NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2406 Analytics I

AY 2025/2026 Semester 1

Submitted by Team 7:

Aloysious Law Jia Jian	**U2420162J**
Bryan Lim How Meng	**U2420547L**
De Souza Alyssa Anne	**U2210429D**
Ng Yong Wei	**U2440621H**

Table of Contents

Executive Summary

Diabetes is one of the fastest-growing chronic diseases worldwide, affecting over 422 million people according to the World Health Organization (2018). Because of its long asymptomatic phase, early detection is crucial for effective management and prevention, yet an estimated 50% of people with diabetes remain undiagnosed. In Singapore, where one in eleven individuals is affected, diabetes poses a growing challenge amid an ageing population. Aligned with the national *Healthier SG* initiative, which focuses on shifting healthcare from reactive treatment to proactive prevention, this project aims to develop a data-driven model to predict diabetes risk using lifestyle and health indicators from the CDC Diabetes Health Indicators dataset.

Our variable analysis showed that all selected predictors were relevant indicators for diabetes prediction. The correlation matrix confirmed the absence of multicollinearity, while external research supported the inclusion of certain predictors even if they were not statistically significant within our dataset. Initially, the dataset exhibited a class imbalance, with 86.1% of cases being non-diabetic (0) and only 13.9% prediabetic or diabetic (1). This imbalance caused both the logistic regression and CART models to produce high false negatives and low true positives, i.e., poor sensitivity.

After rebalancing the data to achieve a 50–50 class distribution, sensitivity improved substantially. For instance, in the logistic regression model, sensitivity increased by 392.85%—from 0.1553 to 0.7654. We further enhanced the model by optimizing the decision threshold from 0.5 to 0.35, which boosted sensitivity from 76.54% to 89.16%, while accuracy decreased only slightly from 74.91% to 74.01%. This trade-off was worthwhile given the priority of identifying at-risk individuals.

Meanwhile, the CART model was refined through cost complexity pruning (CP) optimization. However, after comparison, the logistic regression model with a threshold of 0.35 demonstrated more stable and superior performance metrics across evaluations. Therefore, we selected this model as the optimal choice for real-time diabetes prediction.

Finally, all the variables form the foundation of our risk-scoring algorithm, which converts individual patient data into a probability-based score ranging from 0 to 100, providing an accessible way for users to assess their likelihood of developing diabetes. Integrated into the LifeSG app, this model empowers residents to self-assess their risk and receive personalized lifestyle recommendations. This analytics solution supports Healthier SG's goal of preventive, community-based healthcare through early detection and convenient engagement with the public.

1. Business Problem

Despite extensive national health campaigns and regular screenings offered by polyclinics and hospitals, diabetes continues to persist in Singapore, largely due to late detection and unhealthy lifestyle choices. According to the Ministry of Health (2023), about 9.5% of adults aged 18 to 69 have been diagnosed with diabetes. The International Diabetes Federation (2024) further estimates that approximately 699,100 adults in Singapore, roughly 11.4% of the population, are living with the condition.

Prevalence is particularly high among older adults, exceeding 20% for those aged 60 to 74. Furthermore, one in three Singaporeans is projected to develop diabetes in their lifetime if current trends persist. Despite these figures, many Singaporeans remain unaware of their personal risk until complications arise. This highlights a pressing need for a simple, accessible, and data-driven tool that enables individuals to assess their potential risk of developing diabetes early on.

2. Objective

To develop a predictive diabetes risk calculator integrated within the LifeSG app that enables Singaporeans to:

a. Input basic health and lifestyle information such as blood pressure, cholesterol, BMI, diet, age, and physical activity levels.
b. Obtain a personalized diabetes risk score derived from a data-driven model.
c. Receive tailored health advice based on their input values to encourage risk reduction and healthier choices.

3. Dataset

We decided on the “CDC Diabetes Health Indicators” dataset from Kaggle, which originally contained 253,680 survey responses and 21 health-related features such as HighBP, HighChol, BMI, Age, GenHlth (general health perception), PhysActivity, Fruits, Veggies, and HvyAlcoholConsump. These factors align closely with Healthier SG’s focus on lifestyle modification and chronic disease prevention. Studies such as Zhou et al., (2024) have shown that diabetic patients also tend to have a mix of the variables mentioned, where our model will try to predict the risk as accurately as possible .

This dataset utilises “0”s for individuals with no diabetes and “1”s for prediabetic or diabetic. Choosing to categorise prediabetic and diabetic together is important as it helps our model focus on assessing the risk of developing diabetes rather than

providing a conclusive medical diagnosis. Our goal is to help potentially diabetic users identify risk levels early, while also ensuring that consulting a doctor or undergoing formal clinic testing is recommended for confirmation and management.

One challenge that could be faced from this dataset was that more than 86.1% of the dataset contained data of individuals who had no diabetes (0) with the remaining 13.9% were people who were prediabetic/diabetic (1). This was a concern because class imbalance makes the model of poor quality during training, and as we have expected, there were more false negatives from our logistic regression and CART models. To address this, we used a balanced dataset instead, titled *diabetes_binary_5050split_health_indicators_BRFSS2015.csv*, which included 70,692 survey responses from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS). This version maintains a 50-50 distribution between prediabetic/diabetic (1) and non-diabetic (0) respondents across 21 health-related features, ensuring reliable model training and evaluation.

4. Analytics Solution

4.1 Data Cleaning and Libraries Used

The dataset was cleaned and prepared by converting categorical variables into factors, addressing missing or inconsistent values, and verifying that all data types were suitable for analysis. Core R packages such as data.table, ggplot2, caTools, and rpart were used for data handling, visualization, data splitting, and building models.

4.2 Variable Analysis: Importance and Statistical Testing

To explore variable importance, both statistical tests and visual techniques were used:

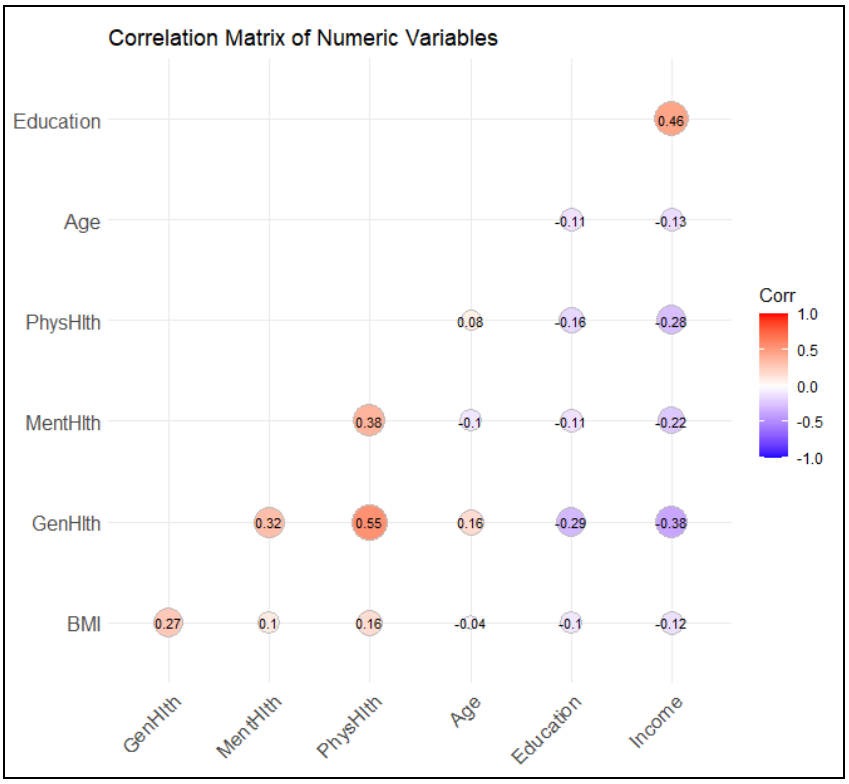


Figure 1: Heatmap of Correlation Matrix

The heatmap of the correlation matrix of only numeric variables using Pearson's correlation coefficients confirmed that no multicollinearity issues existed, as all the predictors' correlations amongst one another were below 0.60.

