HEALTHIER SG

DiaScope

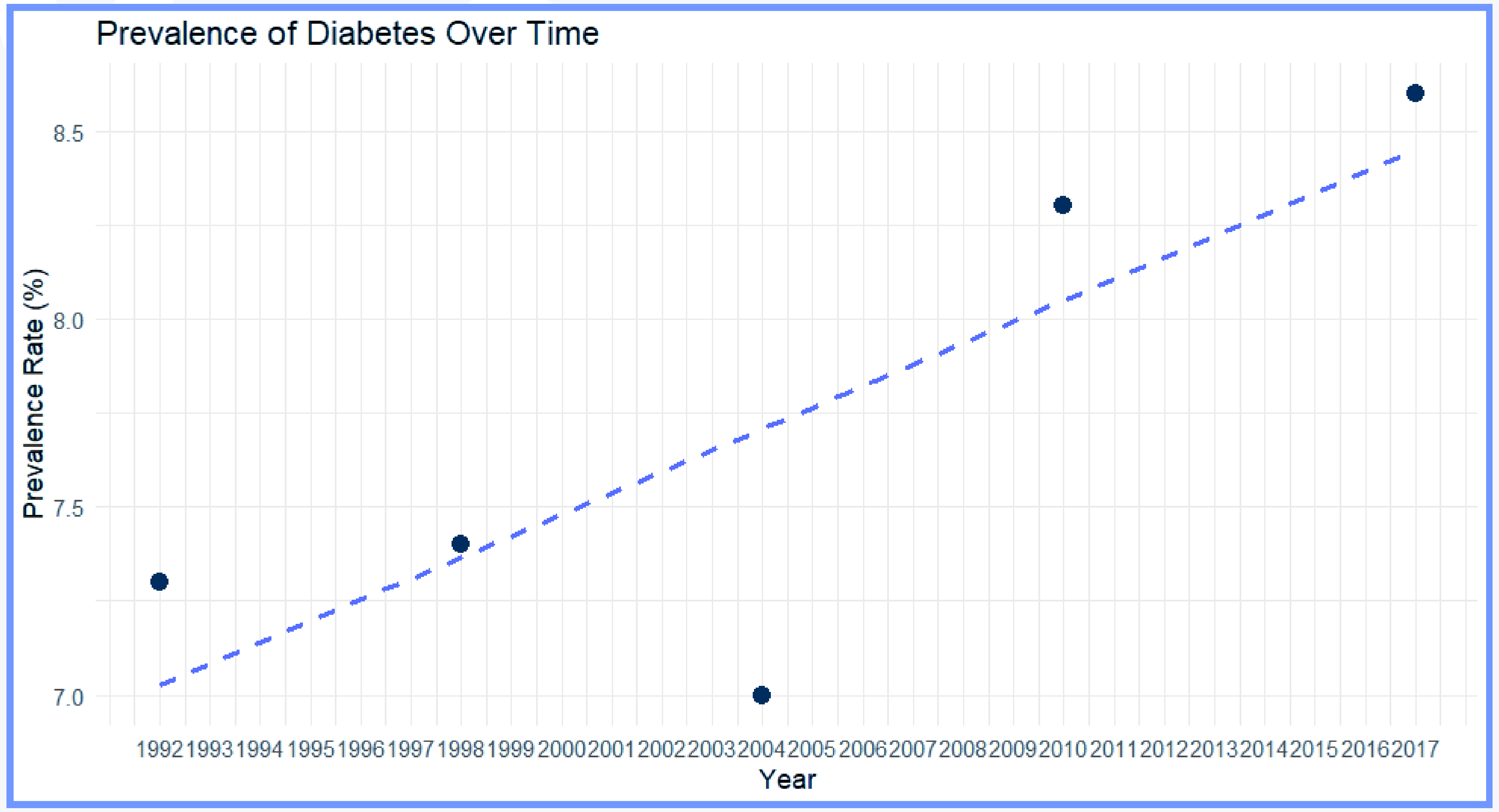
Strengthening Healthier SG's preventive care ecosystem with personalised lifestyle interventions and dashboards using predictive models

BC2406 Group 3:

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

Table of Content

Diabetes Mellitus



What is Diabetes Mellitus?

- Chronic metabolic disease that occurs when the body is unable to use insulin effectively, resulting in high blood glucose levels over time.
- Can lead to serious complications (cardiovascular disease, kidney failure, and blindness).

Prevalence

- Increasing over time.
- By 2017, about 14% of Singaporeans aged 18 to 69 years were diagnosed with pre-diabetes (MOH, 2017).
- By 2017, about 8.6% of Singaporeans aged 18 to 69 years were diagnosed with T2DM (MOH, 2024).

Anastasia Koesoemo

T2DM: Healthier SG

Aim

To shift the healthcare system from solely treatment focused to preventive care

How

- Enrol with family doctors
- Adopt personalised health plans
- Embrace healthier lifestyles (HealthHub and Healthy365)

Our Goals

- Implement intervention using predictive analytics to reduce the likelihood of developing diabetes
- Help Singapore achieve more efficient screening, better resource allocation and improved population health outcomes in the future.

Anastasia Koesoemo

Problem Statement

Primary Care Networks

Regular Health Screening

Current Measures

Community Programmes

Population-Based Prediction

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**.

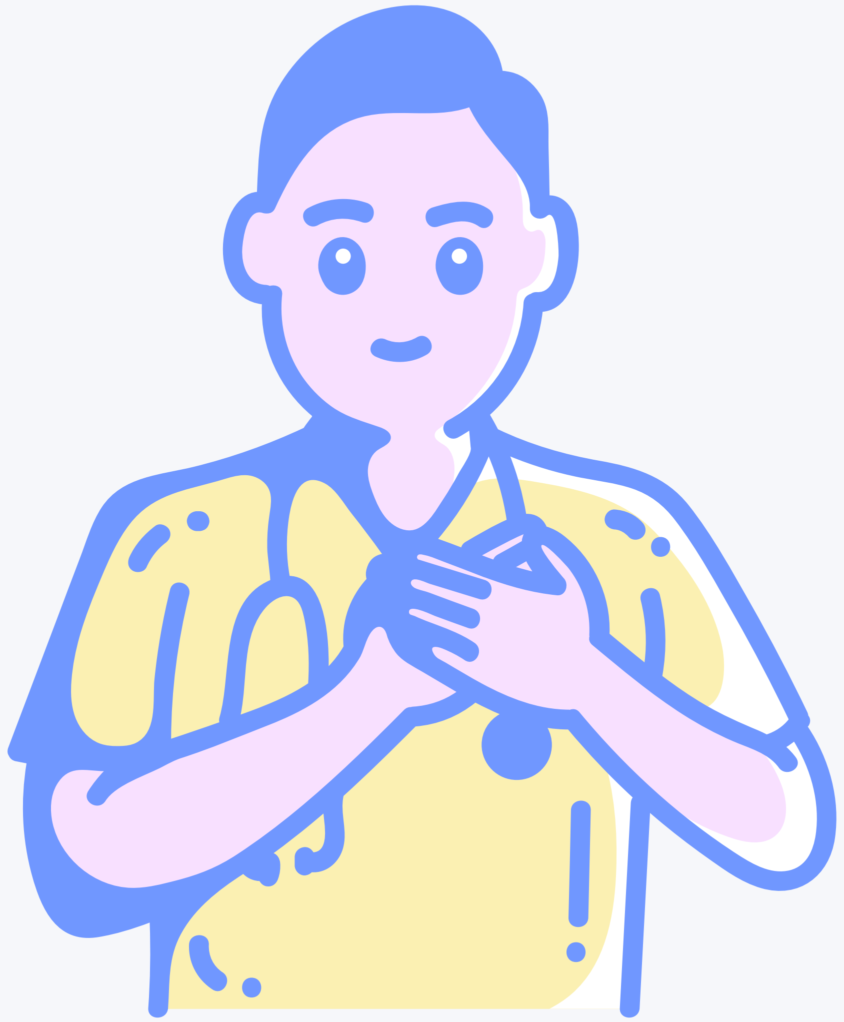
Through this project, we aim to leverage on **predictive modelling techniques** such as regression and CART to innovate on further solutions. Thereby **enhancing prevention**, enabling doctor engagement and **supporting Healthier SG's goal** towards proactive community based care.

Anastasia Koesoemo

Datasets

Early Symptoms

- 520 observations
- 17 variables
- early warning signs that may indicate potential onset of diabetes.



Health Indicators

- 253,680 observations
- 22 variables
- measurable wellness and fitness related metrics.

