likelihood of developing diabetes. Studies have shown that structured lifestyle interventions can reduce the risk of T2DM by 30 - 40% over 5 years (WHO, 2023). By applying data-driven approaches within the Healthier SG framework,

Singapore can achieve more efficient screening, better resource allocation and improved population health outcomes in the future.

2. Methodology

2.1 Datasets

We found 2 datasets to model early detection and prevention strategies. They are, Early Symptoms and Health Indicators, which will be introduced in this section.

2.1.1 Early Symptoms

The Early Symptoms dataset has 520 observations and 17 variables on individuals' early warning signs that may indicate potential onset of diabetes (Andrewmvd, 2021). This dataset is made through collating questionnaires and diagnosis results from the patients in Sylhet Diabetes Hospital in Sylhet, Bangladesh. It includes qualitative and behavioural variables such as frequent thirst, frequent urination, weight loss etc (Appendix A Table 1).

This dataset is used to identify early behavioural and symptomatic predictors of diabetes before medical diagnosis.

2.1.2 Health Indicators

The Health Indicators dataset has 253,680 observations and 22 variables on measurable wellness and fitness related metrics (Alex, 2021). This dataset is made through a yearly survey, Behavioural Risk Factor Surveillance System (BRFSS) by Centres for Disease Control and Prevention (CDC). BRFSS collects information from Americans on topics such as health risk behaviours, chronic conditions and the use of preventive health services. It includes key features such as blood pressure, blood cholesterol, physical activity frequency etc and demographic variables such as gender, age, education and income (Appendix A Table 2).

This dataset captures modifiable risk factors and life determinants that contribute to onset diabetes.

2.2 Integration Across Datasets

The integration of these 2 datasets provides a holistic view of diabetes risk, combining behavioural, lifestyle and clinical dimensions. By linking early symptoms and lifestyle indicators with clinical outcomes, the model can stimulate how HealthierSG's multilevel preventive strategies, from individual selfcare to doctor led interventions can be enhanced through predictive and clustering analytics.

3. Data Processing and Exploration (Early Symptoms)

The Data transformation for this dataset before model training was relatively straightforward.

3.1 Missing Data

No missing values were detected in any column.

3.2 Duplicate Records

269 duplicate rows were identified and removed to avoid model bias. This reduced the dataset count to 251 unique records.

3.3 Class Distribution

The dataset is moderately imbalanced but not too extreme. It is acceptable for binary classification without resampling.

Class	Count	Proportion (%)
1 (Diabetic)	173	68.9
0 (Non-diabetic)	78	31.1

3.4 Outlier Treatment

No outliers were found in this dataset after analysing the variable distribution. The plots confirmed that all binary features contained valid (0/1) entries while age values were within expected adult ranges (20 - 80 years) (Appendix B).

3.5 Categorical Variables Conversion

All predictor variables except for age were converted to factors. This ensures correct handling in logistic regression and CART models which depend on classification algorithms.

4. Data Processing and Exploration (Health Indicators)

The data transformation process for this dataset was more extensive due to its larger size and containing mixed variable types (binary, ordinal and continuous). Several cleaning and standardisation steps were performed before model training.

4.1 Missing Data

No missing values were detected in any column.

4.2 Duplicate Records

A total of 23,899 duplicate rows were identified and removed to avoid data redundancy and over representation of certain records. After duplicate removal, the dataset size was reduced from 253,680 rows to 225,152 unique records.

4.3 Class Distribution

The target Y variable, Diabetes_012, originally contained 3 categories:

- 0 = Non-diabetic
- 1 = Prediabetic
- 2 = Diabetic

For binary classification, the prediabetic cases were removed and the diabetic cases were re-coded as 1 (Appendix B).

The resulting class distribution is shown below:

Class	Count	Proportion (%)
1 (Diabetic)	190,055	84.4
0 (Non-diabetic)	35,097	15.6

The dataset is quite imbalanced, which will later be addressed using random sampling before model training.

4.4 Outlier Treatment

From the variable distribution plots in Appendix B, the BMI column contained several extreme values. For example, the maximum was 98. These outliers were handled using the IQR method (Winsorisation). These values were obtained by calculating the upper and lower limit using Q1, Q3 and IQR (Appendix B).

	Minimum	Maximum
Before	12.0	98.0
After	13.5	41.5

4.5 Categorical Variables Conversion

All categorical variables except continuous and ordinal ones were converted to factors to ensure proper treatment in classification algorithms.

4.6 Data Balancing

To address the class imbalance, a balanced subset of 2,000 observations (1,000 diabetic + 1,000 non-diabetic) were drawn randomly for model training and testing.

Subset	Records	Class Split
Trainset	1400	700 : 700
Testset	600	300 : 300

5. Modelling and Performance

5.1 Train-Test Split

The Train-Test Split method separates the dataset into training and testing subsets. The model learns from the training data and is tested on new instances to ensure accuracy in making predictions. This technique prevents overfitting and evaluates the model's real-world

applicability. This is also known as supervised learning. For this project, we have used a 70-30 split: 70% for training and 30% for testing.

5.2 Logistic Regression

Logistic regression is a supervised classification algorithm used to predict the probability of a binary outcome based on one or more predictor variables. It models the log odds of an event occurring and applies a logistic function to map predicted values between 0 and 1. This model is chosen in particular because the response variable is categorical whereby individuals are labelled either as diabetic or non diabetic. We applied this model on R using the glm( ) function.

5.2.1 Selection of Variables to Model

Variable selection ensures that only significant predictors are retained in the model, improving accuracy and reducing overfitting.

We performed variable selection by using the Stepwise Algorithm which implements model selection based on Akaike Information Criterion (AIC). AIC balances model fit and complexity as it penalises unnecessary predictors that do not improve predictive power. By using step( ) on R, the algorithm iteratively removes variables, selecting the combination which minimises AIC value. Variables retained in the final model are those that contribute significantly to predicting diabetes outcome. The final model after enhancement using step Stepwise Algorithm is shown in Appendix C Table 1.

Additionally, a key assumption for logistic regression is no multicollinearity which means that the independent variables should not be highly correlated to one another. High collinearity can distort the estimated coefficients, inflate standard errors and make it difficult to determine the individual effect of each predictor on the outcome variable. To assess this, the Variance Inflation Factor (VIF) was calculated for all retained predictors using vif( ) on R. In this model, all predictors had VIF value below 5 suggesting that multicollinearity is not a concern (Appendix C Table 2). Hence, we can safely interpret the model's coefficients as independent effects.

5.2.2 Training and Testing the Model

The train dataset was used to fit the logistic regression model while the test dataset was used to validate its predictive ability. The model output produces a predicted probability for each

individual. Predictions were classified into diabetic or non diabetic using a probability threshold of 0.5 whereby values above 0.5 were predicted as diabetic cases.

5.2.3 Model Evaluation

The table below presents the model development process and goodness-of-fit statistics for logistic regression applied to both datasets.

Dataset	Initial Predictors	Final Predictors	Initial AIC	Final AIC	Null Deviance	Residual Deviance
Early Symptoms	16	7 Significant	102.63	90.75	218.62 (df=175)	72.75 (df=167)
Health Indicators	21	7 Significant	1538.0	1516.4	1940.8 (df=1399)	1500.4 (df=1392)

Refer to Appendix C for significant predictors' Odds Ratio (OR) and Confidence Interval (CI) distributions.

Rank	Early Symptoms	OR	Health Indicators	OR
1	polydipsia	68.48	CholCheck	5.16
2	polyuria	33.51	HighBP	2.15
3	genital_thrush	11.67	HighChol	1.98
4	sudden_weight_loss	5.92	GenHlth	1.58
5	partial_paresis	4.81	AgeBrac	1.19
6	gender	0.11	BMI	1.10
7	itching	0.06	HvyAlcoholConsump	0.28

Rank is based on effect on outcome. The stronger the effect the higher the rank.

5.3 CART

The Classification and Regression Tree (CART) is a supervised machine learning algorithm that builds decision trees for classification and regression problems. Since we are predicting for a binary outcome, we will be using a classification tree model. This will be done using the 'class' method in the rpart( ) function.

5.3.1 Model Building

In building the CART model, we first grew out the tree to reach the maximum number of terminal nodes. We then used pruning techniques to optimise the model's statistical accuracy while minimising the complexity. After pruning, the optimal tree will be derived.

5.3.2 Pruning

We utilised the 1 Standard Error (SE) rule guideline in choosing the simplest tree that would be statistically equivalent with the minimum Cross-Validation (CV) error, based on a 10-fold cross validation. This approach aims to balance model simplicity with predictive accuracy, preventing overfitting and over complex trees.

5.3.3 Model Evaluation

The table below shows that the Early Symptoms tree is simpler and more compact This indicates strong and direct predictors that classify diabetes with minimal depth. In contrast, the Health Indicators tree is deeper and more complex. This suggests that the predictors are more indirect and require finer pruning. Overall, the symptoms-based model offers higher interpretability, while the indicators-based model captures more nuanced health risk patterns.

Dataset	Splits	Terminal Nodes	CP
Early Symptoms	5	6	0.045
Health Indicators	9	10	0.006

The table below displays the top 5 variable importance scores for both CART models.

Rank	Early Symptoms	Importance Score	Health Indicators	Importance Score
1	Polyuria	25.92	HighBP	91.15
2	Polydipsia	22.07	GenHlth	49.92
3	Sudden weight loss	9.25	AgeBrac	37.71
4	Partial paresis	9.04	BMI	31.44
5	Polyphagia	8.14	HighChol	14.78

Refer to Appendix C for the complex parameter (CP) plots, optimal tree illustrations and full variable importance distributions.

5.4 Model Performance Comparison

To evaluate predictive accuracy, we generated a confusion matrix to compare predicted and actual outcomes in the test dataset. The confusion matrix summarises the results in four categories:

- True Positives (TP): Correctly predicted diabetic cases
- True Negative (TN): Correctly predicted diabetic cases
- False Positives (FP): Predicted diabetic but actually non diabetic
- False Negatives (FN): Predicted non diabetic but actually diabetic

These measures help to provide a comprehensive understanding of the model's prediction performance.

After the derivation of the confusion matrix, we computed the following evaluation metric to evaluate our logistic regression models and CART models.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	Shows how well model classifies individuals
$Sensitivity = \frac{TP}{TP+FN}$	Shows how well model detects actual diabetes
$Specificity = \frac{TN}{TN+FP}$	Shows how well model avoid false alarms
$Precision = \frac{TP}{TP+FP}$	Shows how reliable positive predictions are

5.4.1 Early Symptom Dataset

Model	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Logistic Regression	49	16	7	3	86.7	94.2	69.6	87.5
CART	50	20	3	2	93.3	96.2	87.0	94.3

The CART model achieved higher overall accuracy than logistic regression with better sensitivity, specificity and precision.