For categorical data, Chi-square tests identified High Blood Pressure, High Cholesterol, and Difficulty Walking as key predictors. T-tests for continuous variables further showed that all factors with p-values below 0.001 were highly significant, suggesting that all 7 continuous variables are significant. These small p-values suggest a statistically significant relationship between these factors and the likelihood of a diabetes diagnosis.

To illustrate these relationships more clearly, visualizations of the top predictors were plotted to highlight their influence on diabetes risk.

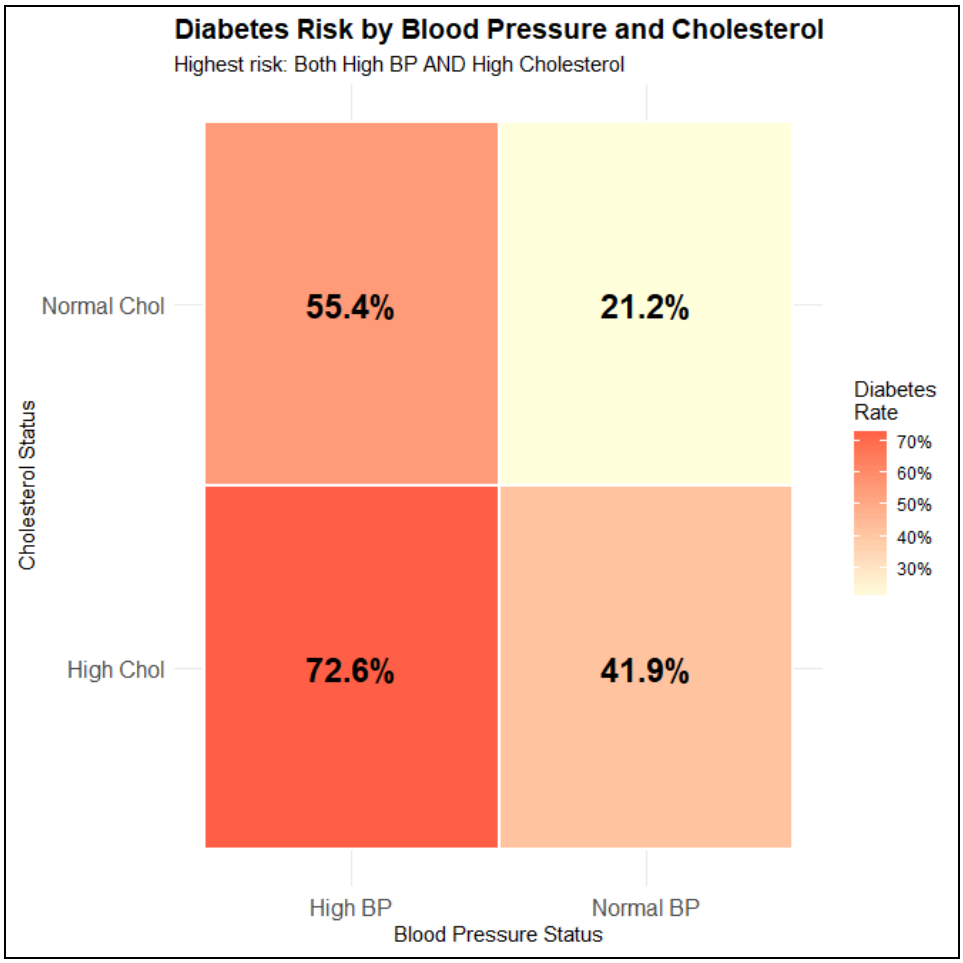


Figure 2: Heatmap of Diabetes Risk by Blood Pressure and Cholesterol

The heatmap shows that people who have both high blood pressure and high cholesterol face the greatest risk of developing diabetes. Those with normal blood pressure or cholesterol have a much lower risk, highlighting how these two conditions work together to significantly increase the chance of diabetes. This suggests both blood pressure and cholesterol as strong predictors of diabetes diagnosis. Supporting this, research by Northwest Integrative Medicine (2014) also showed that high blood pressure and high cholesterol often occur together as part of the metabolic syndrome, a combination of conditions that can lead to diabetes. High blood pressure and diabetes share similar biological roots, such as insulin resistance and inflammation in the body. Likewise, high cholesterol can disrupt normal metabolism, making it easier for diabetes to develop.

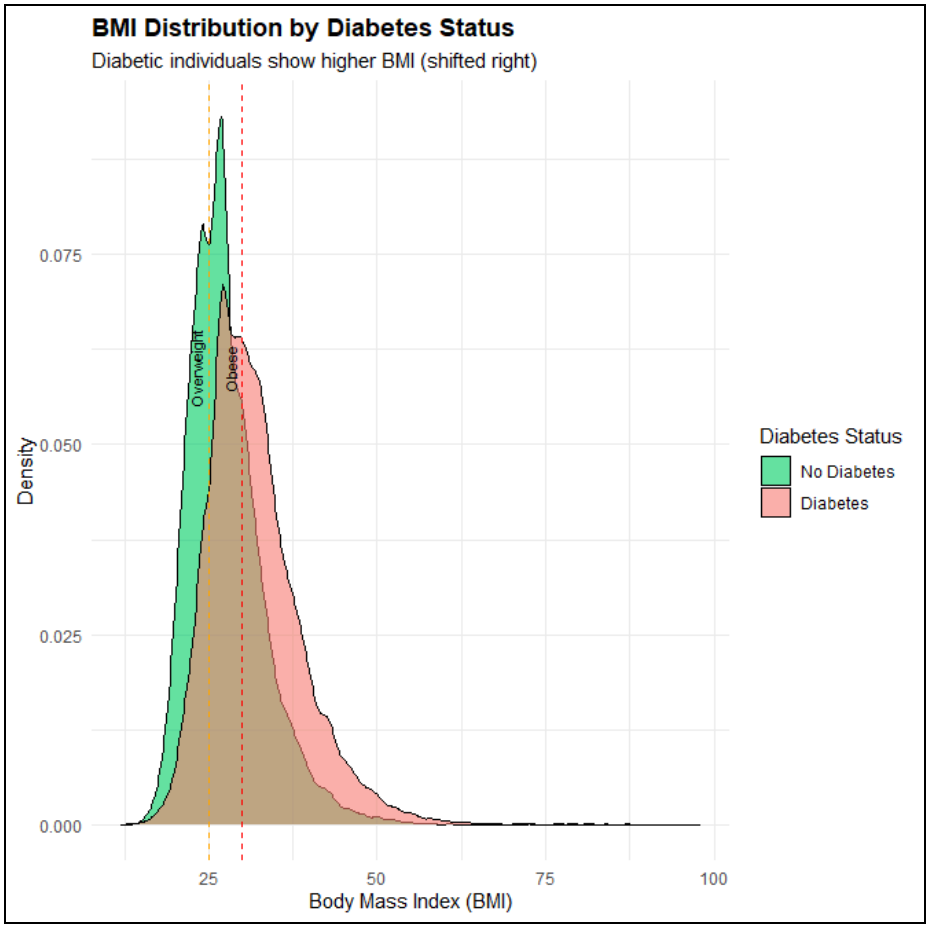


Figure 3: Diabetes Against BMI Distribution

The BMI density plot shows that individuals with diabetes generally have higher BMIs than those without, with a noticeable rightward shift in the diabetic group. This highlights the strong association between excess body weight and diabetes risk. Overweight (BMI 25–29.9) and obese (BMI ≥30) individuals face a significantly higher likelihood of developing diabetes, as excess body fat reduces the body's ability to use insulin effectively and strains the insulin-producing cells. Numerous studies, including Arshad Mohamed Channanath et al. (2014), have consistently validated obesity—reflected by BMI—as one of the strongest modifiable predictors of type 2 diabetes, linking it to insulin resistance and beta cell dysfunction across populations.