Stephen Michael Lee

Data Pipeline

1

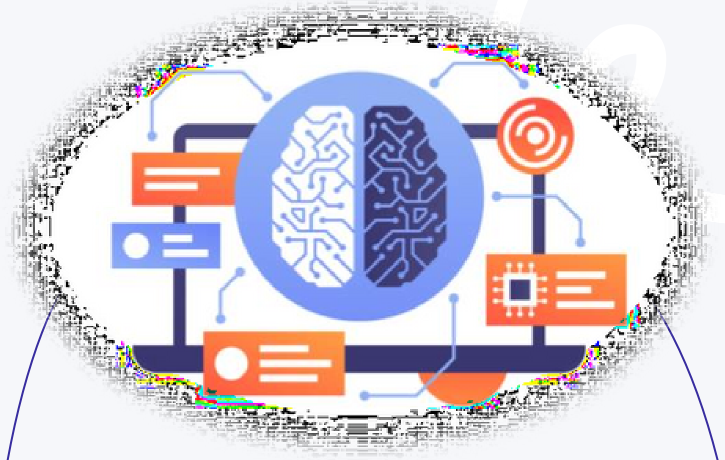
Data Collection
Choosing the 2 best datasets from Kaggle

2

Data Exploration
Exploratory visualisations & correlations

3

Data Preprocessing
Duplicate handling, NA values, outlier treatment, class balancing

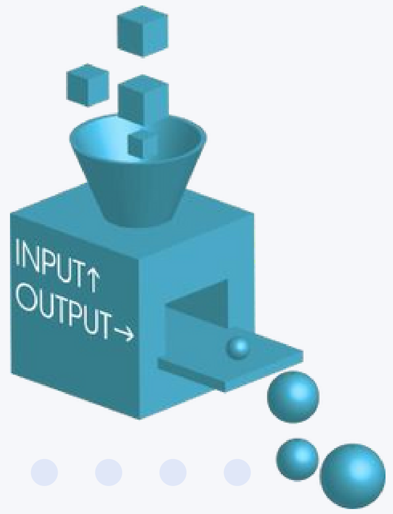


4

Model Training
Optimise each model & select the best with predefined metrics

5

Implementations
Propose data-driven solutions based on model results



Data Exploration

Total Unique Patients

251

(269 duplicates removed)