Model	Significant Predictors
Logistic Regression	gender, polyuria, polydipsia, sudden_weight_loss, genital_thrush, itching, partial_paresis
CART	polyuria, polydipsia, sudden_weight_loss, partial_paresis, polyphagia

These are the significant predictors from our models which align with recognised early warning signs of diabetes.

5.4.2 Health Indicators Dataset

Model	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Logistic Regression	239	209	91	61	74.7	79.7	69.7	72.4
CART	237	191	109	63	71.3	79.0	63.7	68.5

Logistic regression outperformed CART with higher accuracy and precision.

Model	Significant Predictors
Logistic Regression	HighBP, HighChol, CholCheck, BMI, HvyAlcoholConsump, GenHlth, AgeBrac
CART	HighBP, GenHlth, AgeBrac, BMI, HighChol

These are the significant predictors from our models which are consistent with established diabetes risk factors.

5.4.3 Overview

CART performs best for symptom based results which is ideal for public facing self assessment modules. Logistic regression is more effective for clinical indicator data, making it suitable for use in primary care analytics. Together, these models complement each other. CART enhances accessibility in preventive self checks while logistic regression supports data driven decision making for doctors.

6. Proposed Solution: DiaScope

By working on significant predictors of diabetes derived from our models, DiaScope aims to align with the goals of Healthier SG and focus on preventing onset diabetes, empowering individuals to actively manage their health and reduce hospital burden. Mainly, we want to integrate predictive analysis and digital health tools into existing platforms such as HealthHub and Healthy365 as well as existing digital health infrastructure (Appendix D Figure 1).

6.1 Individual Support

6.1.1 Personalised Plans

The first part of DiaScope focuses on modifiable risk factors such as high blood pressure, cholesterol, BMI and alcohol consumption which can be identified by our models in Section 5. These are lifestyle related risk that can be improved through behavioural changes which is why our solution encourages active participation in personalised diet and exercise programmes via Healthy365.

Predictive Risk Alerts

We plan to integrate our analytical solution into Healthy365, this means that the system can automatically send predictive-risk alerts to individuals who fall into higher-risk categories based on their BMI, glucose readings, and activity levels. These alerts, delivered through mobile notifications, would promptly recommend prevention actions, such as scheduling health screenings or joining community wellness.

The notifications are designed to be timely, contextual, and realistic. For instance, during meal times (around 8 a.m., 12 p.m., and 5 p.m.), users may receive messages such as: “Remember to opt for lower-sugar food options for lunch”. Similarly, activity reminders will encourage physical movement, such as “Take a walk during your break instead!” or “Alight one bus stop earlier to save money and clock extra steps!”. These nudges would help users make healthier decisions throughout the day and reinforce good habits over time.

Risk-Targeted Engagement

The risk-targeted engagement system extends the predictive alerts by tailoring the type and frequency of notifications based on the user’s characteristics and risk level. High risk users

would receive more frequent and focused guidance, such as invitation to nutrition webinars, fitness programmes or reminders to go for health screenings. Lower risk users would then receive general wellness tips or light reminders to maintain healthy routines. This level of personalisation ensures that interventions are meaningful and relevant, increasing user receptiveness and long term engagement.

Gamified Health Challenges

Lastly, to further drive motivation and participation, the proposed solution would integrate gamified health challenges into Healthy365. Identified diabetes-risk users can participate in themed challenges, such as “Sugar Smart Month” or “10,000 Steps a Day”, and earn health-focused rewards like discounts on healthier food options, fitness classes, or Health Promotion Board (HPB) e-vouchers. Having an engaging and rewarding preventive measure helps sustain behaviour change beyond one-off actions.

6.1.2 Community Specific Programmes

Mass Cooking Workshops on Diabetic Friendly Meals

This initiative would be conducted in collaboration with HPB and local community centres. These workshops teach participants how to prepare affordable, low sugar, and balanced meals using ingredients that are easily available in local markets. They also encourage family participation, helping households build sustainable healthy eating habits together.

Group Based Physical Activities

The second initiative focuses on Group-Based Physical Activities, such as community walks or jogs organised weekly by neighbourhood health ambassadors. These group sessions aim to promote regular physical activity while fostering a sense of community and mutual encouragement among participants. In order to maintain engagement, the activities would also be linked to the Healthy365 app (Figure 2) through team challenges or small rewards for consistent participation.

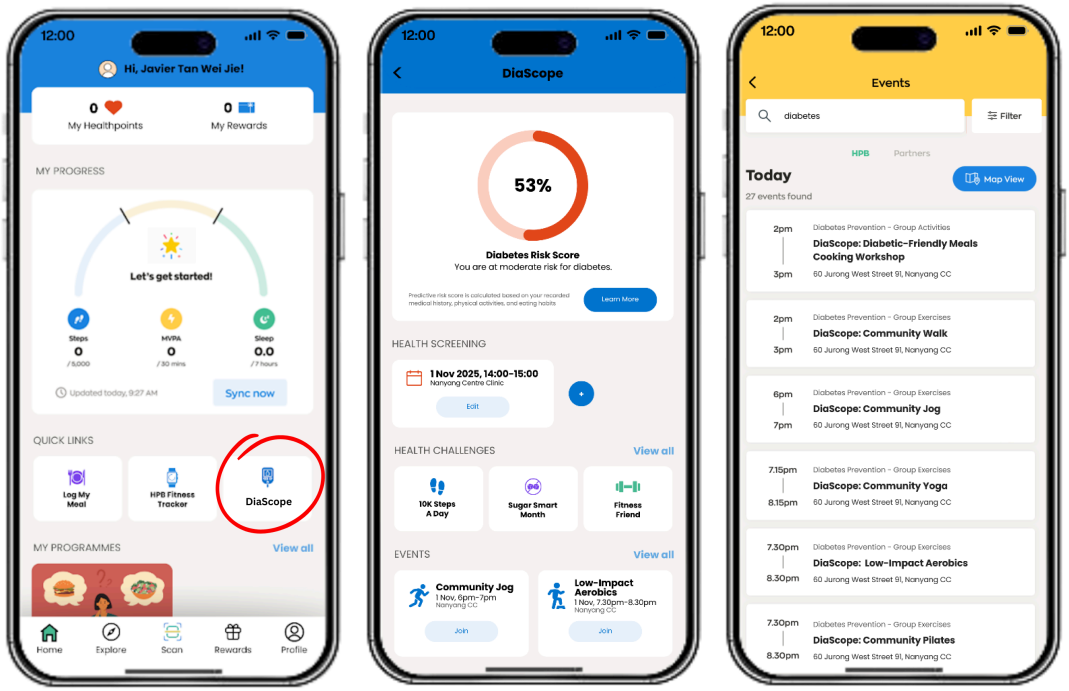


Figure 2: Integration of DiaScope into Healthy365 application

6.2 Primary Care Support

The second part of DiaScope is a tool that's designed for primary care support, mainly family doctors. This will include a Predictive Clinical Dashboard (PCD), and it will serve as a simple software tool for the doctor's office (Appendix D). The dashboard mainly extracts data from our predictive Logistic Regression and CART models to make treatment in clinics more proactive than reactive, which is inline with the Healthier SG vision.

Who to Help First Model

The first pillar of the solution acts as a decision based system, where it decides "who to help first". One key problem stakeholders in primary care support is the large influx of patients and a lack of a robust decision making matrix on who to help first, or to follow up with. Therefore, our predictive models will be used to generate a simple T2DM Risk Score, categorised into Low, Moderate or High for every patient enrolled into the clinic. Therefore, instead of viewing one long list of patients, doctors simply look at the dashboard which will automatically rank all registered patients by their diabetic risk, allowing doctors and the care team to focus on the more important cases. This enables efficiency, as it provides doctors and their team with a focus on the highest risk patients, dedicating their limited resources to them.

Priority will then go into scheduling these individuals for check-ins and health plan discussions, supporting the Healthy365 programme currently envisioned.

Integrated and Continuous Data Collection and Monitoring

The use of a risk score indicates a need for it to be constantly updated with new data. The PCD is designed to automatically sync and update two key data sources: Patient generated data from their Healthy365 App and their home devices, as well as clinical data from health screening and lab tests. This enhances data sharing across platforms, and allows for a continuous care monitoring system in which primary care support can be used. For example, if a high-risk patient's home blood sugar suddenly spikes, the dashboard can then send an automated alert to the clinic, which then allows the team to intervene weeks or even months before the patient's next scheduled visit. This proactive monitoring shifts the focus away from reactive solutions, and is the key to preventing complications and further directly addresses a key long term goal for Healthier SG, which is to reduce avoidable Emergency Department attendance.

7. Business Application

7.1 Approach: Integrating Predictive Modeling

To operationalise the proposed diabetes prevention plans under Healthier SG, DiaScope will adopt a four stage implementation framework (Appendix E Figure 1).

Each stage details how predictive analysis can be embedded into existing digital health infrastructure ensuring clinical feasibility, governance and scalability.

7.1.1 Phase 1: Data Collection and Governance

Data will be collected through partnership with public hospitals and primary care clusters (Appendix E Table 1). Two complementary streams will be established:

- Individual reported data via HealthHub questionnaires capturing lifestyle habits and early diabetes symptoms. Questions asked will be mainly on the significant predictors we have derived earlier as seen in Appendix E Table 2.

- Clinician reported data including blood pressure and cholesterol levels extracted from electronic medical records.

Consent and privacy will follow PDPA guidelines with data sharing agreements, deidentification and role based access. Questionnaires can be completed digitally before clinic visits while clinical measures are drawn from routine check ups, keeping the process low burden and feasible.

7.1.2 Phase 2: Generation of Diabetes Risk Scores

A combined dataset will train logistic regression and CART models to predict an individual's probability of developing diabetes. Both models will be cleaned, calibrated and validated using techniques shown in Section 2 and 3 to ensure reliability before integration into the HealthHub ecosystem.

Predicted probabilities will be categorised into 3 operational tiers:

Tier	Probability Threshold	Measures
High Risk	$\geq$ 0.30	• HealthHub push notification to book screening • Flagged on doctor dashboard
Moderate Risk	0.15 - 0.29	• Recommended to join Healthy365 wellness programmes • Reassessment in 6-12 months
Low Risk	$\leq$0.14	• Maintain lifestyle plan • Receive standard preventive reminders

The probability threshold is determined on our models in Section 5 and adjusted to align with preventive healthcare objectives. Although our logistic regression model used a threshold of 0.5, a lower threshold (0.3) is applied to our system to ensure that prediabetic individuals are also flagged as high risk. This approach prioritises early intervention as prediabetes is largely irreversible and timely lifestyle changes are crucial to prevent progression to diabetes.