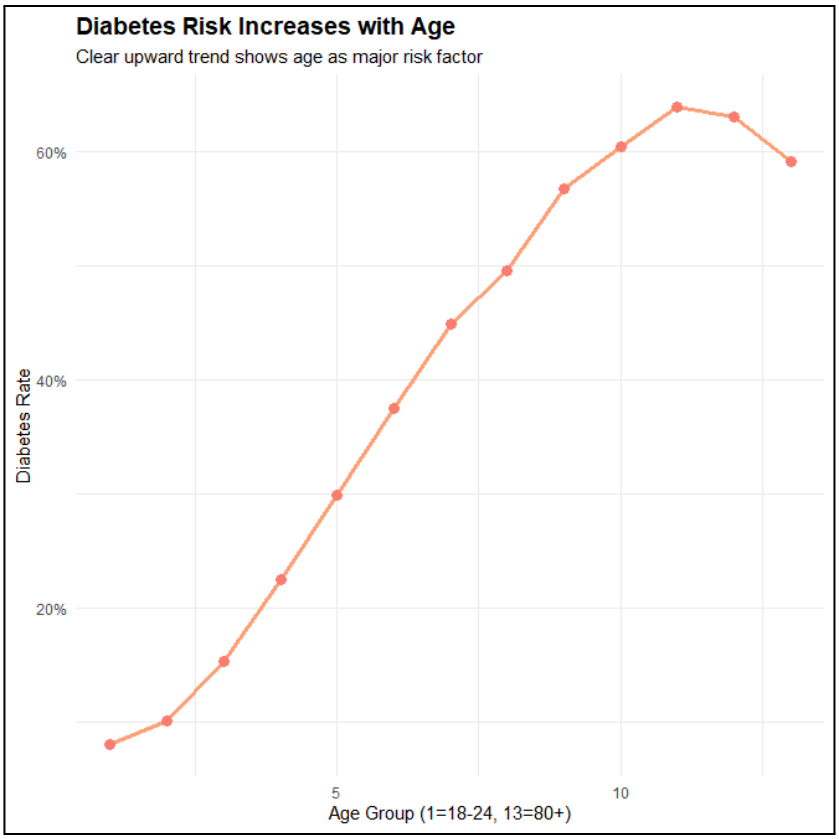


Figure 4: Diabetes Rate Against Age

The age groups on the x-axis are divided into 5-year intervals. For instance, age group 9 represents individuals aged 60–64, while age group 10 corresponds to those aged 65–69. The only exceptions are group 1, which includes all individuals aged 18–24, and group 13, which represents those aged 80 and above. The line plot shows that diabetes rates rise steadily with age, confirming that age is a major risk factor. As people get older, their metabolism slows down, the body becomes less responsive to insulin, and the cells that produce it start to wear out. However, after the rate peaks in older adults, it then dips slightly in the oldest age group, a pattern likely due to survival bias, competing mortality, and underdiagnosis. People with diabetes are more likely to die younger from complications like heart and kidney disease, so fewer survive into advanced age. Furthermore, among the elderly, only the diabetes cases that are healthier or better managed tend to persist. Additionally, diabetes may be underdiagnosed in older adults because symptoms can be mistaken for normal aging, and screenings are less common, which can hide the true number of cases in this group.

Beyond the primary visualisations, other variables also emerged as relevant predictors.

Self-rated general health (GenHlth) turned out to be a strong indicator of diabetes risk as validated by Moa Lugner et al. (2024). People who described their overall health as poor were more likely to have or develop diabetes. This finding matches Moa Lugner et al. (2024) research showing that how people perceive their health often reflects both diagnosed conditions and underlying issues not yet identified.

Difficulty walking (DiffWalk) or climbing stairs, while subjective, can reveal important signs of deterioration in health. It may indicate nerve damage, reduced fitness, or complications linked to diabetes. Although it may not predict the start of diabetes on its own, it remains an important marker for identifying people with more advanced disease or mobility challenges.

Physical activity (PhysActivity) remains a key predictive factor for diabetes risk. Individuals who are less physically active face a higher likelihood of developing diabetes, as inactivity contributes to weight gain and reduces the body's sensitivity to insulin. Consistent with existing research, So Hyun Cho et al. (2025) found that lack of regular exercise significantly increases diabetes risk by promoting obesity and insulin resistance.

Smoking (Smoker) is also a significant contributor to diabetes risk. It elevates inflammation and oxidative stress in the body, disrupting metabolic balance even among individuals who are not overweight or inactive. Independent of BMI and physical activity, smoking has been shown to increase the likelihood of developing type 2 diabetes through these inflammation and oxidative stress pathways (K. Patja et al., 2005).

Sociodemographic factors such as education and income also play an important role in diabetes risk. Individuals with lower socioeconomic status often face greater barriers to accessing healthcare, affording nutritious foods, and sustaining healthy lifestyle habits. These disparities contribute to the higher prevalence of diabetes among lower-income and less-educated populations. As noted by Better (2023), lower socioeconomic status and education levels are strongly associated with increased diabetes risk due to limited healthcare access, lower health literacy, and less healthy dietary and lifestyle behaviors.

Sociodemographic factors such as education and income also play an important role in diabetes risk. Individuals with lower socioeconomic status often face greater barriers to accessing healthcare, affording nutritious foods, and sustaining healthy lifestyle habits. These disparities contribute to the higher prevalence of diabetes among lower-income and less-educated populations. As noted by Better (2023), lower socioeconomic status and education levels are strongly associated with increased diabetes risk due to limited healthcare access, lower health literacy, and less healthy dietary and lifestyle behaviors.

Mental and physical health (MentHlth, PhysHlth) are closely linked to diabetes risk. Individuals experiencing poor mental well-being or frequent physical illnesses are often less able to maintain healthy routines and may experience heightened stress, which disrupts blood sugar regulation. As noted by Science Direct (2025), poor mental and physical health can indirectly increase diabetes risk by impairing self-care, elevating stress levels, and undermining metabolic control.

Access to healthcare (AnyHealthcare, NoDocbcCost) influences diabetes management more than its initial development. People who lack regular medical access or avoid care due to cost are often diagnosed later, allowing complications to progress.

Finally, diet and alcohol consumption (Fruits, Veggies, HvyAlcoholConsump) also influence diabetes risk, though their effects are less pronounced when considered in isolation. Consuming more fruits and vegetables and limiting heavy alcohol intake can help reduce the likelihood of developing diabetes, particularly when combined with other healthy habits. According to the American Diabetes Association (n.d.), these modifiable lifestyle factors play a meaningful role in diabetes prevention, though their individual predictive power is generally lower than that of biometric indicators such as BMI and blood pressure.

Overall, this analysis highlights a complex combination of biomedical, lifestyle, and social factors shaping diabetes risk. Blood pressure, cholesterol, weight, age, and general health remain the most powerful predictors, supported by evidence from research. Meanwhile, variables like physical activity, mobility, diet, and smoking offer valuable insight into how everyday choices and health perceptions influence the development and management of diabetes.

4.3 Logistic Regression Model

A Logistic Regression model was developed using predictors selected through correlation and Chi-square tests. After training, model coefficients were analysed to assess each variable's effect on diabetes risk, and predictions were evaluated using a confusion matrix.

Using the default threshold of 0.5, the model achieved an accuracy of 74.91% with a sensitivity of 76.54%. However, the model still produced 2488 false negatives, indicating that approximately 23.46% of diabetic cases were missed. Thus, further optimisation was needed to reduce false negatives, as it is critical in a healthcare context where missing a positive case has serious consequences.

To enhance sensitivity, threshold optimization was performed by testing values from 0.1 to 0.9. The optimal threshold of 0.35 maximised the F1 score, achieving 89.15%

sensitivity while maintaining reasonable overall performance. This reduced false negatives to 1151 cases, a 46.26% reduction, making the model more suitable for early diabetes screening.