Diabetes Cases

173

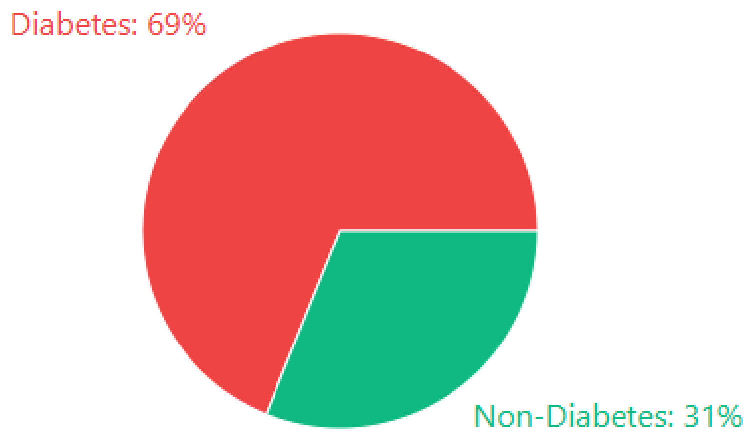
68.9%

Non-Diabetes Cases

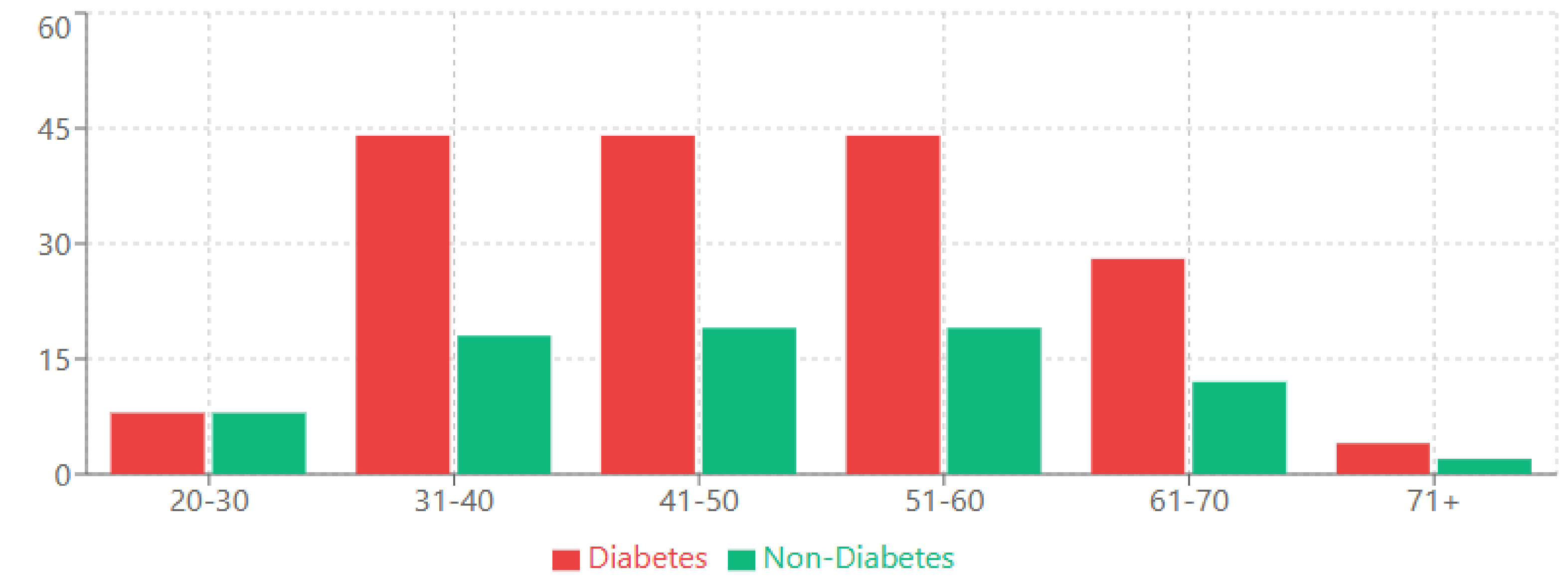
78

31.1%

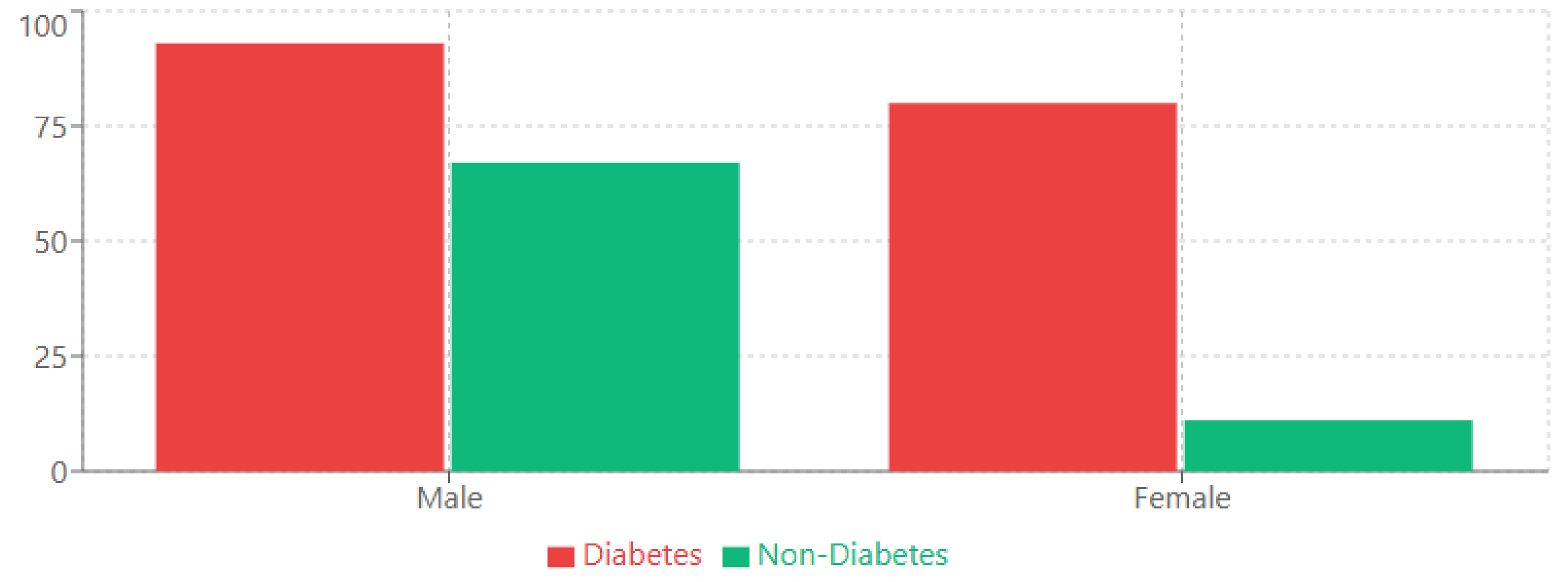
Class Distribution



Age Distribution



Gender Distribution

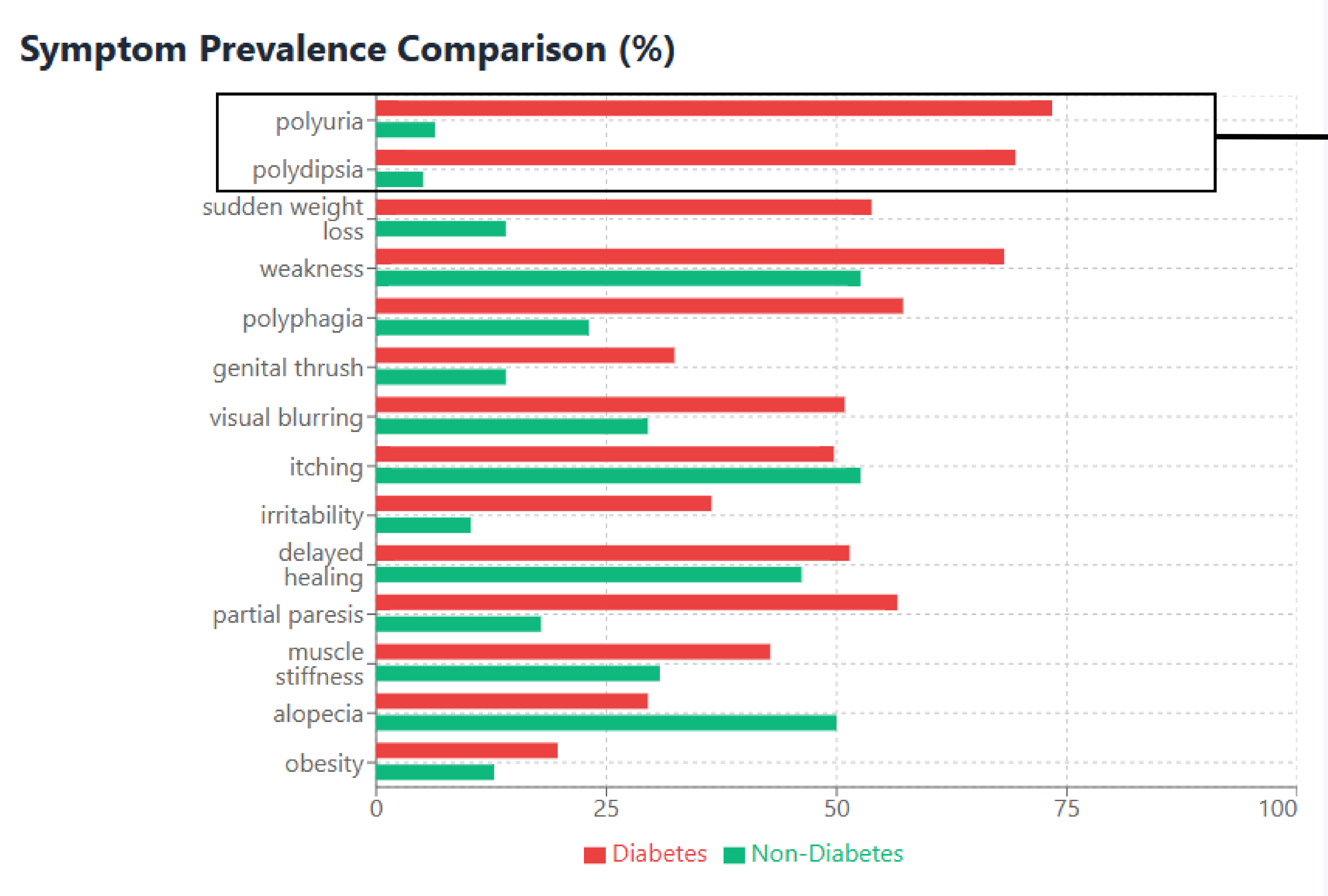


Insights:

- Slightly imbalanced towards diabetics → No sampling required
- Dataset shows higher proportion of females have diabetes
- After 30 years old diabetes start to become more prevalent