Each individual's risk tier and key contributing factors will be displayed within HealthHub. Individuals can view their score on the app with corresponding tips to improve on their score.

Doctors access numeric risk, contributing variables and clinical recommendations through the Clinical Dashboard. The risk score engine directly supports Healthier Sg objectives by enabling personalised health plans, doctor engagement and targeted community programme referrals.

7.1.3 Phase 3: Dynamic Updates and Automation

Risk scores will evolve automatically as new behavioural and clinical data become available. We will mainly take into account 2 updates:

- Behavioural updates: Participation in Healthy365 challenges, verified activity levels or healthier meal logs will show estimated improvements in risk. For instance, after completing a health challenge, individuals will receive a notification “if sustained 8 weeks, risk likely decrease 5%”
- Clinical updates: Annual screenings and new laboratory results will trigger recalculation of the official risk score which will be updated on the HealthHub app timely.

This continuous feedback mechanism ensures that the predictive models remain current and responsive to changes in user behaviour and health status. It allows individuals to visualise tangible progress in their health journey while enabling doctors to receive near real time updates on individuals whose risk levels have increased, ensuring high risk cases are prioritised for follow up.

7.1.4 Phase 4: Clinical Regulation

With our collaborators, we will set up a multidisciplinary Model Oversight Committee comprising clinicians and data scientists who will review model outputs quarterly. Special attention will be paid to false negative cases to avoid missing high risk individuals and to false positive cases to manage clinical workload. Model Interpretability will be maintained by providing coefficient summaries from logistic regression and decision rule paths from CART, fostering clinician trust. Key performance indicators for evaluation will include evaluation metric as stated in Section 4.

7.2 Strengths

The main strength of DiaScope is that it provides a holistic, reactive data driven ecosystem that directly supports both patients and doctors, effectively allowing the healthcare model to shift from a reactive model to a proactive model.

Digital Patient Tools

The current approach is a broad, self-directed "one-size-fits-all" app. It is used for tracking steps, diet, and accessing general community activities. DiaScope is an active and personalized health companion that uses predictive analytics to send personalized, predictive risk alerts to high-risk individuals. It recommends risk-targeted engagement to push the right intervention to its target audience: For example a diet challenge vs. an activity challenge to the user who needs it most.

Primary Care Management

Family doctors are tasked with delivering preventive care and developing health plans for their entire enrolled population. This is a significant new mandate that relies on annual check-ins and existing clinical data. Diascope provides an actionable, data-driven clinical dashboard. It employs a “who to treat” feature, allowing the care team to efficiently focus its limited resources on the highest-risk patients and manage resources and time allocation efficiently It also delivers a comprehensive, real-time data view by integrating patient-generated app data with clinical records, acting as a "secondary brain" for the doctor.

Social Prescriptions

Doctors make general social prescriptions referring residents to a wide range of community activities. This process is manual and reliant on the doctor's general knowledge and is non specific. DiaScope is an automated and targeted recommendation engine, where E. This allows for a far more efficient use of community resources (e.g., recommending a cooking class to one user vs. an educational course to another) and increases the likelihood of patient adherence.

7.3 Limitations

There are a few limitations with our current proposal that will be discussed here.

Unrepresentative Datasets

Firstly, we agree that the data is not representative. Our models were built largely on publicly available data sets from other countries, and are not representative of Singapore's multiracially genetic diverse makeup. Therefore, while our solution currently demonstrates methodology and a proof of concept, it is currently not an accurate diagnosis and accurate sight picture of the local context. This will be discussed further under section 6.4.

Narrow Scope

Secondly, our solution focuses on one chronic disease, diabetes. However, healthier SG tackles a wide range of common chronic conditions such as hypertension, lipid disorders etc. The dashboard to flag diabetes may be too simplistic currently, and has a narrow range of applications.

User Uptake

Lastly, one of the main challenges presented is the uptake of DiaScope by Singaporean Citizens. Individuals need to have the motivation to want to use the app in order for DiaScope to function properly. Therefore, by integrating DiaScope into current apps, we can circumvent this issue by allowing pre existing users of said apps to use DiaScope as part of an in app functionality, instead of using an entirely new app which creates friction. Furthermore, apps like Healthy365 are government backed and based, and can give users the peace of mind about their data protection.

7.4 Future Improvements

Future improvements to solve the current limitations are very feasible and immediately implementable.

Firstly, the model can be validated and trained with local datasets to improve models accuracy and classification, and to get a better view of the local population's health needs. This will benefit DiaScope significantly as the application will be able to provide accurate model representation of the population it is going to be implemented in, thereby improving efficiency and accuracy.

Secondly, DiaScope should be expanded to include other chronic diseases. This will allow DiaScope to become a holistic platform to target most chronic illnesses present in Singapore,

therefore slowly expanding the scope of solution and reaching more patients. Doing this will allow all Singaporeans suffering from at least one chronic illness reap the benefits of DiaScope, allowing our solution to be more comprehensive and adaptable.

8. Conclusion

This project shows how predictive analytics can help strengthen Singapore's Healthier SG initiative by supporting early detection and prevention of diabetes. Using logistic regression and CART models, we were able to identify key behavioural and clinical predictors which are all consistent with established medical articles.

Our results show that CART is more effective for symptom based data and can be used in public facing tools for early self assessment while logistic regression offers clearer insights for clinical decision support. Together, these models provide a complementary framework that combines accuracy with transparency.

The proposed solution, DiaScope, integrates these predictive models into existing platforms like HealthHub and Healthy 365, to create personalised health plans for individuals and data driven dashboards for family doctors. This supports Healthier SG's shift from reactive to proactive healthcare and their goal of prevention, empowerment and reduced hospital burden.

While current models are based on international datasets and serve as a proof of concept, future work should focus on local dataset validation, testing with healthcare clusters and expansion to other chronic diseases.

In conclusion, DiaScope highlights how data driven solutions can make preventive care more targeted, efficient and sustainable for both individuals and healthcare providers in Singapore.

9. Bibliography

Larxel, A. (2021, December 6). *Early classification of diabetes* [Dataset]. Kaggle. https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification/data

Ministry of Health. (2024, June 6). *Prevalence of Hypertension, Diabetes, High Total Cholesterol, Obesity and Daily Smoking* [Dataset]. data.gov.sg. https://data.gov.sg/datasets/d_efea9966c502767122171c61d88062db/view

Ministry of Health. (2017). *Studying Measures to Better Support Persons with Pre-Diabetes*. Ministry of Health. https://www.moh.gov.sg/newsroom/ministry-of-health-studying-measures-to-better-support-persons-with-pre-diabetes/

Teboul, A. (2021, November 8). *Diabetes Health Indicators Dataset* [Dataset]. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

World Health Organization. (2024, November 14). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

10. Appendix

Appendix A: Data Dictionary

Variable Name	Description
age	Age
gender	Gender
polyuria	0: Never experience excessive urination 1: Experienced excessive urination
polydipsia	0: Never experience excessive thirst/excess drinking 1: Experienced excessive thirst/excess drinking
sudden_weight_loss	0: Never experience sudden weight loss 1: Experienced sudden weight loss
weakness	0: Never experience feeling weak 1: Experienced feeling weak
polyphagia	0: Never experience excessive/extreme hunger 1: Experienced excessive/extreme hunger
genital_thrush	0: Never had yeast infection 1: Had yeast infection
visual_blurring	0: Never experience having blurred vision 1: Experienced having blurred vision
itching	0: Never experience having itch 1: Experienced having itch
irritability	0: Never experience having itch 1: Experienced having irritability
delayed_healing	0: Never observed delayed healing when injured 1: Observed delayed healing when injured
partial_paresis	0: Never experience weakening of a muscle 1: Experienced weakening of a muscle
muscle_stiffness	0: Never experience having stuff muscles 1: Experienced having stuff muscles
alopecia	0: Never experience hair loss

	1: Experienced hair loss
obesity	0: Not obese 1: Obese
class	0: No diabetes 1: Diabetic

Table 1: Variables in Early Symptoms Dataset

Variable	Description
Diabetes_012	0: No diabetes 1: Prediabetes 2: Diabetes
HighBP	0: No high blood pressure 1: High Blood Pressure
HighChol	0: No high cholesterol 1: High cholesterol
CholCheck	0: No cholesterol check in 5 years 1: Had cholesterol check in 5 years
BMI	Body Mass Index
Smoker	0: Has never smoked at least 100 cigarettes in their entire life 1: Smoked at least 100 cigarettes in their entire life
Stroke	0: Did not have stroke 1: Had stroke
HeartDiseaseorAttack	0: Has no coronary heart disease or myocardial infarction 1: Has coronary heart disease or myocardial infarction
PhysActivity	0: Did not have physical activity in the past 30 days 1: Had physical activity in the past 30 days
Fruits	0: Did not consume fruits 1 or more times a day 1: Consumed fruits 1 or more times a day
Veggies	0: Did not consume vegetables 1 or more times a day 1: Consumed vegetables 1 or more times a day
HvyAlcoholConsump	0: drank less than 14 drinks per week (adult men)

	drank less than 7 drinks per week (adult women) 1: drank more than 14 drinks per week (adult men) drank more than 7 drinks per week (adult women)
AnyHealthcare	0: Does not have healthcare coverage 1: Has healthcare coverage
NoDocbcCost	0: Was not able to see a doctor when needed in past 12 months 1: Was able to see a doctor when needed in past 12 months
GenHlth	1: Excellent 2: Very Good 3: Good 4: Fair 5: Poor
MentHlth	Scale 1-30 days based on how many days during the past 30 days has the respondent's mental health been not good
PhysHlth	Scale 1-30 days based on how many days during the past 30 days has the respondent's physical health been not good
DiffWalk	0: Does not have serious difficulty walking or climbing stairs 1: Has serious difficulty walking or climbing stairs
Sex	0: Female 1: Male
Age	13 level age category with 5 years interval 1: 18-24 13: 80 or older
Education	1: Never attended school or only kindergarten 2 : Grades 1 through 8 (Elementary) 3: Grades 9 through 11 (Some high school) 4: Grade 12 or GED (High school graduate) 5: College 1 year to 3 years (Some college or technical school) 6: College 4 years or more (College graduate)
Income	Scale 1-8 1: less than $10.000 5: less than $35,000 8: $75,000 or more

Table 2: Variables in Health Indicators Dataset

Appendix B: Data Cleaning and Exploration

Early Symptoms Dataset:

NA checks:

```
> sum(is.na(early.symptoms.dt))
[1] 0
> colSums(is.na(early.symptoms.dt))
                age              gender            polyuria          polydipsia sudden_weight_loss
                  0                   0                   0                   0                   0
           weakness          polyphagia      genital_thrush     visual_blurring            itching
                  0                   0                   0                   0                   0
      irritability     delayed_healing     partial_paresis     muscle_stiffness           alopecia
                  0                   0                   0                   0                   0
            obesity               class
                  0                   0
```

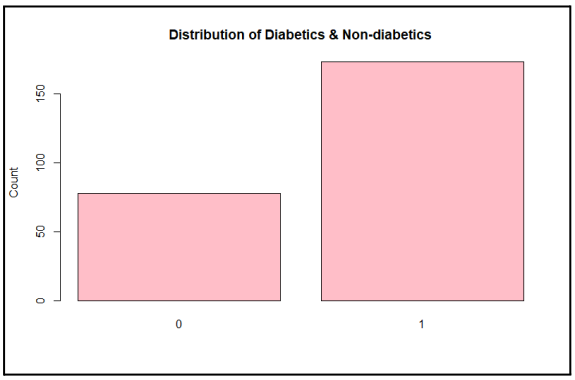
Duplicate checks:

```
> sum(duplicated(early.symptoms.dt))
[1] 269
```

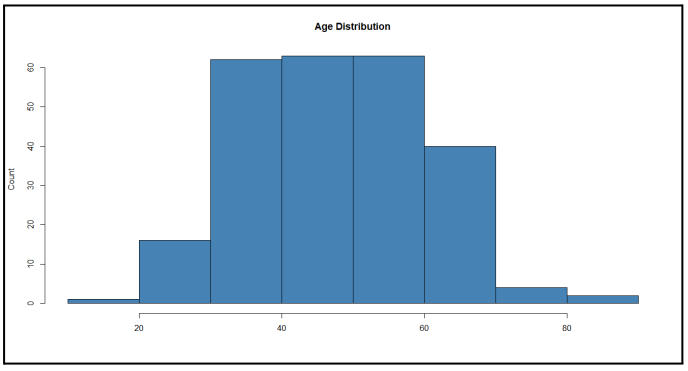
Handling duplicates:

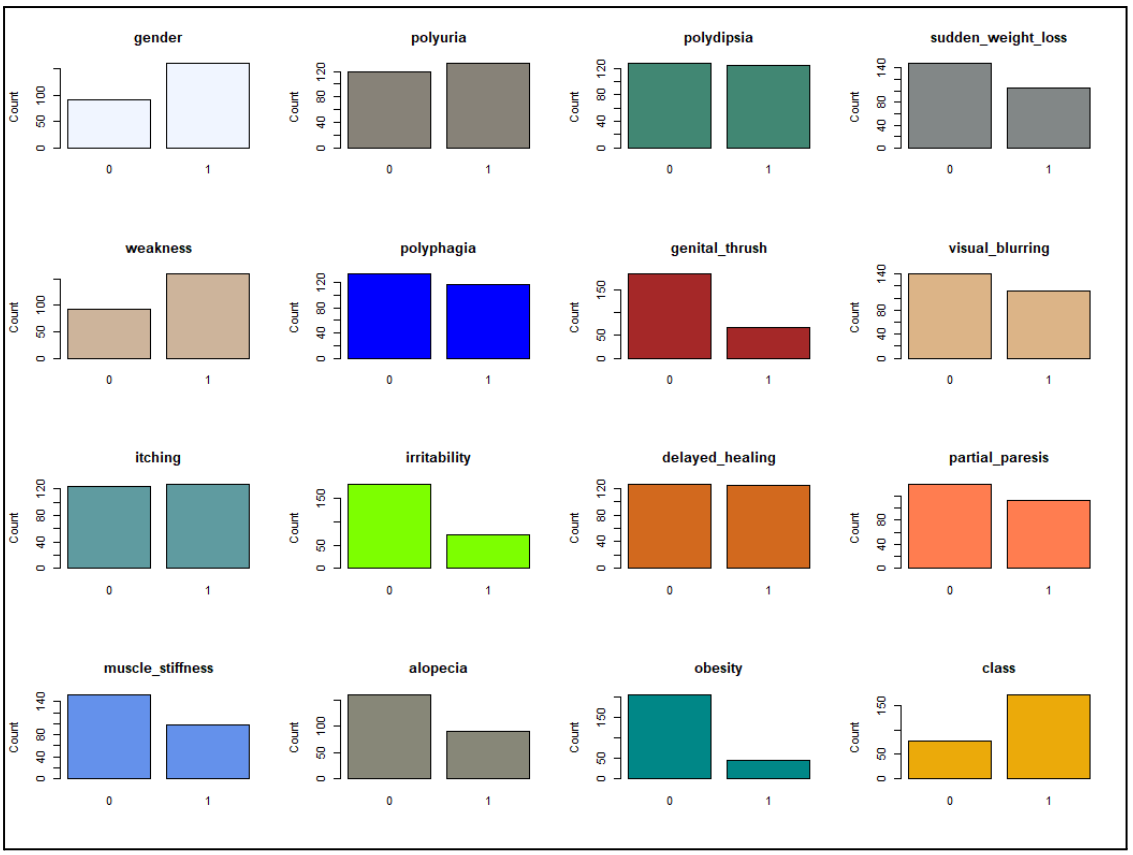
```
> early.symptoms.dt <- unique(early.symptoms.dt)
> dim(early.symptoms.dt)
[1] 251  17
```

Class distribution:

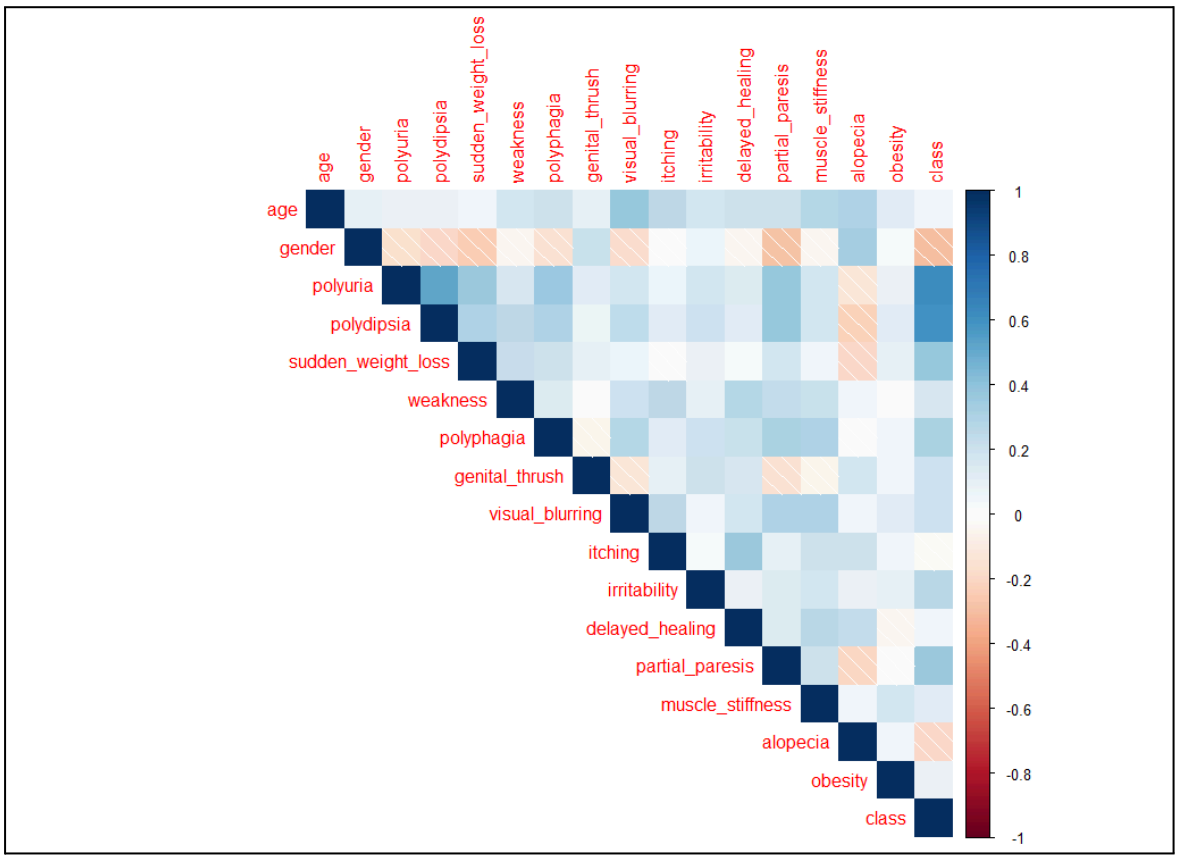


Variable distributions:





Correlation Heatmap:



Health Indicators Dataset:

NA checks:

```
> colSums(is.na(health.ind.dt))
       Diabetes_012              HighBP            HighChol           CholCheck                 BMI
                  0                   0                   0                   0                   0
             Smoker              Stroke HeartDiseaseorAttack        PhysActivity              Fruits
                  0                   0                   0                   0                   0
            Veggies    HvyAlcoholConsump        AnyHealthcare        NoDocbcCost             GenHlth
                  0                   0                   0                   0                   0
           MentHlth             PhysHlth            DiffWalk                 Sex                 Age
                  0                   0                   0                   0                   0
          Education              Income
                  0                   0
```

Duplicate checks:

```
> sum(duplicated(health.ind.dt))
[1] 23899
```

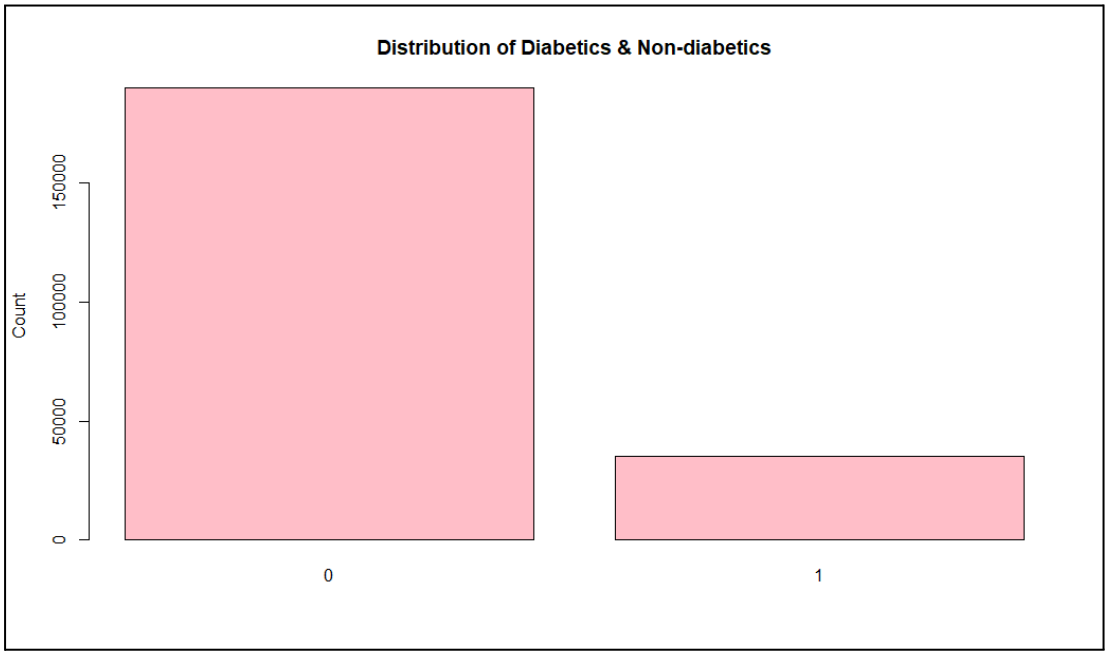
Handling duplicates:

```
> # create copy to clean
> clean.dt <- copy(health.ind.dt)
> # remove duplicates
> clean.dt <- unique(clean.dt)
> dim(clean.dt)
[1] 229781     22
```