4.4 CART Model

A CART model was developed to capture non-linear relationships and variable interactions that logistic regression might miss. The maximal tree was pruned using the 1-standard-error rule based on cross-validation results (optimal CP = 0.000475) to prevent overfitting while maintaining predictive power.

The pruned CART model demonstrated strong performance on the balanced test set, producing 2439 false negatives, outperforming the default logistic regression model (FN = 2488) but not reaching the optimised threshold logistic regression (FN = 1151). The decision tree's interpretability makes it valuable for clinical communication, as healthcare providers can easily explain the prediction logic through simple if-then rules. Variable importance analysis from CART confirmed General Health, High BP, and BMI as the top predictors, validating findings from statistical tests.

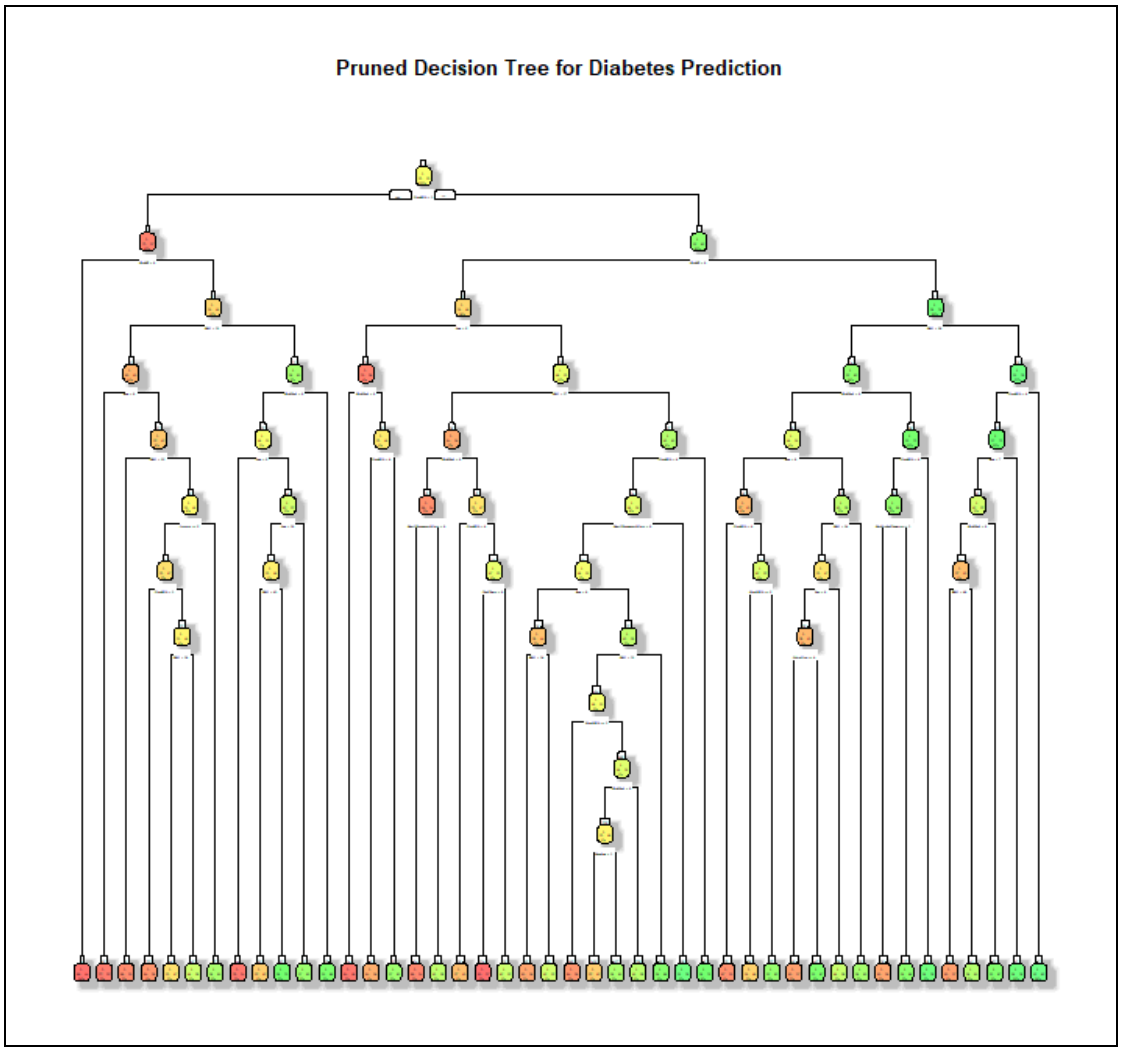


Figure 5: Pruned Decision Tree

4.5 Performance Metrics

Models were evaluated with the following metrics:

- **Accuracy**: The overall correctness of the model
- **Sensitivity** (Recall): The ability to correctly identify diabetic cases
- **Specificity**: Correctly identifying non-diabetic cases
- **F1 Score** (harmonic mean of precision and recall): Reflection of class imbalance impact
- **AUC** (Area Under ROC Curve): Overall discriminatory power of the model
- **FN** (False Negatives): Cases where diabetic individuals are incorrectly predicted as non-diabetic.

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC	FN
Logistic Regression (default threshold 0.5)	**0.7491**	0.7654	**0.7327**	0.7531	**0.8248**	2488
Logistic Regression (optimal threshold 0.35)	0.7401	**0.8915**	0.5886	**0.7742**	**0.8248**	**1151**
CART (pruned)	0.7444	0.7700	0.7189	0.7508	0.7987	2439

Logistic regression, with an optimised threshold of 0.35 instead of the default 0.5, significantly improved sensitivity and reduced false negatives (FN = 1151), in expense of a slight decrease in accuracy. It resulted in the highest F1 score, demonstrating a better balance between precision and recall, which is crucial for imbalanced health datasets.

4.6 Model Comparison

Logistic Regression with optimised threshold (0.35) achieved the highest F1 score (77.4%) and best AUC (0.8248), demonstrating superior overall balance between precision and recall with the strongest discriminative ability. It also achieved the highest sensitivity (89.15%), making it most effective at catching diabetes cases. By reducing false negatives by 46.26%, this approach minimises missed diagnoses, which is critical in preventive healthcare where early intervention significantly improves outcomes.

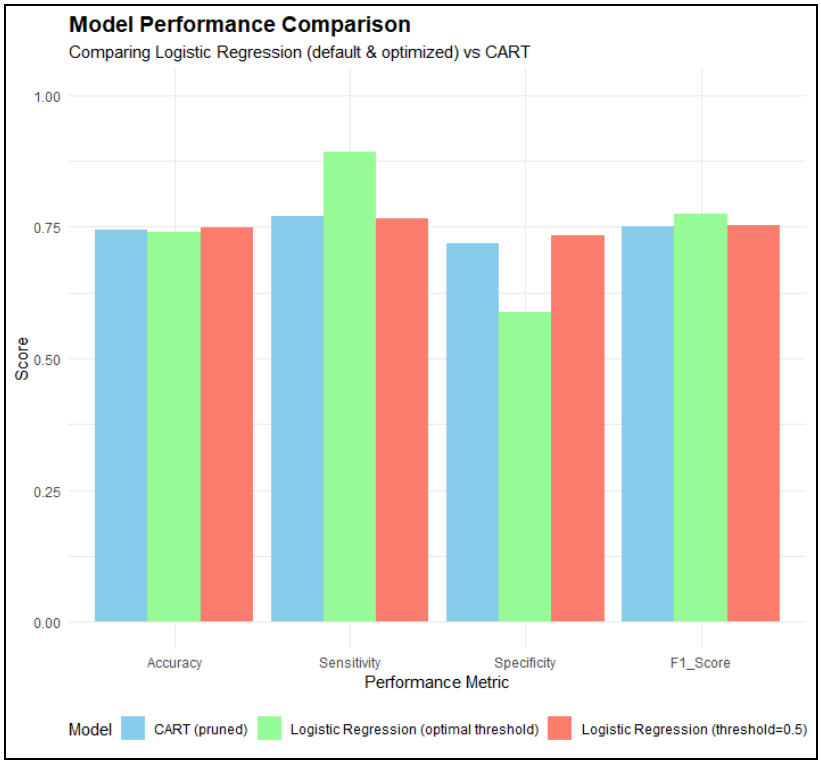


Figure 6: Model Performance Comparison Bar Chart

4.7 ROC Curves Visualization

ROC curves compare the trade-off relationship between true positive and false positive rates at various thresholds. Logistic regression (AUC = 0.8248) outperforms CART (AUC = 0.7987), with both models side by side for comparison.