Stephen Michael Lee

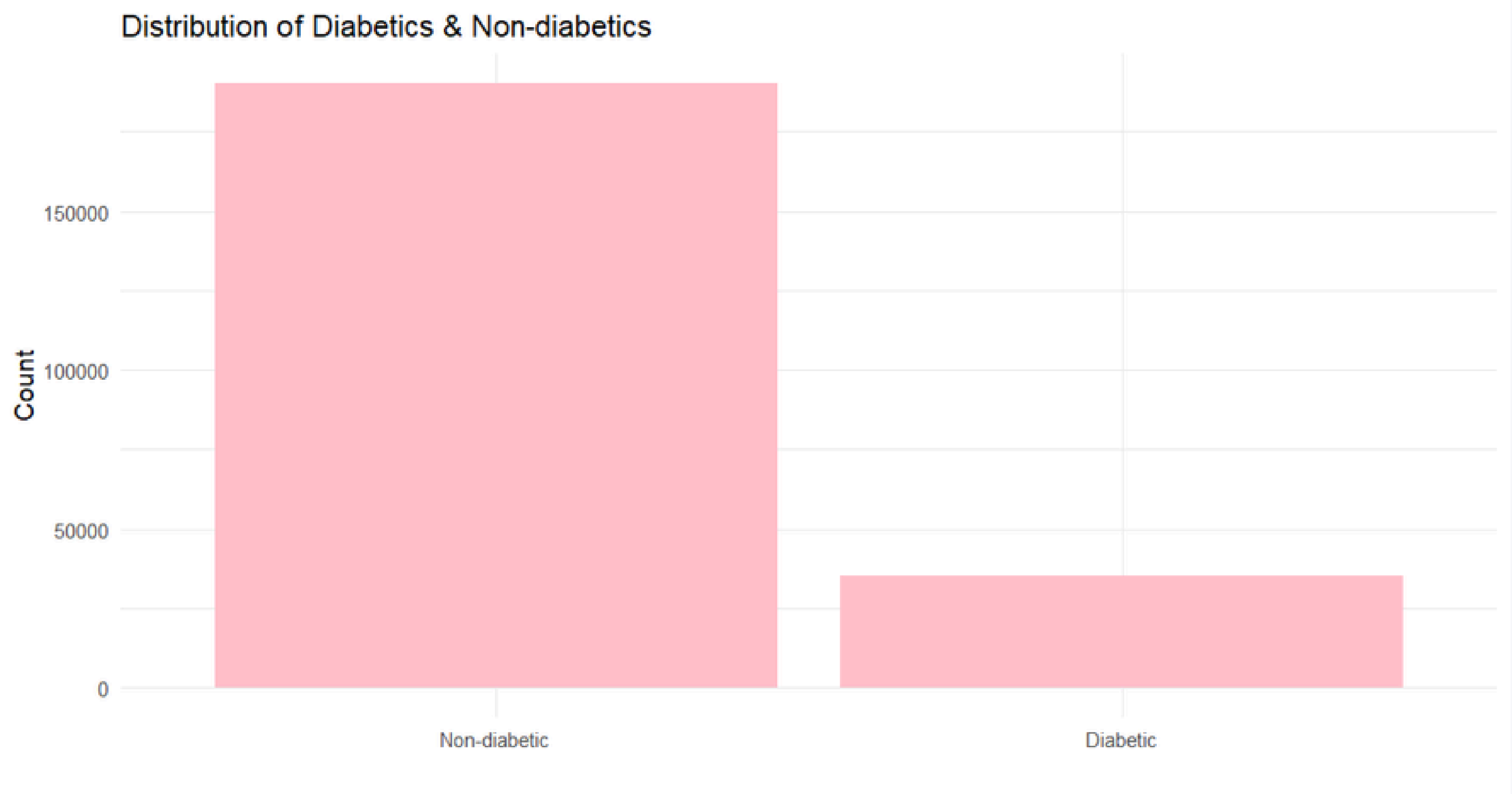
Data Exploration

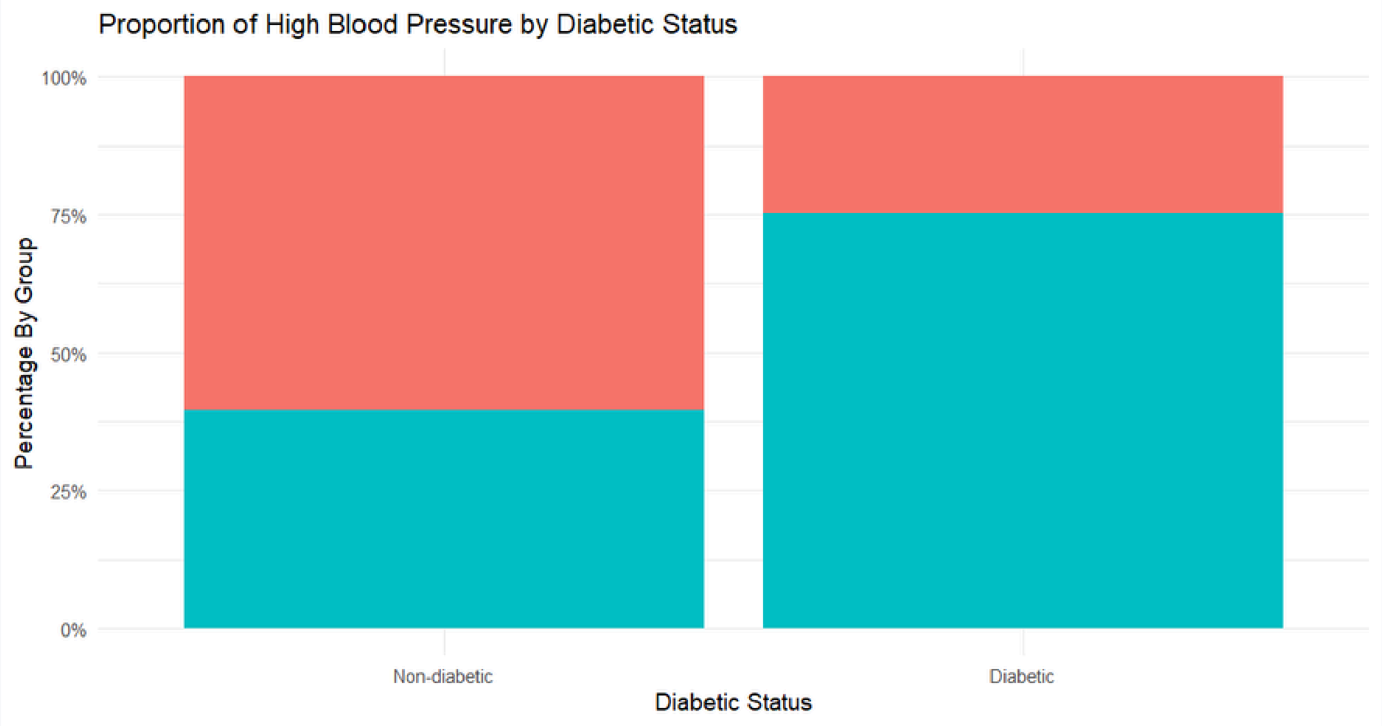


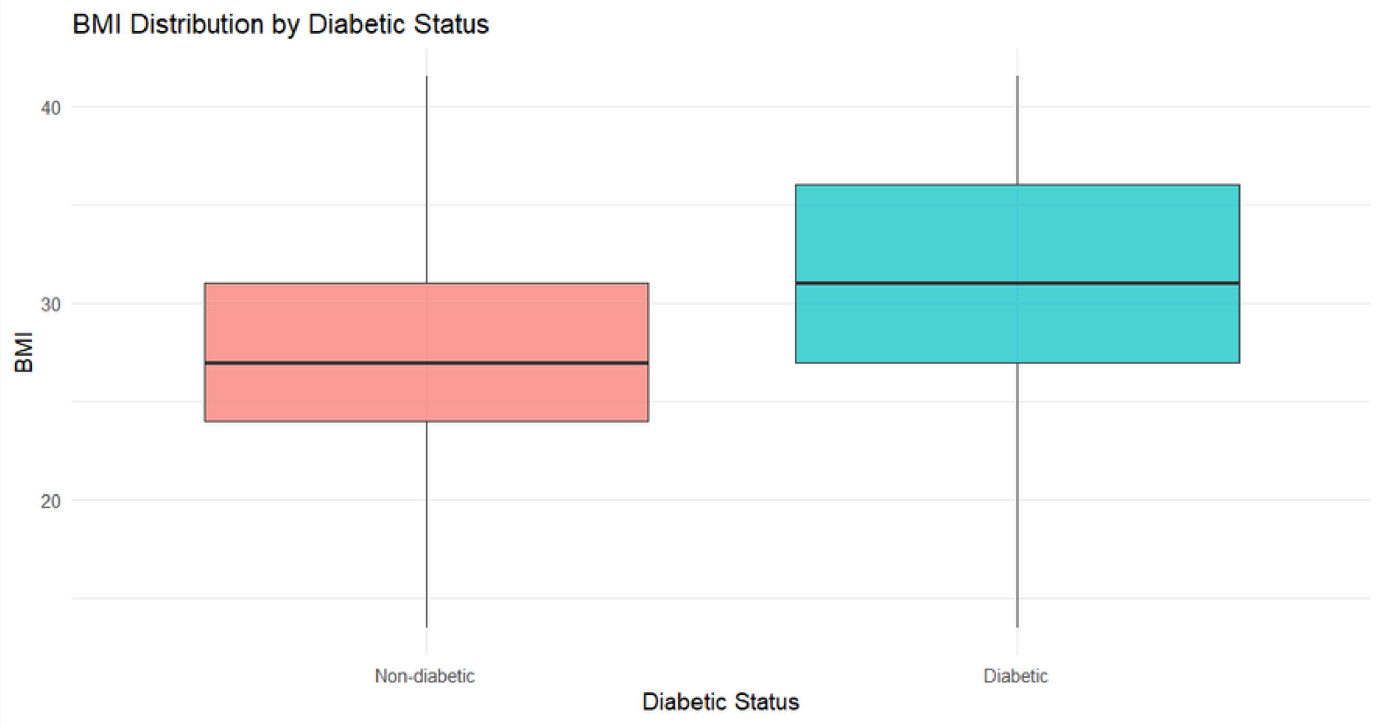
Most Distinctive Symptoms

Symptom	Definition
Polyuria	Excessive peeing
Polydipsia	Excessive thirst

Data Exploration







Insights:

- Quite imbalanced for diabetics (~ 85%) → Sampling required
- ~75% of diabetics have High BP while non-diabetics are ~40%
- Diabetics have a median BMI of ~31 while non-diabetics are ~27

Pre Modelling

Data Cleaning → **Train-Test Split** → **Performance Metric**

Data Cleaning

Early Symptoms: Duplicate removal

Health Indicators: Duplicate removal, binary reclassification of target Y variable, Outlier treatment for BMI

Train-Test Split

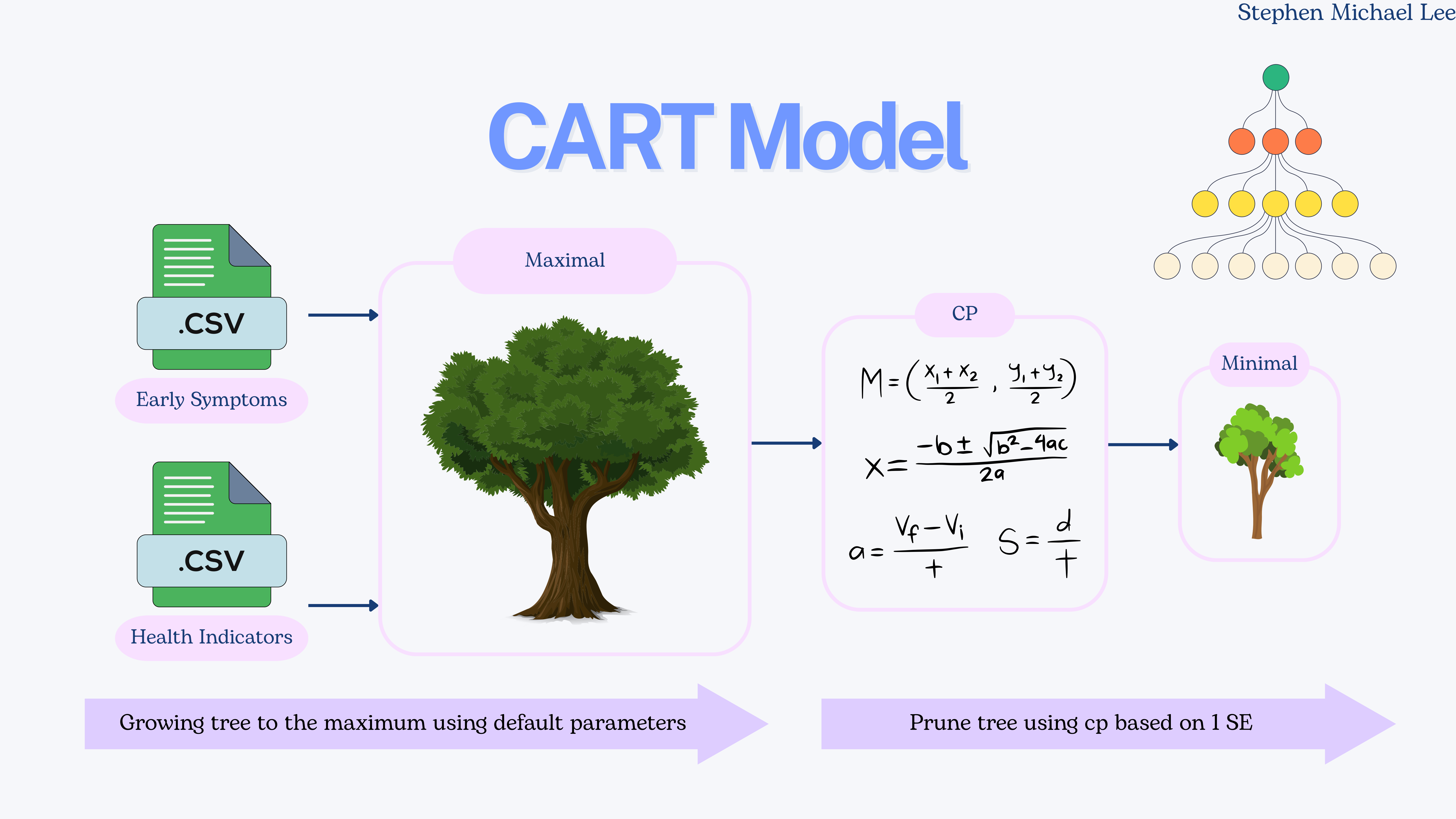
The datasets were randomly divided into training and test sets in a ratio of **7:3**.