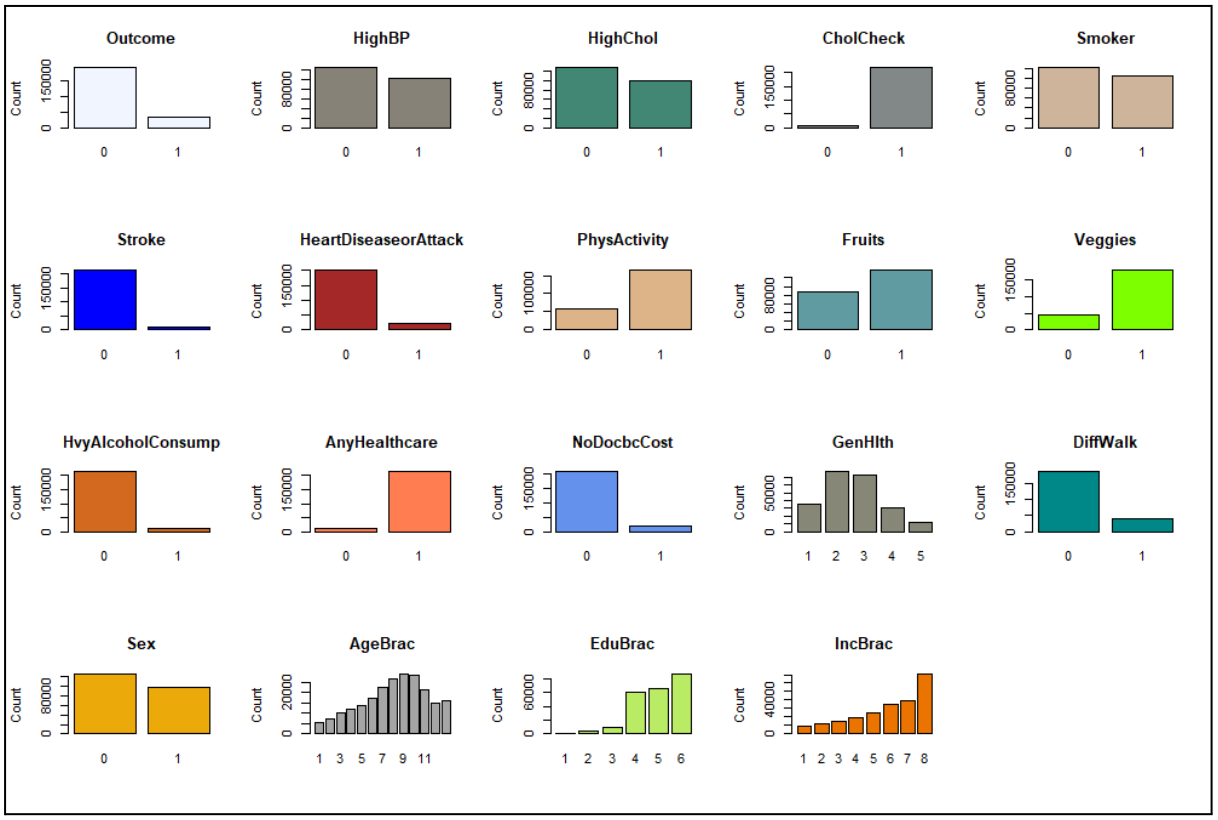
Make target Y variable binary:

```
> # remove pre diabetes (make it only yes or no)
> clean.dt <- clean.dt[Diabetes_012 != 1]
> clean.dt[Diabetes_012 == 2, Diabetes_012 := 1]
```

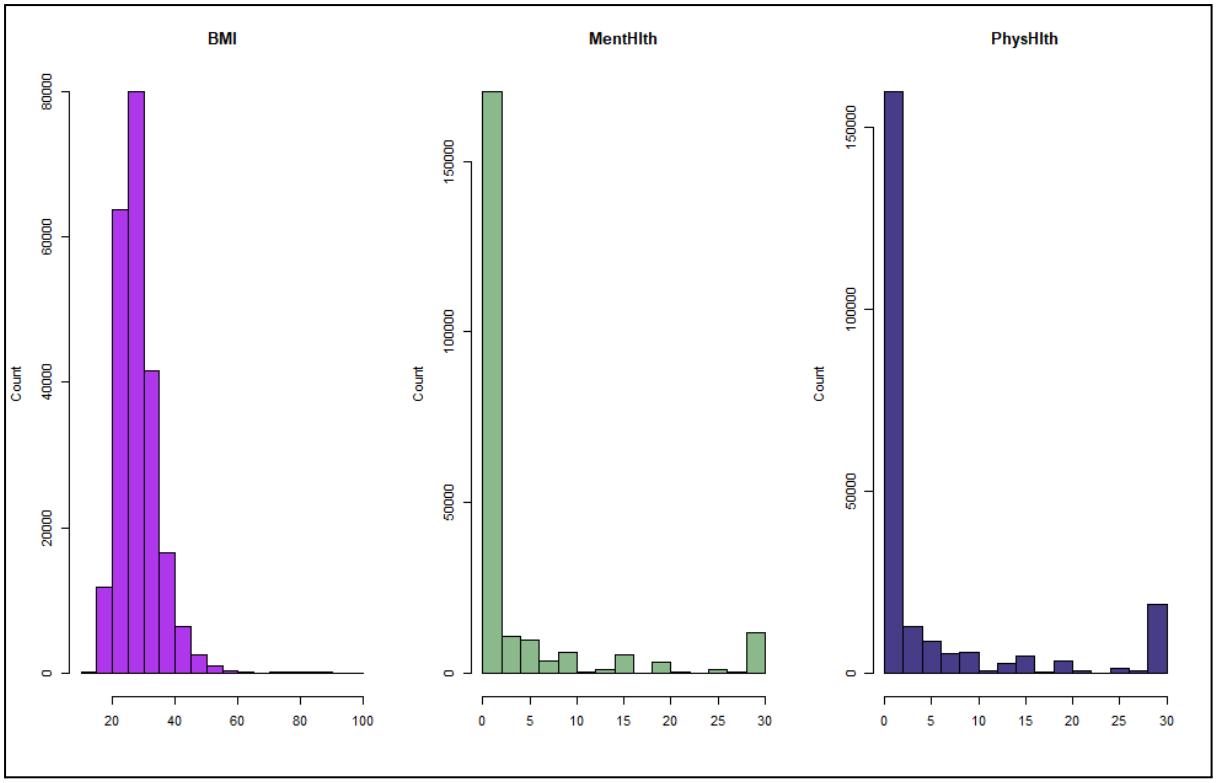
Class distribution:



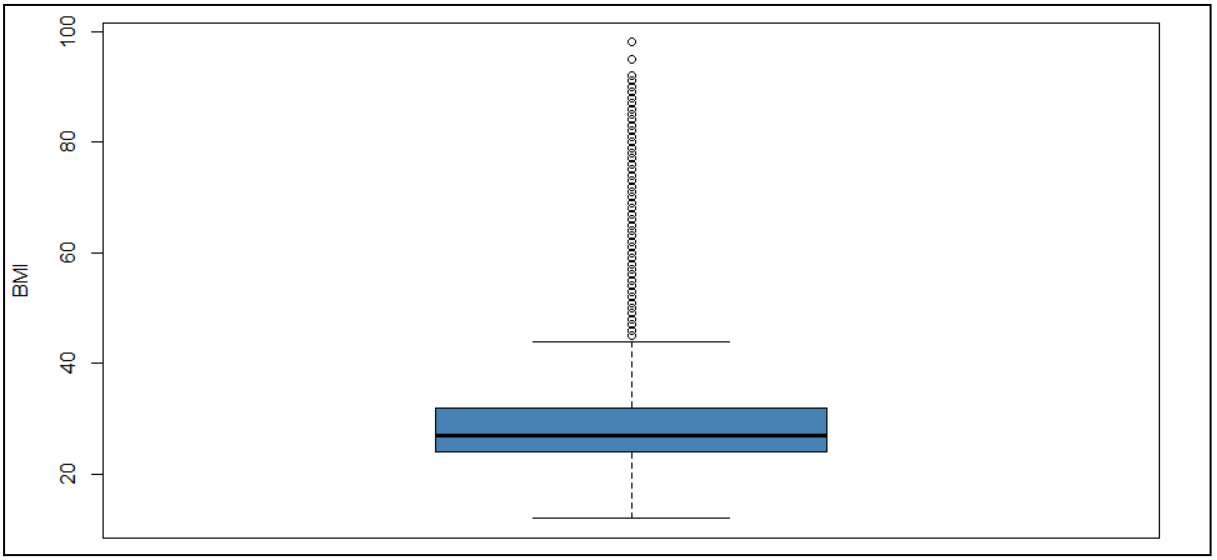
Bar charts of categoricals:



Histograms of continuous:



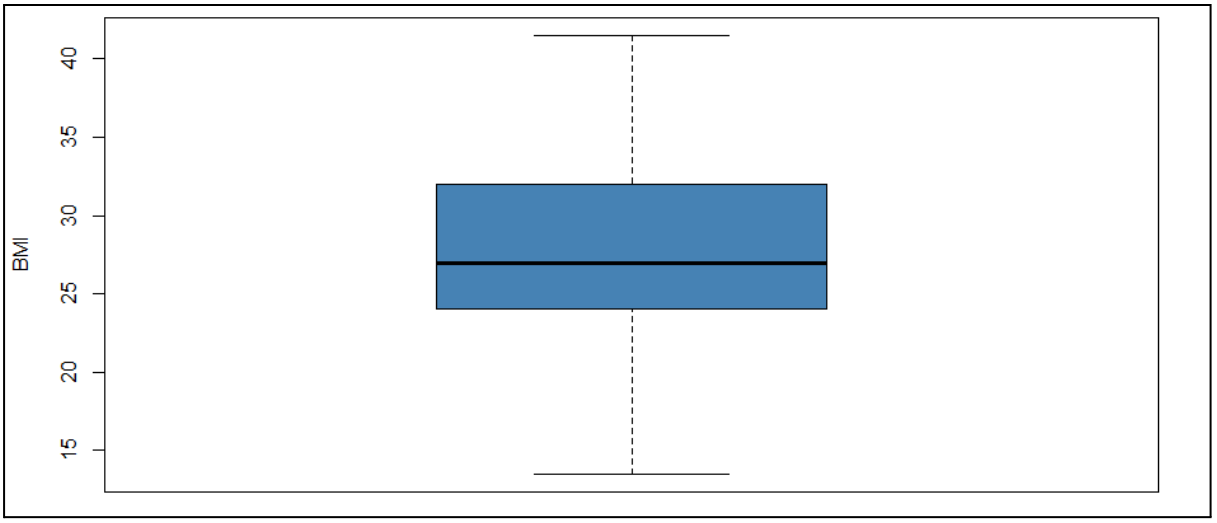
Boxplot of BMI with outliers:



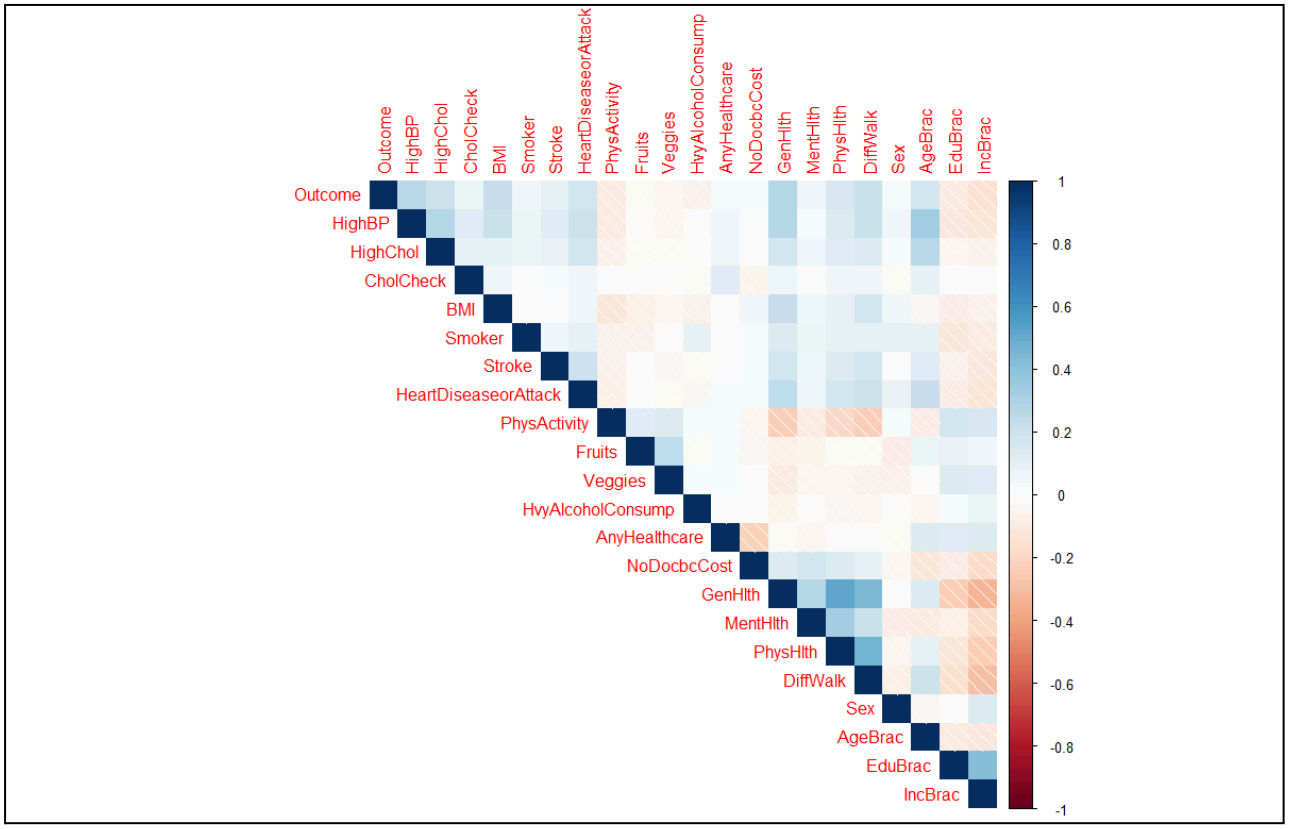
Removal of outliers:

```
> summary(health.ind.dt$BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   24.00   27.00   28.38   31.00   98.00
> Q1 <- quantile(health.ind.dt$BMI, 0.25)
> Q3 <- quantile(health.ind.dt$BMI, 0.75)
> IQR <- Q3 - Q1
> UL <- Q3 + (1.5 * IQR)
> LL <- Q1 - (1.5 * IQR)
> clean.dt[BMI < LL, BMI := LL]
> clean.dt[BMI > UL, BMI := UL]
> summary(clean.dt$BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.50   24.00   27.00   28.34   32.00   41.50
```

Boxplot of BMI after removal:



Correlation Heatmap:



Random Sampling:

```
> # subset each class
> set.seed(100)
> dia <- clean.dt[Outcome == 1]
> non.dia <- clean.dt[Outcome == 0]
> # randomly select 1000 samples each
> dia.sample <- dia[sample(.N, size = 1000)]
> non.dia.sample <- non.dia[sample(.N, size = 1000)]
> balanced.dt <- rbind(dia.sample, non.dia.sample)
> # check balance
> prop.table(table(balanced.dt$Outcome))

  0   1
0.5 0.5
> summary(balanced.dt$Outcome)
   0    1
1000 1000
```

Appendix C: Models

Early Symptom Dataset

```
Call:
glm(formula = class ~ gender + polyuria + polydipsia + sudden_weight_loss +
    genital_thrush + visual_blurring + itching + partial_paresis,
    family = "binomial", data = trainset)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.2174     0.6040  -0.360 0.718924
genderMale          -2.1722     0.7083  -3.067 0.002163 **
polyuria1            3.5118     0.9919   3.540 0.000400 ***
polydipsia1          4.2266     1.0497   4.026 5.66e-05 ***
sudden_weight_loss1  1.7777     0.7541   2.357 0.018400 *
genital_thrush1      2.4571     0.8595   2.859 0.004253 **
visual_blurring1     1.4401     0.7843   1.836 0.066335 .
itching1            -2.8071     0.8433  -3.329 0.000873 ***
partial_paresis1     1.5717     0.7798   2.016 0.043836 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 218.622  on 175  degrees of freedom
Residual deviance:  72.752  on 167  degrees of freedom
AIC: 90.752

Number of Fisher Scoring iterations: 8
```

Health Indicators Dataset

```
Call:
glm(formula = Outcome ~ HighBP + HighChol + CholCheck + BMI +
    HvyAlcoholConsump + GenHlth + AgeBrac, family = "binomial",
    data = trainset)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.89515    0.76029 -10.384  < 2e-16 ***
HighBP1             0.76683    0.13507   5.677 1.37e-08 ***
HighChol1           0.68377    0.12828   5.330 9.81e-08 ***
CholCheck1          1.64129    0.57924   2.834 0.004604 **
BMI                 0.09107    0.01177   7.735 1.03e-14 ***
HvyAlcoholConsump1 -1.28696    0.37428  -3.438 0.000585 ***
GenHlth             0.45852    0.06509   7.044 1.87e-12 ***
AgeBrac             0.17429    0.02556   6.820 9.09e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1940.8  on 1399  degrees of freedom
Residual deviance: 1500.4  on 1392  degrees of freedom
AIC: 1516.4

Number of Fisher Scoring iterations: 5
```

Table 1: Logistic Regression Model after step( ) for both Datasets

Early Symptom Dataset

```
vif(e2)
          gender          polyuria        polydipsia sudden_weight_loss     genital_thrush    visual_blurring
        1.305998          1.267411          1.631159          1.059397          1.390348          1.602398
         itching    partial_paresis
        2.014813          1.131832
```

Health Indicators Dataset

```
vif(m2)
          HighBP          HighChol         CholCheck               BMI HvyAlcoholConsump           GenHlth
        1.101178          1.030035          1.007725          1.115464          1.003999          1.047713
         AgeBrac
        1.138127
```

Table 2: Variance Inflation Factor (VIF) of Retained Predictors of both Datasets

Early Symptom Dataset

```
OR <- exp(coef(e2))
OR
      (Intercept)          genderMale           polyuria1         polydipsia1 sudden_weight_loss1
       0.80461652          0.11392461         33.50963042         68.48101295           5.91636218
  genital_thrush1    visual_blurring1            itching1    partial_paresis1
      11.67121304          4.22111175          0.06038082          4.81491113
```

Health Indicators Dataset

```
> OR <- exp(coef(m2))
> OR
      (Intercept)           HighBP1          HighChol1         CholCheck1                BMI HvyAlcoholConsump1
     0.0003725472        2.1529404001        1.9813277269        5.1618028034        1.0953488333        0.2761088684
           GenHlth            AgeBrac
     1.5817254789        1.1904052448
```