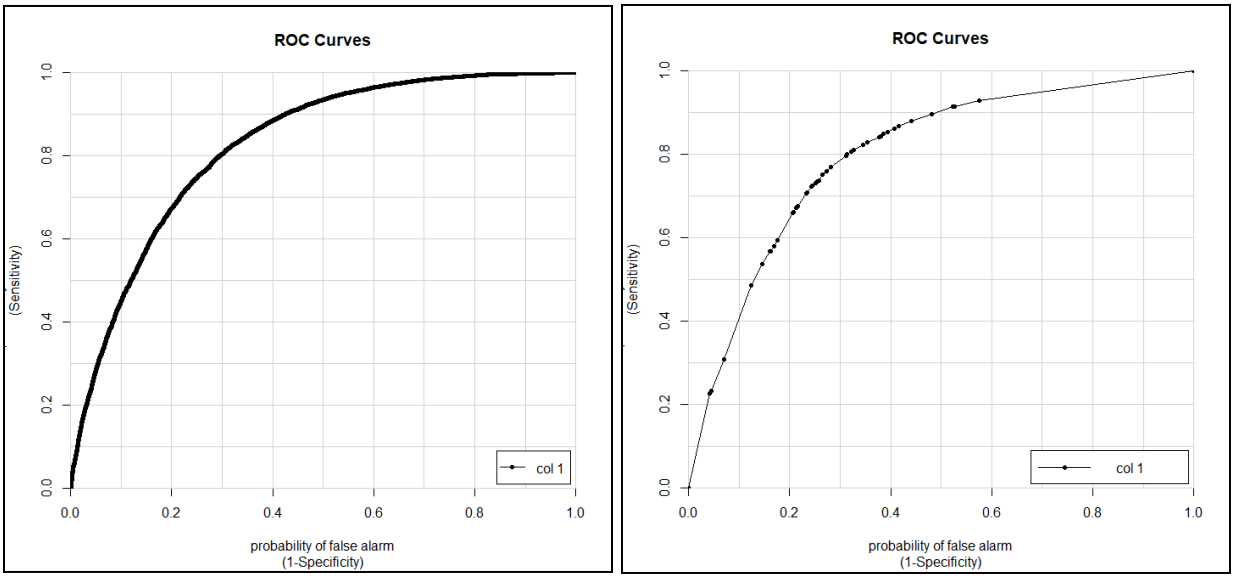


Figure 8 & 9: Logistic Regression's & CART's ROC Curves

4.8 Facet Charts (Multi-Dimensional Visualisation)

Facet charts visualise the interaction between risk factors, such as Age vs BMI, with stratification by HighBP and HighChol status. These provide rich, multi-dimensional views of how high-risk subgroups cluster, guiding targeted interventions.

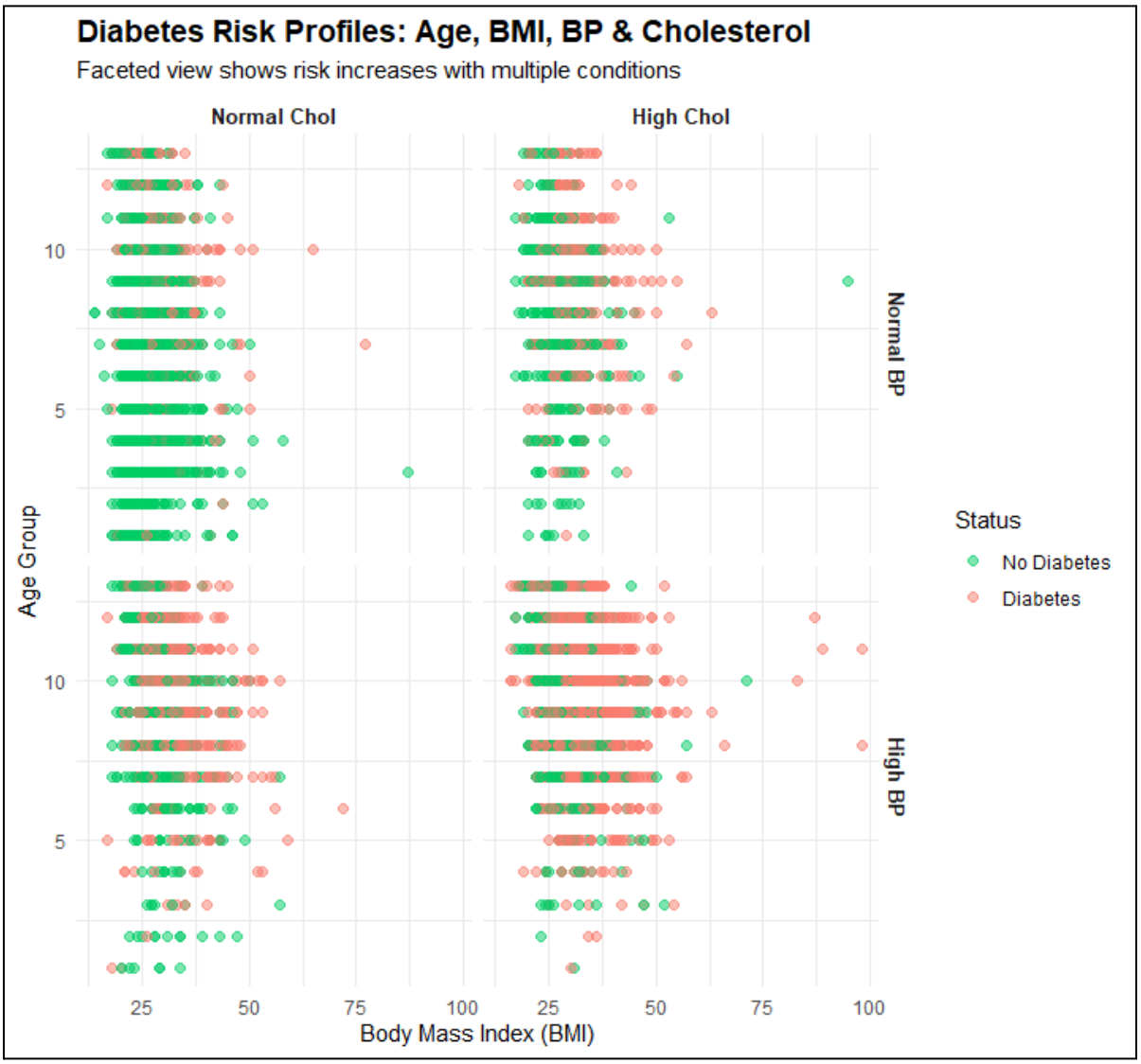


Figure 7: Facet Chart

4.9 Coefficient Extraction and Risk Scoring

To create a practical screening tool, we extracted coefficients from the trained logistic regression model and developed a risk scoring function that calculates personalised diabetes risk scores on a 0-100 scale.

The model coefficients represent each predictor's contribution to diabetes risk in log-odds terms. We calculated odds ratios by exponentiating the coefficients to quantify the relative impact of each factor. Variables with odds ratios greater than 1 increase diabetes risk, while those below 1 are protective.