Sampling was required for Health Indicators due to the imbalanced class distribution

Performance Metric

Classification Accuracy
Percentage of correct predictions (TP + TN)

False Negative Rate
Error of misclassifying high-risk populations as low risk


Stephen Michael Lee
CART Model
.CSV
Early Symptoms
.CSV
Health Indicators
Maximal
CP
$M=\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right)$
$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$
$a=\frac{V_f-V_i}{t}$
$S=\frac{d}{t}$
Minimal
Growing tree to the maximum using default parameters
Prune tree using cp based on 1 SE

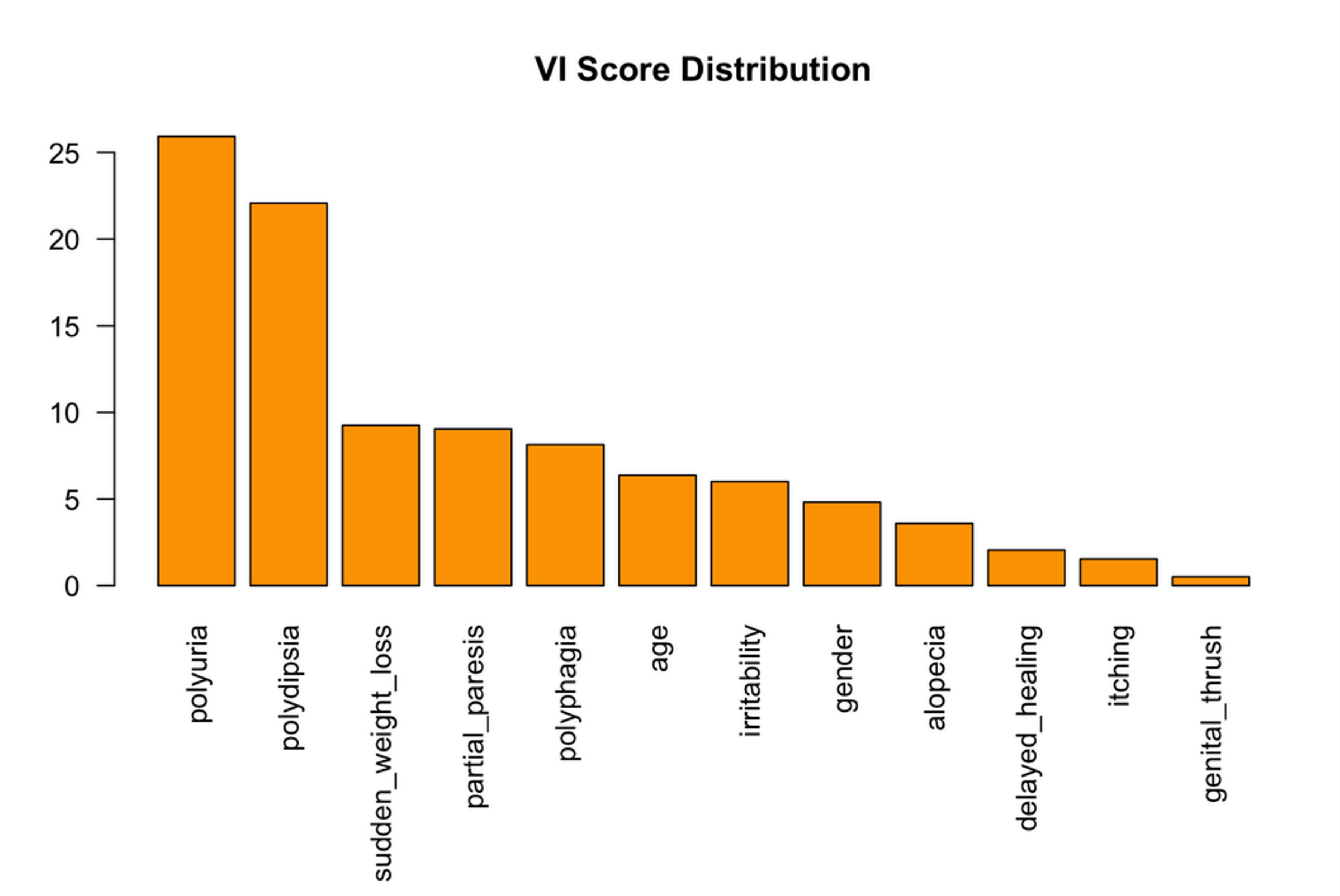
Optimal Tree Analysis

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

Optimal Tree Analysis

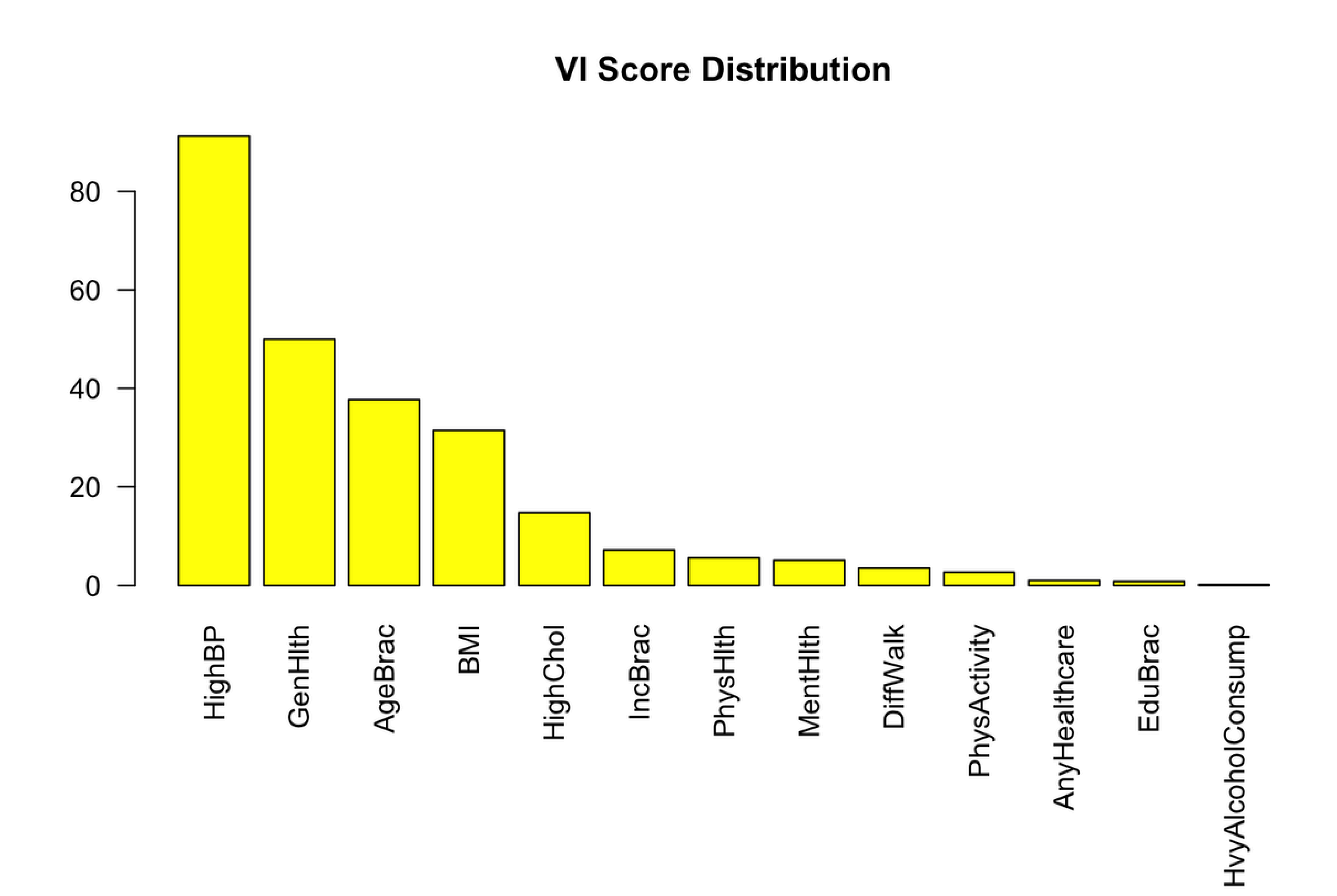
Dataset	Advantage	Limitations
Early Symptoms	• Simpler tree, easy to interpret and visualise • Fewer splits reduce risk of overfitting • Performs well when key early symptom variables are strong indicators	• May underfit, limited depth can miss subtle patterns • Slightly lower predictive accuracy due to small dataset and fewer variables • Less generalisable to diverse populations
Health Indicators	• Captures interactions among multiple health indicators • Higher accuracy due to richer data and lower CP, allowing finer splits • Better generalisation to real-world health datasets	• Risk of overfitting due to deeper tree and more variables • Harder to interpret clinically • Requires more data and computational power

Variable Importance (ES)



Rank	Predictor	Score
1	Polyuria	25.92
2	Polydipsia	22.07
3	Sudden Weight Loss	9.25
4	Partial Paresis	9.04
5	Polyphagia	8.14

Variable Importance (HI)