Table 3: Odds Ratio (OR) of Retained Predictors of both Datasets

Early Symptom Dataset

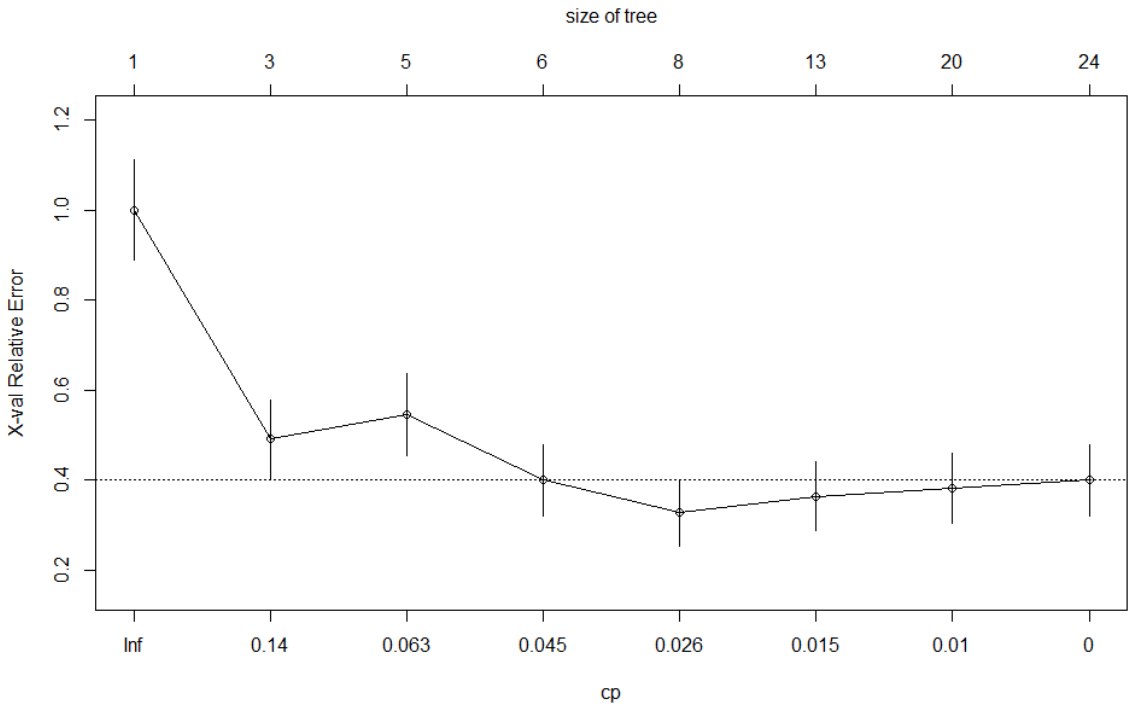
```
> OR.CI
                          2.5 %      97.5 %
(Intercept)          0.239412513   2.6762130
genderMale           0.025622228   0.4276879
polyuria1            5.914944675 308.7299685
polydipsia1         10.827291427 696.9103254
sudden_weight_loss1  1.441139936  29.0024789
genital_thrush1      2.400171266  72.8769536
visual_blurring1     0.962706740  21.9228677
itching1             0.009557894   0.2726853
partial_paresis1     1.100547817  24.5635046
```

Health Indicators Dataset

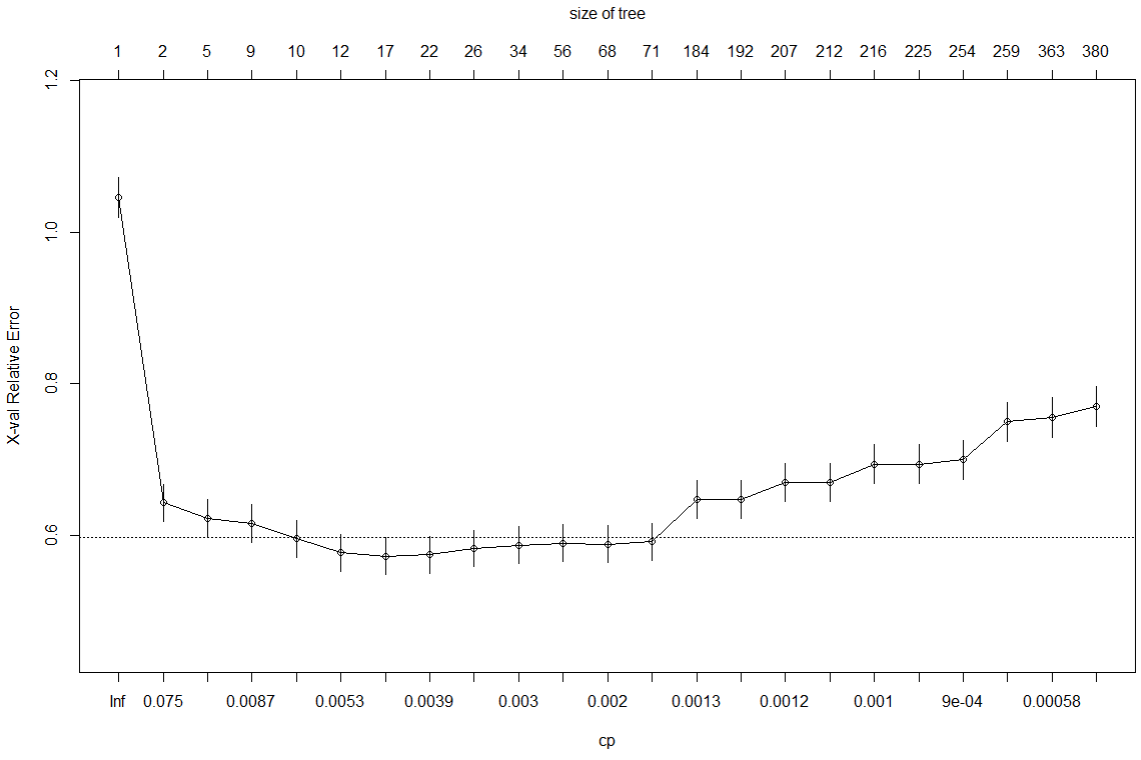
```
> OR.CI
                          2.5 %       97.5 %
(Intercept)        7.742745e-05  0.001543047
HighBP1            1.652157e+00  2.806079424
HighChol1          1.541198e+00  2.548868538
CholCheck1         1.779189e+00 17.700180899
BMI                1.070605e+00  1.121211872
HvyAlcoholConsump1 1.281440e-01  0.561282878
GenHlth            1.393847e+00  1.799342825
AgeBrac            1.132857e+00  1.252312452
```

Table 4: Confidence Interval (CI) of Retained Predictors of both Datasets

Early Symptoms Dataset



Health Indicators Dataset



Automated computation of optimal CP based on 1 SE rule:

```
> # compute cp
> CVerror.cap <- m.cart$cptable[which.min(m.cart$cptable[, "xerror"]), "xerror"] + m.cart$cptable[which.min(m.cart$cptable
[, "xerror"]), "xstd"]
> i <- 1
> while (m.cart$cptable[i, 4] > CVerror.cap) {
+   i <- i + 1
+ }
> cp.opt <- ifelse(i > 1, sqrt(m.cart$cptable[i, 1] * m.cart$cptable[i-1, 1]), 1)
```

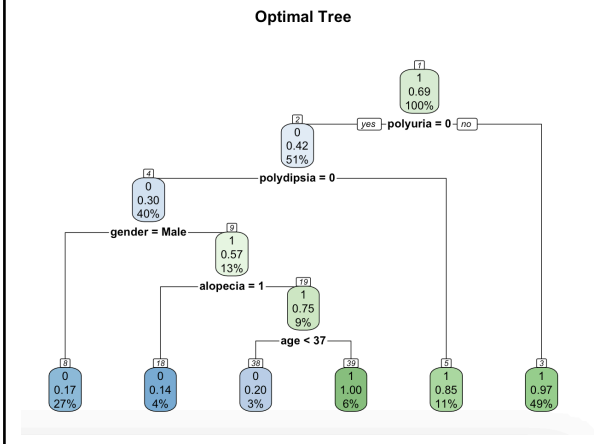
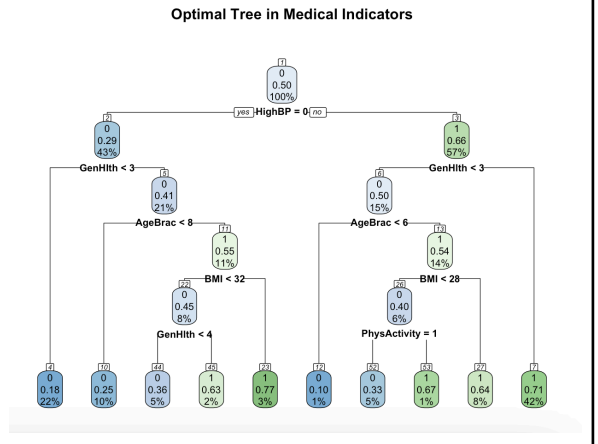
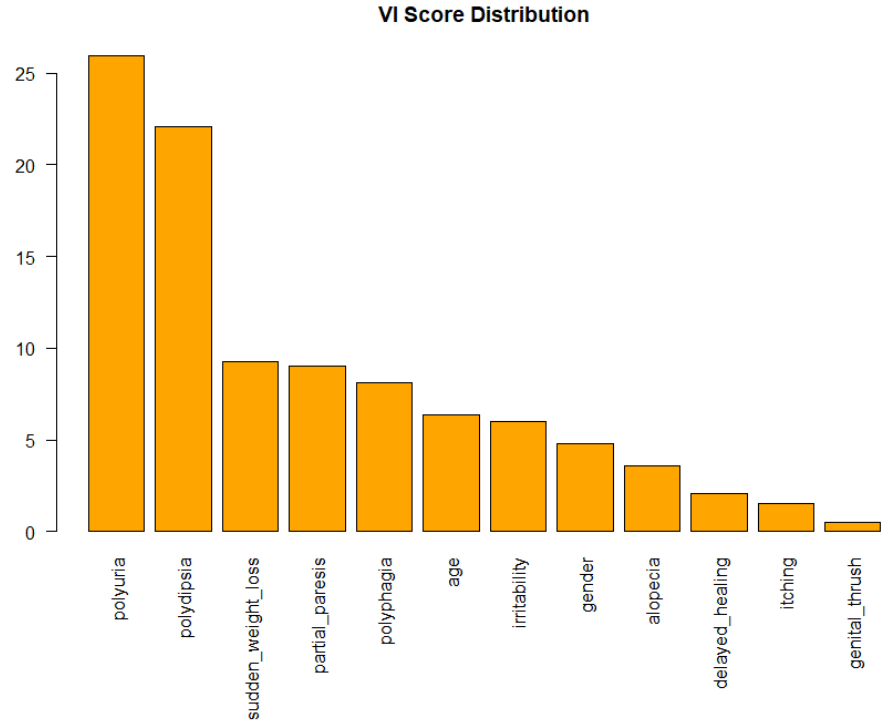
Early Symptom Dataset	Health Indicators Dataset
	