The risk scoring function implements the logistic regression equation:

$$\text{Risk Score} = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_i x_i)}} \times 100$$

Figure 8: Risk Scoring Function

This three-step process calculates log-odds, converts to probability using the logistic function, and scales to 0-100 for intuitive interpretation.

We applied this function to the test dataset (n = 21208), generating risk scores for each individual. The score distribution showed clear separation between diabetic and non-diabetic cases, validating the calculator's discriminative ability.

5. Implementation Concept

This model can be integrated into LifeSG under the Health & Wellness section. This integration introduces a new and improved ecosystem where government applications such as LifeSG, HealthHub, and Healthy 365 would be able to sync. Users can enter self-reported data or sync with HealthHub/Healthy365 devices. This would ensure a seamless data flow among stakeholders such as individuals, healthcare providers, and the national health system.

Under the existing ecosystem, government health applications are not synced, risking data overlaps. In the new ecosystem:

- **LifeSG App**: The centralised portal where our model will run. Individuals will fill in information on their demographic and health according to the prompts, and a

summary of their diabetes risk assessment will be generated immediately by our model

- **HealthHub / Healthy 365 Apps**: The results of the risk assessment on the LifeSG app would be synced to Healthier SG apps. Individuals may also choose to allow these apps to provide real-time health data to the LifeSG app.
- **Hospitals and clinic databases**: These clinics would be able to receive and access summarised information on patients' risk assessments. High-risk individuals would be flagged, and healthcare professionals would be able to decide on the next course of action.

The immediate outputs from our model running the diabetes risk assessment are:

- **Personalised risk score**: a visual indicator of an individual's risk assessment score. It will be presented through traffic light colours, indicating Low, Medium, and High risk.
- **Recommended actions**: Individuals would receive personalised recommendations in accordance with Healthier SG goals. This includes recommendations for dietary and physical activity enhancements, as well as suggested health screenings at the clinic.
- **Collaboration**: Data sharing with family doctors for proactive follow-up, especially for high-risk individuals.

6. Techniques Used

We have applied a combination of techniques to predict diabetes and translate the data into actionable insights for end-users.

1. **Logistic regression:** This was used to model how probable a diabetes diagnosis is (diabetes_binary = Y) based on variables like lifestyle, demographics, and health profile. Logistic regression was used as our target was a binary variable. Using the 70-30 train-test split, we trained the model on our cleaned dataset, thus quantifying the effect of each variable on the probability of developing diabetes. Through this technique, we identified the significant predictors in identifying diabetes.
2. **Feature Selection:** Based on model coefficients and variable significance, we identified the variables that significantly influence the probability of a diabetes diagnosis and included them into the model. This helped to reduce the complexity of the model as only meaningful predictors were used. Our selected features were high blood pressure, high cholesterol, BMI, general health and age.

3. **Risk Scoring Algorithm**: To make our output more user-friendly for the general public, we used a risk scoring algorithm to translate the regression output into an interpretable 0–100 point scale. Hence, all users are able to quickly interpret their risk level and their suggested preventive action.

7. Expected Outcomes

7.1 Technical Outcomes

We expect to have a validated diabetes risk model, with the logistic regression model achieving >74% accuracy. This will ensure our model is perceived as a reliable tool in identifying and classifying individuals based on the likelihood of them developing diabetes. A risk-scoring system will be deployed via LifeSG based on the diabetes risk model. Following the results, actionable insights regarding diet and lifestyle habits will be displayed as key preventive levers.

These outcomes directly contribute to our goal of lowering the diabetes rate in Singapore by having a predictive gauge and guiding individuals on preventive measures they can take.

7.2 Social Outcomes

Data gathered from the risk assessment will be captured and stored, and population-level insights will be used by health authorities such as the Ministry of Health (MOH) to run Healthier SG campaigns. This will lead to improved public awareness and early intervention rates under Healthier SG.

8. Conclusion

Our diabetes risk scoring model demonstrates how data analytics has transformed many aspects of our lives, and in this case, preventive healthcare. By making use of lifestyle and biometric indicators, like blood pressure, cholesterol, BMI, diet, etc., we created a model that helps to convert complex statistical output into a clear and user-friendly risk score. This helps to bridge the gap between individual measures and crowd research, giving Singaporeans the chance to understand their health status better and take proactive and preventive actions.

Since it will be integrated into the LifeSG app, the predictor acts as a “digital health companion” for Singaporeans, where data from their everyday lives can be conveniently input or synced from their HealthHub or Healthy 365 to get a diabetic risk score right

from their phones. This appropriately aligns with Healthier SG's preventive focus to build and maintain engagement between Singaporeans, GPs, and community health programmes. Looking at this from a systemic standpoint, combined random risk data can be collected to help identify emerging health concern hotspots, revise health campaigns, and dedicate resources more accurately. For GPs, this model can provide a reliable and data-centred reference point to bolster clinical conversations and prompt lifestyle alterations.

Finally, our model and project shows how predictive data analysis can assist in Singapore's transition from reactive treatment to proactive prevention. With the implementation of real-world data and understandable risk thresholds (high, medium, low risk), it will help Healthier SG to improve early detection, motivate healthier lifestyles and decelerate the national increment in chronic disease. Therefore, shifting Singapore closer to a more sustainable and citizen-centred healthcare community.

References

TEBOUL, A. (2022). Diabetes Health Indicators Dataset. www.kaggle.com. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Northwest Integrative Medicine. (2024, July 18). *Connecting Blood Sugar, Cholesterol & Blood Pressure - Northwest Integrative Medicine*. https://nwim.org/blood-sugar-cholesterol-blood-pressure/

Cigolle, C. T., Blaum, C. S., Lyu, C., Ha, J., Kabeto, M., & Zhong, J. (2022). Associations of Age at Diagnosis and Duration of Diabetes With Morbidity and Mortality Among Older Adults. *JAMA Network Open*, *5*(9), e2232766. https://doi.org/10.1001/jamanetworkopen.2022.32766

Kaptoge, S., Seshasai, S., Sun, L., Walker, M., Bolton, T., Spackman, S., Ataklte, F., Willeit, P., Bell, S., Burgess, S., Pennells, L., Altay, S., Assmann, G., Ben-Shlomo, Y., Best, L., Björkelund, C., Blazer, D., Brenner, H., Brunner, E., & Dagenais, G. (2023). Life expectancy associated with different ages at diagnosis of type 2 diabetes in high-income countries: 23 million person-years of observation. *The Lancet Diabetes & Endocrinology*, *11*(10), 731–742. https://doi.org/10.1016/S2213-8587(23)00223-1

Why Diabetes Risk Increases After 60 and How to Prevent It. (2025). Wellness We Care. https://www.megawecare.com/wellness-we-care/healthy-aging/risk-of-diabetes-after-60

Lugner, M., Rawshani, A., Helleryd, E., & Eliasson, B. (2024). Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-52023-5

Channanath, A. M., Farran, B., Behbehani, K., & Thanaraj, T. A. (2014). Impact of Hypertension on the Association of BMI with Risk and Age at Onset of Type 2 Diabetes Mellitus: Age- and Gender-Mediated Modifications. *PLoS ONE*, *9*(4), e95308. https://doi.org/10.1371/journal.pone.0095308

Hyun, C. S., Kim, G., Lee, K., Oh, R., Yoon, K. J., Jang, M., Lee, Y.-B., Jin, S.-M., Yeon, H. K., Han, K., & Hyeon, K. J. (2025). Impact of smoking and physical activity on cardiovascular outcomes in type 2 diabetes with metabolic dysfunction-associated steatotic liver disease: a nationwide study. *Endocrine Abstracts*. https://doi.org/10.1530/endoabs.110.ep373

PATJA, K., JOUSILAHTI, P., HU, G., VALLE, T., QIAO, Q., & TUOMILEHTO, J. (2005). Effects of smoking, obesity and physical activity on the risk of type 2 diabetes in middle-aged Finnish men and women. *Journal of Internal Medicine*, *258*(4), 356–362. https://doi.org/10.1111/j.1365-2796.2005.01545.x

BETTER. (2023, June 15). *How does Socioeconomic Status Impact the Risk of Type 1 Diabetes-Related Complications? – BETTER*. BETTER. https://type1better.com/en/how-does-socioeconomic-status-impact-the-risk-of-type-1-diabetes-related-complications/

Qasrawi, R., Thwib, S., Issa, G., Abu Ghoush, R., & Amro, M. (2025). Type 2 Diabetes Risk Prediction Using Glycemic Control Metrics: A Machine Learning Approach. *Human Nutrition & Metabolism*, 200341. https://doi.org/10.1016/j.hnm.2025.200341

American Diabetes Association. (2024). *Alcohol & Diabetes | ADA*. Diabetes.org. https://diabetes.org/health-wellness/alcohol-and-diabetes

Appendix

Declaration of Academic Integrity

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

Please insert an "X" within the square bracket below to indicate your selection.