Rank	Predictor	Score
1	High Blood Pressure	91.15
2	General Health Rating	49.92
3	Age Bracket	37.71
4	BMI	31.44
5	High Cholestrol	14.78

Logistic Regression Model

- Supervised classification algorithm
- Chosen in particular because the response variable is categorical

Input (x)

Dataset with predictor variables

Logistic regression model

$$\ln(\text{odds}) = \beta_0 + \beta_1\chi_0 + \ldots + \beta_p\chi_p$$

$$\text{where odds} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{P(Y=1)}{P(Y=0)}$$

Output (y)

0: non diabetic

Output (y)

1: diabetic

Lim Li Xuan Valencia

Running Logistic Regression Model

early symptoms dataset

```
Call:
glm(formula = class ~ ., family = "binomial", data = trainset)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.37902    1.66075   0.830 0.406335
age                  -0.04223    0.03930  -1.075 0.282486
genderMale           -2.53992    0.80508  -3.155 0.001606 **
polyuria1             3.06571    1.03652   2.958 0.003099 **
polydipsia1           4.37363    1.17824   3.712 0.000206 ***
sudden_weight_loss1   2.01569    0.96428   2.090 0.036585 *
weakness1            -0.23001    0.79443  -0.290 0.772174
polyphagia1           1.00244    0.92026   1.089 0.276025
genital_thrush1       2.33981    0.98314   2.380 0.017315 *
visual_blurring1      1.37721    0.96893   1.421 0.155206
itching1             -3.05658    1.16210  -2.630 0.008533 **
irritability1         0.82750    0.97647   0.847 0.396751
delayed_healing1      0.19574    1.11572   0.175 0.860734
partial_paresis1      1.95581    0.89314   2.190 0.028537 *
muscle_stiffness1     0.05228    1.06289   0.049 0.960771
alopecia1             0.66103    0.93860   0.704 0.481263
obesity1              0.23131    0.85169   0.272 0.785941
---
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:  68.633  on 159  degrees of freedom
AIC: 102.63

Number of Fisher Scoring iterations: 8
```

health indicators dataset

```
glm(formula = Outcome ~ ., family = "binomial", data = trainset)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.887398   0.904938  -8.716  < 2e-16 ***
HighBP1               0.763272   0.137026   5.570 2.54e-08 ***
HighChol1             0.693552   0.129406   5.360 8.34e-08 ***
CholCheck1            1.564735   0.580420   2.696 0.007021 **
BMI                   0.088224   0.012038   7.329 2.32e-13 ***
Smoker1              -0.032019   0.132088  -0.242 0.808464
Stroke1              -0.125240   0.269333  -0.465 0.641931
HeartDiseaseorAttack1 0.118960   0.194484   0.612 0.540755
PhysActivity1        -0.213625   0.144380  -1.480 0.138981
Fruits1               0.048647   0.135414   0.359 0.719409
Veggies1             -0.063234   0.163572  -0.387 0.699065
HvyAlcoholConsump1   -1.322211   0.383085  -3.451 0.000558 ***
AnyHealthcare1        0.260770   0.310072   0.841 0.400348
NoDocbcCost1         -0.055018   0.229497  -0.240 0.810537
GenHlth               0.491263   0.080756   6.083 1.18e-09 ***
MentHlth             -0.004628   0.008702  -0.532 0.594797
PhysHlth             -0.005373   0.008253  -0.651 0.515000
DiffWalk1             0.031796   0.172981   0.184 0.854163
Sex1                  0.173744   0.133580   1.301 0.193372
AgeBrac               0.162351   0.027307   5.945 2.76e-09 ***
EduBrac               0.013830   0.070378   0.197 0.844207
IncBrac              -0.002008   0.036274  -0.055 0.955848
---
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: 1494.0  on 1378  degrees of freedom
AIC: 1538
```

key:
significant predictors
($p<0.05$)

Variable Selection:

Stepwise Algorithm

- Unnecessary predictors that do not improve predictive power are removed
- Akaike Information Criterion (AIC) is used to measure how well a model fits
  - Model with the lowest AIC has the best fit
- Final model keeps only significant predictors

$$AIC_1 > AIC_2 > AIC_3 \qquad \ldots \qquad Lowest\ AIC$$

Model 1

Dataset with n variables

→

Model 2

Dataset with n-1 variables

→

Model 3

Dataset with n-2 variables

...

Final Model

Dataset with n-m variables

Lim Li Xuan Valencia

After Stepwise Algorithm Enhancement

early symptoms 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
```

key: significant predictors (p<0.05)

Lim Li Xuan Valencia

Checking Variables:

Variance Inflation Factor

- Key assumption for logistic regression:
  - Independent variables should not be highly correlated to one another
- High collinearity can distort the estimated coefficients and increase errors
- Variance Inflation Factor (VIF) was calculated for all retained predictors
  - VIF < 5: multilinearity is not a concern

early symptoms 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 indicator dataset

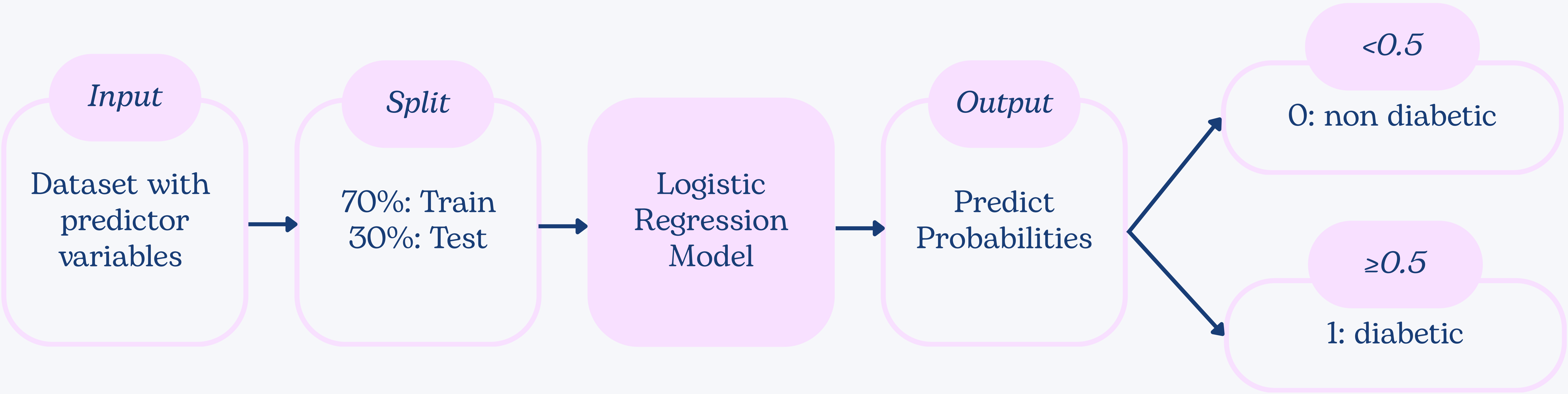
```
> vif(m2)
         HighBP          HighChol         CholCheck               BMI HvyAlcoholConsump           GenHlth
       1.101178          1.030035          1.007725          1.115464          1.003999          1.047713
        AgeBrac
       1.138127
```

✔ All predictors have VIF<5

Lim Li Xuan Valencia

Output Classes:

Training and Testing the Model



Model Evaluation:

Odds Ratio (OR)

- OR shows how much more (or less) likely the outcome is to occur when a particular predictor is present, holding other variables constant
- Larger OR means a stronger effect

early symptoms dataset

```
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 indicator dataset

```
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
```

Lim Li Xuan Valencia

Model Evaluation:

Odds Ratio (OR)

Rank of Significant Predictors according to OR
The stronger the effect the higher the rank

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

Lim Li Xuan Valencia

CART vs Logistic Regression:

Evaluation Metric

From Confusion Matrix:

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

→

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

Lim Li Xuan Valencia

CART vs Logistic Regression:

Early Symptoms Dataset

Significant Predictors

CART

- polyuria
- polydipsia
- sudden_weight_loss
- partial_paresis
- polyphagia

Logistic Regression

- polyuria
- polydipsia
- sudden_weight_loss
- partial_paresis
- gender
- genital_thrush
- itching

Lim Li Xuan Valencia

CART vs Logistic Regression:

Early Symptoms Dataset

Model Perfomance

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

By results, CART model performs better.

Lim Li Xuan Valencia

CART vs Logistic Regression:

Health Indicators Dataset

Significant Predictors

CART

- HighBP
- GenHlth
- AgeBrac
- BMI
- HighChol

Logistic Regression

- HighBP
- GenHlth
- AgeBrac
- BMI
- HighChol
- CholCheck
- HvyAlcoholConsump

Lim Li Xuan Valencia

CART vs Logistic Regression:

Early Symptoms Dataset

Model Perfomance

Model	TP	TN	FP	FN	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
CART	237	191	109	63	71.3	79.9	63.7	68.5
Logistic Regression	239	209	91	61	74.7	79.7	69.7	72.4

From results, Logistic Regression model performs better.