Table 5: Optimal trees after pruning for both Datasets

Early Symptoms Dataset



Health Indicators Dataset

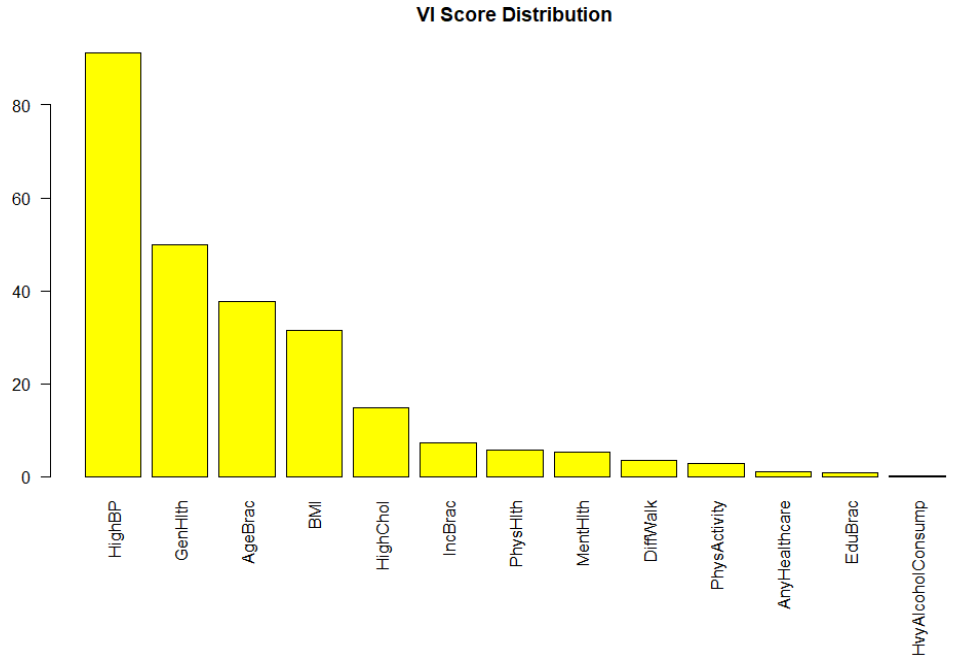


Table 6: VI Score Distributions for both Datasets

Appendix D: Solutions

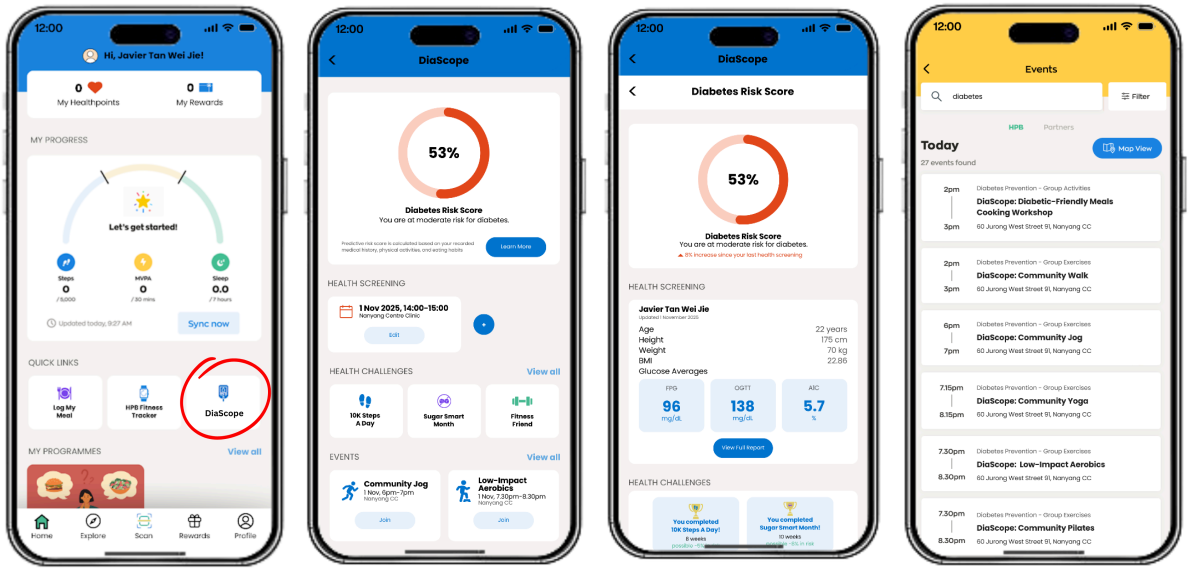
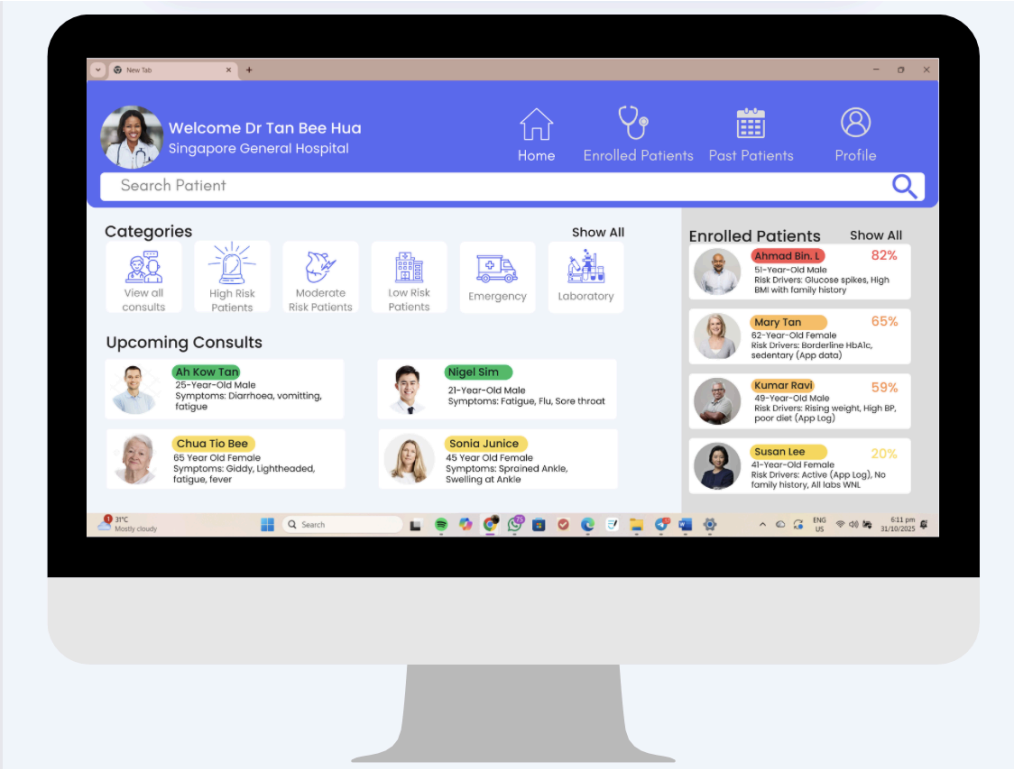


Figure 1: Integration of DiaScope into Healthy365 application



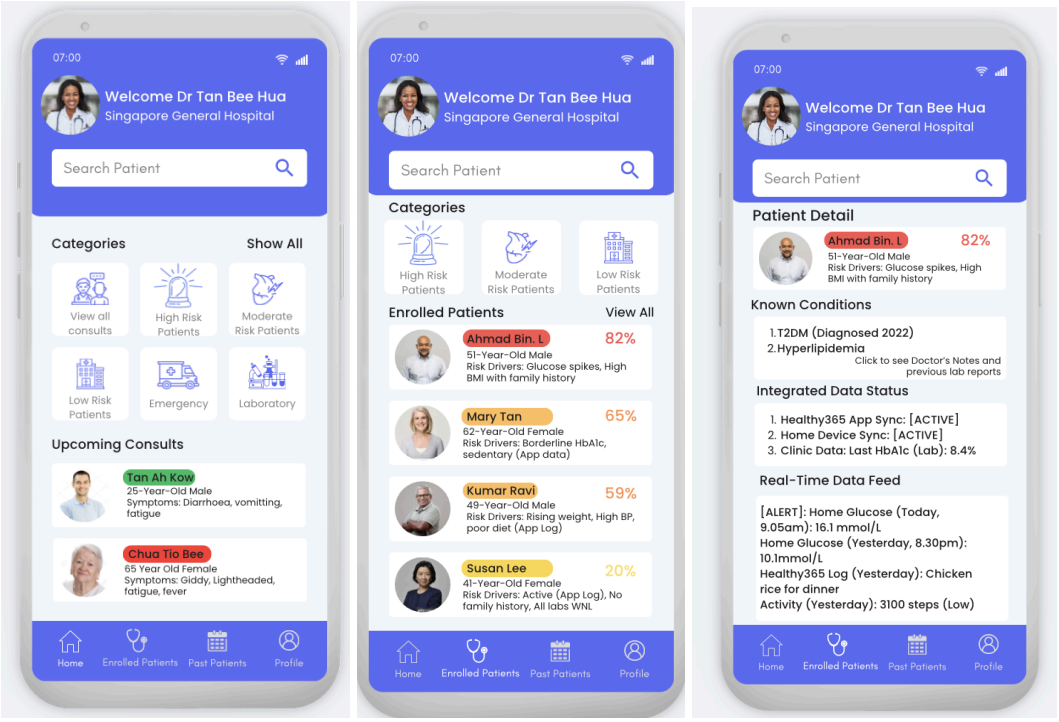


Figure 2: DiaScope Data Dashboard Interface for Healthcare Professionals

Appendix E: Business Application

Partners	Strength
National University Health System (NUHS): • National University Hospital • Ng Teng Fong General Hospital • National University Polyclinics	• academic research environment • established data sharing frameworks with NUS
SingHealth Cluster: • Singapore General Hospital • Changi General Hospital • Sengkang General Hospital	• largest patient base in Singapore

• SingHealth Polyclinics	
National Healthcare Group (NHG): • Tan Tock Seng Hospital • Khoo Teck Phuat Hospital • National Healthcare Group Polyclinics	• strong chronic disease management and population health data
Raffles Medical Group	• extensive GP network for voluntary participation in data collection

Table 1: Possible Partners to Collaborate with on Data Collection

No.	Question	Target Predictor
1	What is your age?	AgeBrac
2	What is your BMI (Body Mass Index)?	BMI
3	Would you say that in general your health is (scale 1-5) 1: excellent 2: very good 3: good 4: fair 5: poor	GenHlth
4	Are you a heavy drinker? No: drank less than 14 drinks per week (adult men) drank less than 7 drinks per week (adult women)	HvyAlcoholConsump

	Yes: drank more than 14 drinks per week (adult men) drank more than 7 drinks per week (adult women)	
5	Have you had your cholesterol checked in 5 years?	CholCheck
6	Have you experienced excessive urination?	polyuria
7	Have you experienced excessive thirst/ excess drinking?	polydipsia
8	Have you experienced an episode of sudden weight loss?	sudden_weight_loss
9	Have you experienced an episode of weakening of a muscle/a muscle group?	partial_paresis
10	Have you experienced an episode of excessive/ extreme hunger?	polyphagia

Table 2: List of Questions for Questionnaire on Individual Reported Data