[X] **I have read and accept the above.**

Declaration on Use of GenAI

1. Name of course: Analytics I: Visual and Predictive Analytics

2. Course Code: BC2406

3. Title of Assignment/Project Submission: Team Project

In relation to the foregoing, I hereby declare that fully and properly in accordance with the Assignment/Project Instructions, I have **(insert an "X" in the relevant square bracket)**:

i. Used GenAI as permitted to assist in generating key ideas. **[ ]**

ii. Used GenAI as permitted to assist in generating a first text. **[ ]**

And/Or

iii. Used GenAI to refine syntax and grammar for correct language submission. **[X]**

Or

iv. Did not use GenAI in any way. **[ ]**

I also declare that I have:

a. Fully and honestly submitted the digital paper trail required under the assignment/project instructions in the appendix of this document; and that

b. Wherever GenAI assistance has been employed in the submission in word or paraphrase or inclusion of a significant idea or fact suggested by the GenAI, I have acknowledged this by a footnote or in-text reference; and that,

c. Apart from the foregoing notices, the submission is wholly my own work

Aloysious Law Jia Jian

Bryan Lim How Meng

De Souza Alyssa Anne

Ng Yong Wei 1/11/2025

……. ……..

Student Name and Signature Date

Model Findings

Logistic Regression (with default threshold of 0.5)

Confusion matrix:

	Predicted Negative	Predicted Positive
Actual Negative	7770	2834
Actual Positive	2488	8116

Performance:

Accuracy	0.7491
Sensitivity	0.7654
Specificity	0.7327
Precision	0.7412
F1 Score	0.7531

Logistic Regression (with optimal threshold of 0.35)

Confusion matrix:

	Predicted Negative	Predicted Positive
Actual Negative	6242	4362
Actual Positive	1151	9453

Performance:

Accuracy	0.7401
Sensitivity	0.8915

Specificity	0.5886
Precision	0.6843
F1 Score	0.7742

CART (pruned)

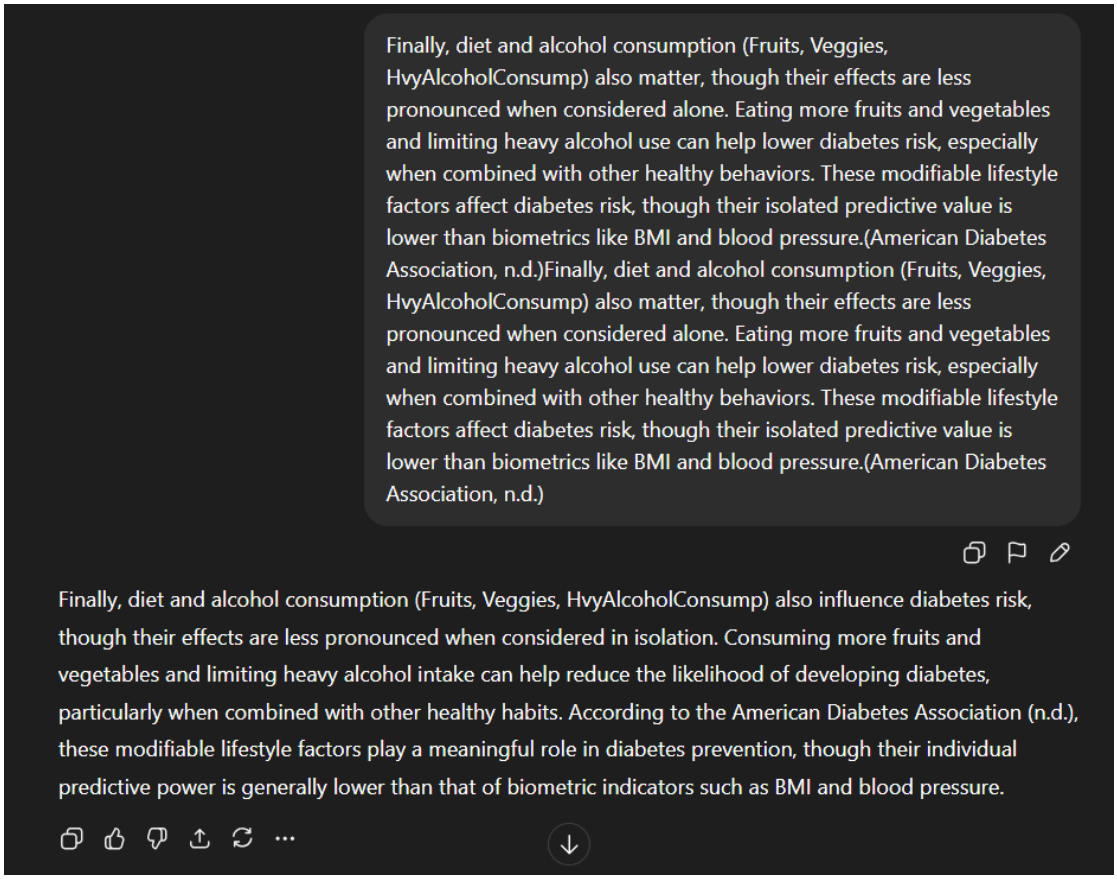
Confusion matrix:

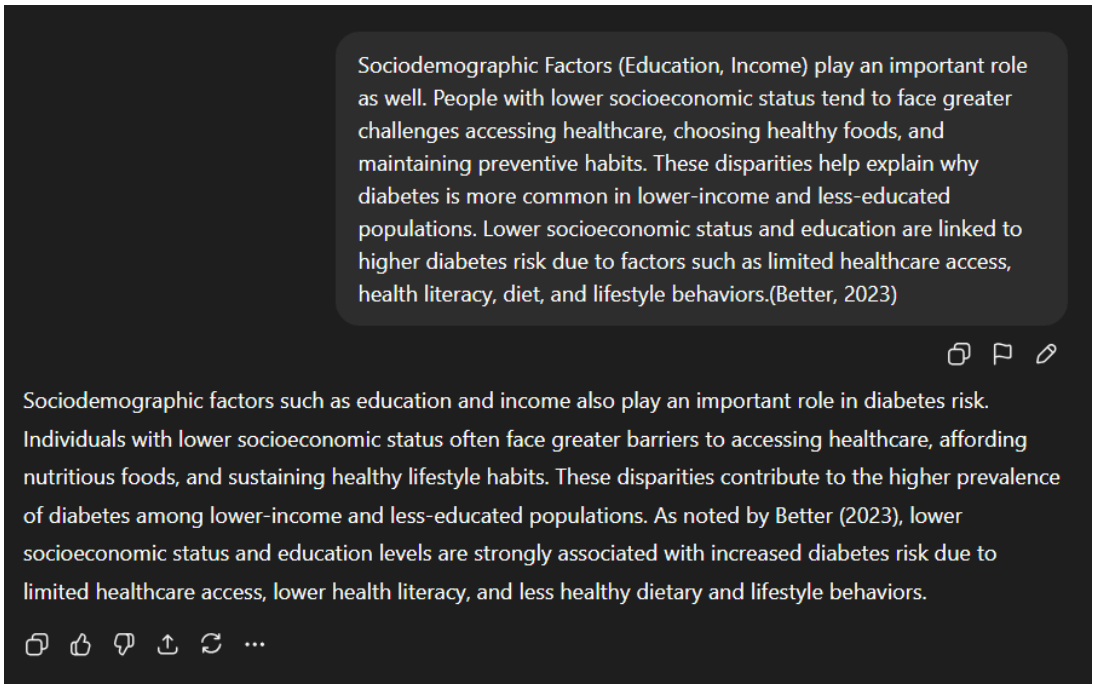
	Predicted Negative	**Predicted Positive**
Actual Negative	7623	2981
Actual Positive	2439	8165