Lim Li Xuan Valencia

CART vs Logistic Regression:

Conclusion

Early Symptoms Dataset

CART > Logistic Regression

Health Indicators Dataset

Logistic Regression > CART

Use both models together

CART: public facing self assessment modules

Logistic Regression: primary care analytics

Rae Ng

HEALTHIER SG

Proposed Solution

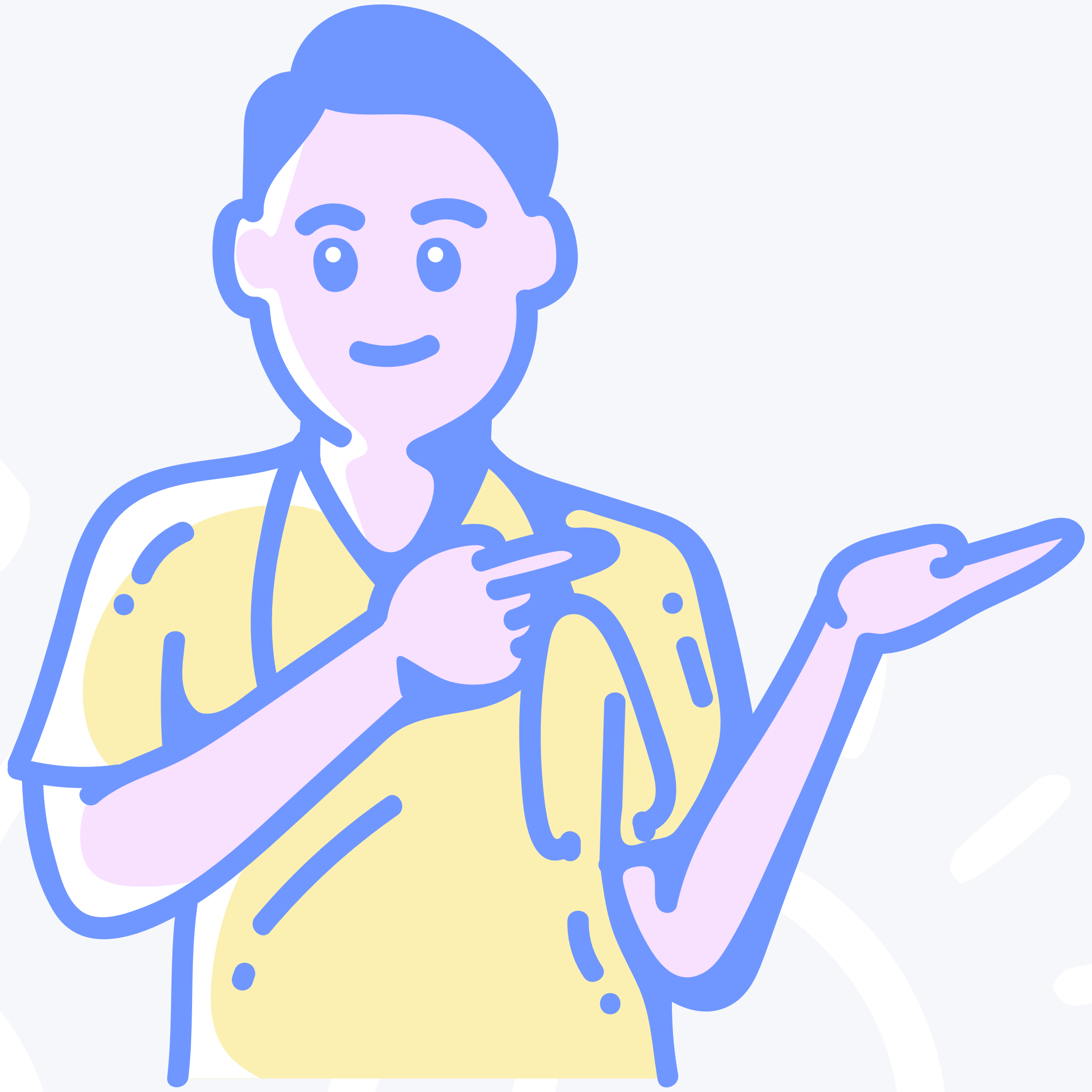
-DiaScope-

INDIVIDUAL SUPPORT

PRIMARY CARE SUPPPORT

Personalised Plan

Predictive Risk Alerts



- Integrated into Healthy365 to send alerts based on BMI, glucose, and activity data.
- Timely and contextual notifications (e.g., during meal times and breaks).
- Encourages small, daily healthy actions like choosing lower-sugar meals or taking short walks.
- Reinforces consistent, healthy habits through regular engagement.

Personalised Plan

Risk-targeted Engagement

- Notifications and guidance are **customised by individual risk level**
- **High-risk users:** frequent, focused tips
- **Low-risk users:** general wellness advice and light reminders
- Ensures **relevant and meaningful interventions** for sustained participation

Personalised Plan

Gamified Health Challenges

- Introduces **interactive themed challenges**
  - e.g. “10000 Steps a Day”, “Sugar Smart Month”
- Rewards users with **HPB e-vouchers, healthy food discounts, and fitness perks**
- Makes preventive care fun, motivating, and **habit-forming**
- Builds long-term engagement beyond one-off actions

Rae Ng

Community-Based Programmes

Mass Cooking Workshops on Diabetic Friendly Meals

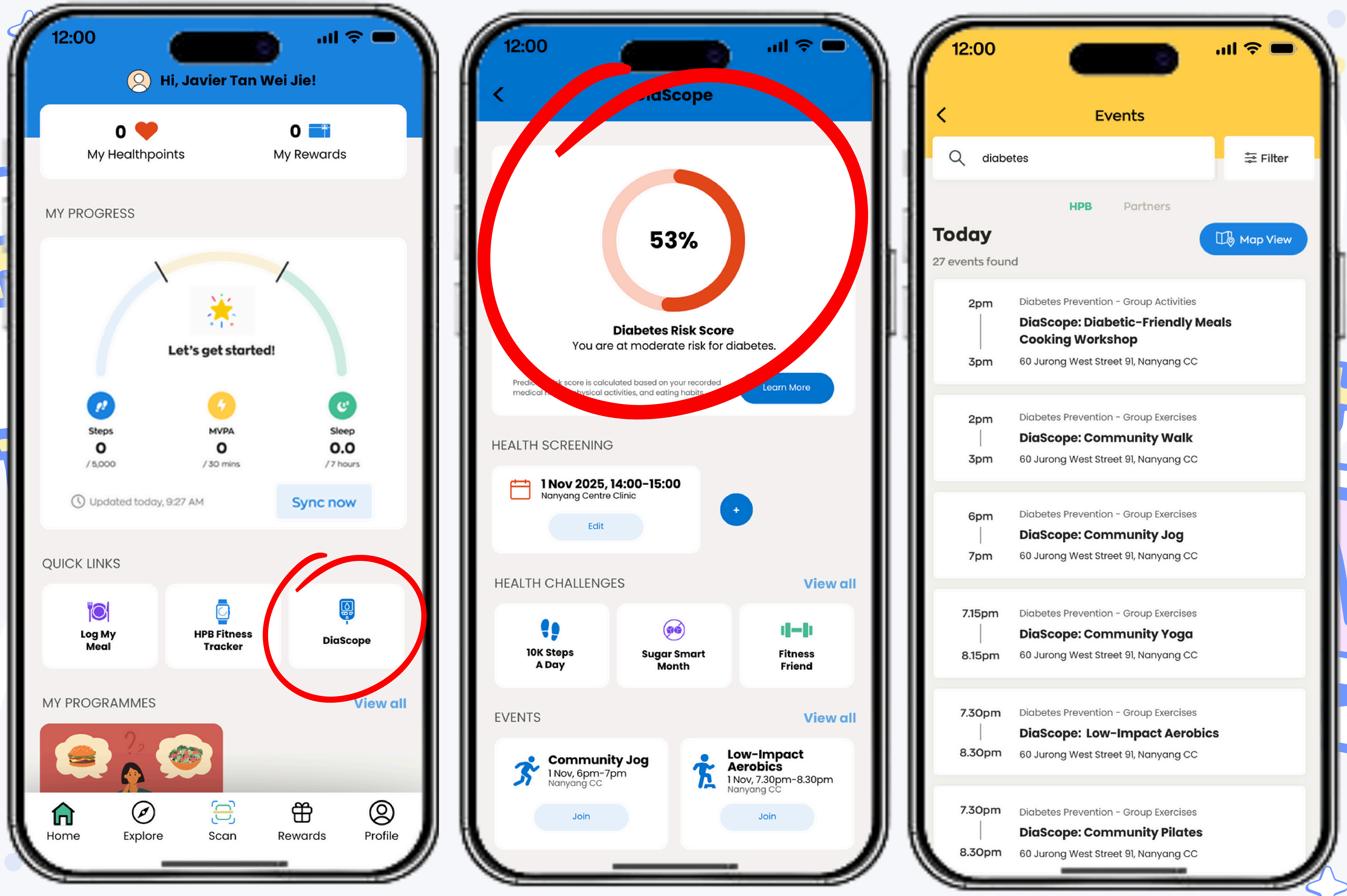
- Partner with HPB and community centres
- teach affordable, diabetic-friendly meals and encourage family participation