Performance:

Accuracy	0.7444
Sensitivity	0.7700
Specificity	0.7189
Precision	0.7325
F1 Score	0.7508

Attached Prompts





Smoking (Smoker) also contributes meaningfully to diabetes risk. It increases inflammation and oxidative stress in the body, which can harm metabolic balance even in people who are not overweight or inactive.Independent of BMI and physical activity, smoking increases risk of type 2 diabetes through inflammation and oxidative stress pathways. (K Patja et al., 2005)

Smoking (Smoker) is also a significant contributor to diabetes risk. It elevates inflammation and oxidative stress in the body, disrupting metabolic balance even among individuals who are not overweight or inactive. Independent of BMI and physical activity, smoking has been shown to increase the likelihood of developing type 2 diabetes through these inflammation and oxidative stress pathways (K. Patja et al., 2005).

Physical activity (PhysActivity) continues to be a key predictive factor. People who are less active face a higher risk of diabetes because lack of exercise leads to weight gain and makes the body less responsive to insulin. Lack of regular physical activity is a well-documented risk factor increasing diabetes risk by promoting obesity and insulin resistance. (So Hyun Cho et al. 2025)

Physical activity (PhysActivity) remains a key predictive factor for diabetes risk. Individuals who are less physically active face a higher likelihood of developing diabetes, as inactivity contributes to weight gain and reduces the body's sensitivity to insulin. Consistent with existing research, So Hyun Cho et al. (2025) found that lack of regular exercise significantly increases diabetes risk by promoting obesity and insulin resistance.

The BMI density plot shows that people with diabetes generally have higher BMIs than those without. The rightward shift of the diabetic group highlights the strong link between excess body weight and diabetes risk. Overweight (BMI 25–29.9) and obese (BMI ≥30) individuals face a much higher risk of developing diabetes. Excess body weight makes it harder for the body to use insulin properly and puts extra pressure on the cells that produce it. Many studies have consistently shown that a higher BMI strongly predicts the likelihood of type 2 diabetes across populations.
Obesity, reflected by BMI, is consistently one of the strongest modifiable predictors of type 2 diabetes worldwide, causing insulin resistance and beta cell dysfunction, validated repeatedly in cohort and mechanistic studies. (Arshad Mohamed Channanath et al., 2014)

combine these two and remove duplicative content where necessary

The BMI density plot shows that individuals with diabetes generally have higher BMIs than those without, with a noticeable rightward shift in the diabetic group. This highlights the strong association between excess body weight and diabetes risk. Overweight (BMI 25–29.9) and obese (BMI ≥30) individuals face a significantly higher likelihood of developing diabetes, as excess body fat reduces the body's ability to use insulin effectively and strains the insulin-producing cells. Numerous studies, including Arshad Mohamed Channanath et al. (2014), have consistently validated obesity—reflected by BMI—as one of the strongest modifiable predictors of type 2 diabetes, linking it to insulin resistance and beta cell dysfunction across populations.

The age groups on the x-axis are in 5-year intervals, for example, age group 9 corresponds to ages 60–64, age group 10 corresponds to 65-69, with the exceptions of group 1 includes all ages 18–24 and group 13 is for ages 80 or older.

The age groups on the x-axis are divided into 5-year intervals. For instance, age group 9 represents individuals aged 60–64, while age group 10 corresponds to those aged 65–69. The only exceptions are group 1, which includes all individuals aged 18–24, and group 13, which represents those aged 80 and above.

Despite the national health campaigns and regular screenings offered by polyclinics and hospitals, diabetes still persists in Singapore mainly due to late detection and poor lifestyle choices. Around 9.5% of adults aged 18-69 are diagnosed with diabetes (MOH, 2023) with a large number of adults living with diabetes at a staggering 699,100 in Singapore alone in 2024, making up roughly 11.4% of the population (IDF, 2024). A majority of Singaporeans still remain oblivious to the personal risks before complications arise. Amongst the older age groups, prevalence is much higher, rising above 20% for those aged 60-74. In addition, one in three Singaporeans is estimated to have a lifetime risk of diabetes emergence if our trends remain stagnant. Despite this, we realised that Singaporeans lack an easy, accessible, and data analytical way to examine themselves for a potential diabetic risk.

polish this

Despite extensive national health campaigns and regular screenings offered by polyclinics and hospitals, diabetes continues to persist in Singapore, largely due to late detection and unhealthy lifestyle choices. According to the Ministry of Health (2023), about 9.5% of adults aged 18 to 69 have been diagnosed with diabetes. The International Diabetes Federation (2024) further estimates that approximately 699,100 adults in Singapore—roughly 11.4% of the population—are living with the condition.

Prevalence is particularly high among older adults, exceeding 20% for those aged 60 to 74. Alarmingly, one in three Singaporeans is projected to develop diabetes in their lifetime if current trends persist. Despite these figures, many Singaporeans remain unaware of their personal risk until complications arise. This highlights a pressing need for a simple, accessible, and data-driven tool that enables individuals to assess their potential risk of developing diabetes early on.

Variable analysis: all were relevant indicators for prediction of diabetes (correlation matrix doesnt show multicollinearity, external research shows relevance for predictors even though they were not statistically significant)
Changed datasets because of class imbalance (class imbalance (86.1% 0s and 13.9% 1s) made the models to have high false negatives and low true positives → low sensitivity, with balanced classes (50-50 split), sensitivity increased by 392.85% from 0.1553 to 0.7654.
Optimizing threshold from 0.5 to 0.35, sensitivity was from 76.54% to 89.16% but accuracy dropped slightly from 74.91% to 74.01% which is worth the trade
Our technique includes using logistic regression to create a model that identifies key health and lifestyle factors in order to translate them into a personalized risk score that can be easily interpreted by the general public, who are end-users of our model. Through our logistic regression model, we deduced that the top predictors, variables that have a statistically significant relationship with a positive diabetes diagnosis, are high blood pressure, high cholesterol, BMI, general health, and age. We then created a risk-scoring algorithm that converts patient data into a probability-based score (0–100).

combine these two together to from paragraphs

Our variable analysis revealed that all selected features were relevant indicators for predicting diabetes. The correlation matrix confirmed the absence of multicollinearity, while external research supported the inclusion of certain predictors even if they were not statistically significant within our dataset. Initially, our dataset suffered from class imbalance, with 86.1% of cases being non-diabetic (0) and only 13.9% diabetic (1). This imbalance led to models producing high false negatives and low true positives, resulting in poor sensitivity. After rebalancing the data to a 50–50 split, sensitivity improved dramatically by 392.85%, increasing from 0.1553 to 0.7654.

To further enhance model performance, we optimized the decision threshold from 0.5 to 0.35. This adjustment increased sensitivity from 76.54% to 89.16%, while accuracy experienced only a minor decline from 74.91% to 74.01%—a worthwhile trade-off given the importance of identifying at-risk individuals. Our technique leverages logistic regression to identify key health and lifestyle factors, translating them into a personalized risk score easily understood by the general public. From our model, we identified high blood pressure, high cholesterol, BMI, general health, and age as the most significant predictors of diabetes. These variables form the foundation of our risk-scoring algorithm, which converts individual patient data into a probability-based score ranging from 0 to 100, providing an accessible way for users to assess their likelihood of developing diabetes.