Group Physical Activities

- Weekly walks/jogs led by community ambassadors to promote regular exercise

- Linked to **Healthy365 team challenges** with small rewards for consistency
- Fosters a sense of **community, motivation** through **shared progress**

Rae Ng


12:00
Hi, Javier Tan Wei Jie!
0
My Healthpoints
0
My Rewards
MY PROGRESS
Let's get started!
Steps
0
/ 5,000
MVPA
0
/ 30 mins
Sleep
0.0
/ 7 hours
Updated today, 9:27 AM
Sync now
QUICK LINKS
Log My Meal
HPB Fitness Tracker
DiaScope
MY PROGRAMMES
View all
Home
Explore
Scan
Rewards
Profile
12:00
53%
Diabetes Risk Score
You are at moderate risk for diabetes.
Learn More
HEALTH SCREENING
1 Nov 2025, 14:00-15:00
Nanyang Centre Clinic
Edit
HEALTH CHALLENGES
View all
10K Steps A Day
Sugar Smart Month
Fitness Friend
EVENTS
View all
Community Jog
1 Nov, 6pm-7pm
Nanyang CC
Join
Low-Impact Aerobics
1 Nov, 7.30pm-8.30pm
Nanyang CC
Join
12:00
Events
diabetes
Filter
HPB
Partners
Today
Map View
27 events found
2pm
3pm
Diabetes Prevention - Group Activities
DiaScope: Diabetic-Friendly Meals Cooking Workshop
60 Jurong West Street 91, Nanyang CC
2pm
3pm
Diabetes Prevention - Group Exercises
DiaScope: Community Walk
60 Jurong West Street 91, Nanyang CC
6pm
7pm
Diabetes Prevention - Group Exercises
DiaScope: Community Jog
60 Jurong West Street 91, Nanyang CC
7.15pm
8.15pm
Diabetes Prevention - Group Exercises
DiaScope: Community Yoga
60 Jurong West Street 91, Nanyang CC
7.30pm
8.30pm
Diabetes Prevention - Group Exercises
DiaScope: Low-Impact Aerobics
60 Jurong West Street 91, Nanyang CC
7.30pm
8.30pm
Diabetes Prevention - Group Exercises
DiaScope: Community Pilates
60 Jurong West Street 91, Nanyang CC

Ngiam Kai Her Darryn

Primary Care Support

The app will also contain primary care supports that targets helping mainly family doctors.

How?

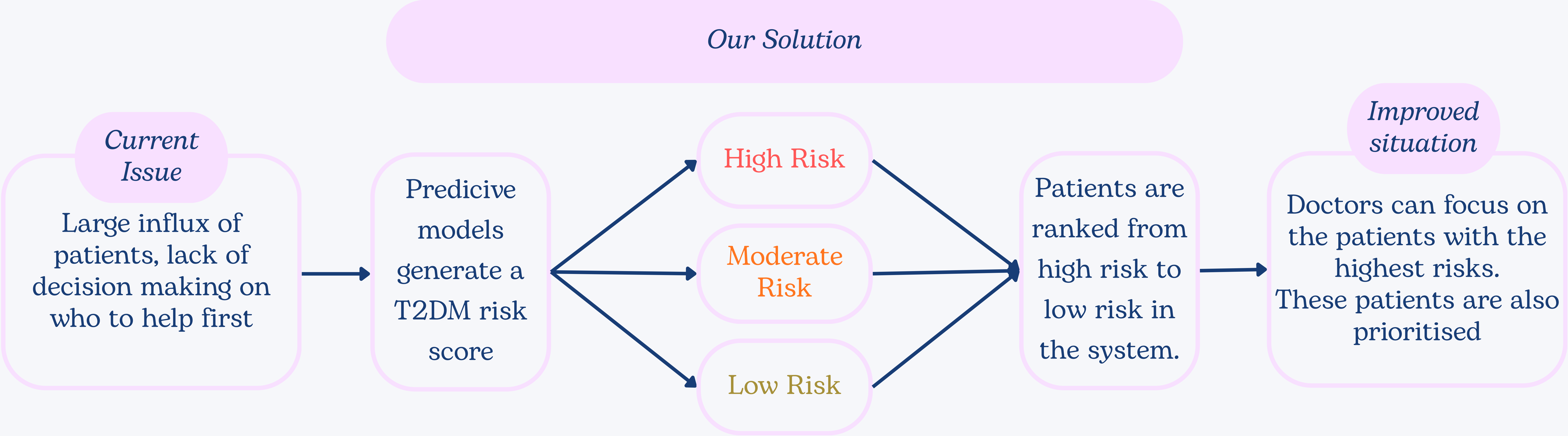
Data is extracted from our predictive Logistic Regression and CART models.

This entails two components:
(a) Who to Help First Model
(b) Integrated and Continuous Data Collection & Monitoring

Ngiam Kai Her Darryn

Who to Help First Model

- Decision-based system
- To prioritise the limited resources in the healthcare system to patients with the highest risk.



Ngiam Kai Her Darryn

Integrated & Continuous Data Collection and Monitoring

Data is constantly updated with patient-generated data from Healthy365 as well as clinical data from health screenings and lab tests.

How it will look like:

Scenario

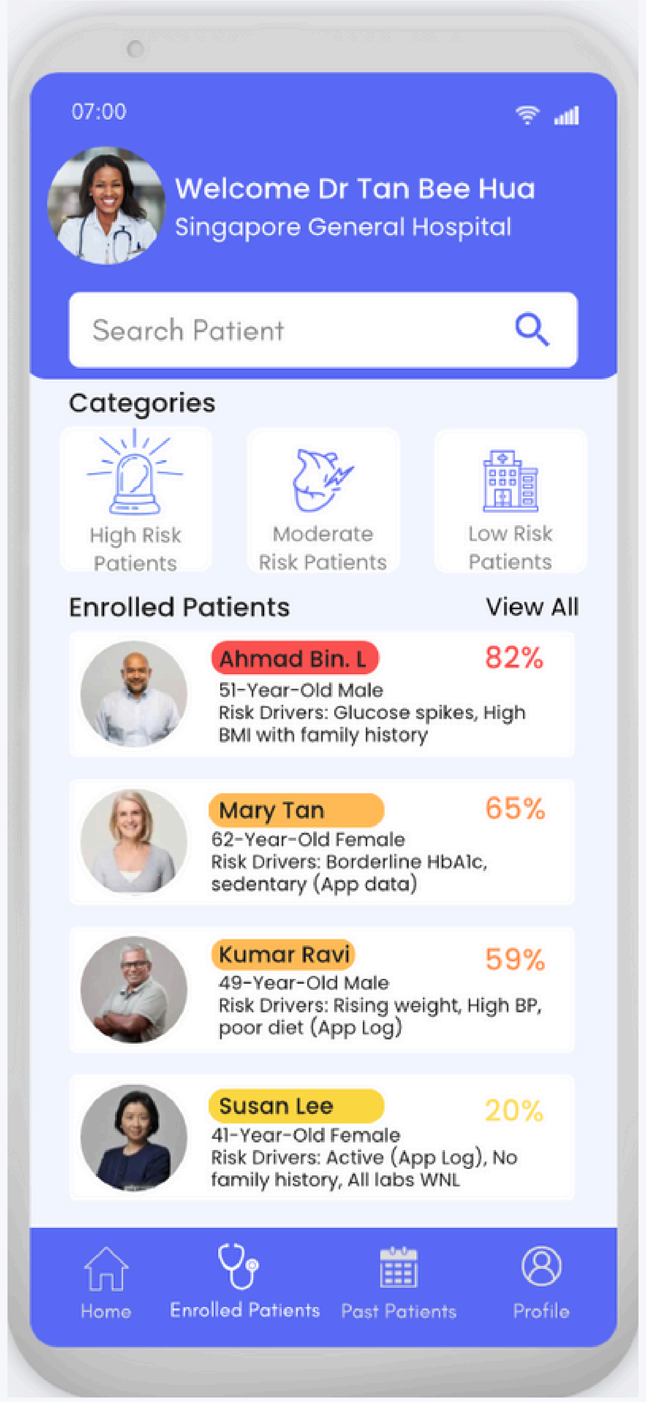
Blood sugar spike in high risk patient.

→ Automated alert sent to clinic

→ Team is able to intervene even before patient's scheduled visit

→ **Outcome**

Prevent complications as well as reduce any avoidable emergency department attendances.



Lim Li Xuan Valencia

Approach (Overview):

Business Application

Goal: Integrating predictive analysis with our solution into existing digital health infrastructure

Phase 1	Phase 2	Phase 3	Phase 4
Data Collection and Governance	Generation of Diabetes Risk Scores	Dynamic Updates and Automation	Clinical Regulation

Phase 1

Data Collection and Governance

- Partner with public hospitals and polyclinics for clinical data
- Collect self reported lifestyle data via HealthHub questionnaires
- Questionnaires can be completed digitally before clinic visits while clinical measures are drawn from routine check ups → low burden process
- Follow PDPA compliant consent, de-identification, and secure sharing to protect patients

Lim Li Xuan Valencia

Phase 1

Lim Li Xuan Valencia

Data Collection and Governance

Possible Partners to Collaborate with on Data Collection

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 • SingHealth Polyclinics	• largest patient base in Singapore
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

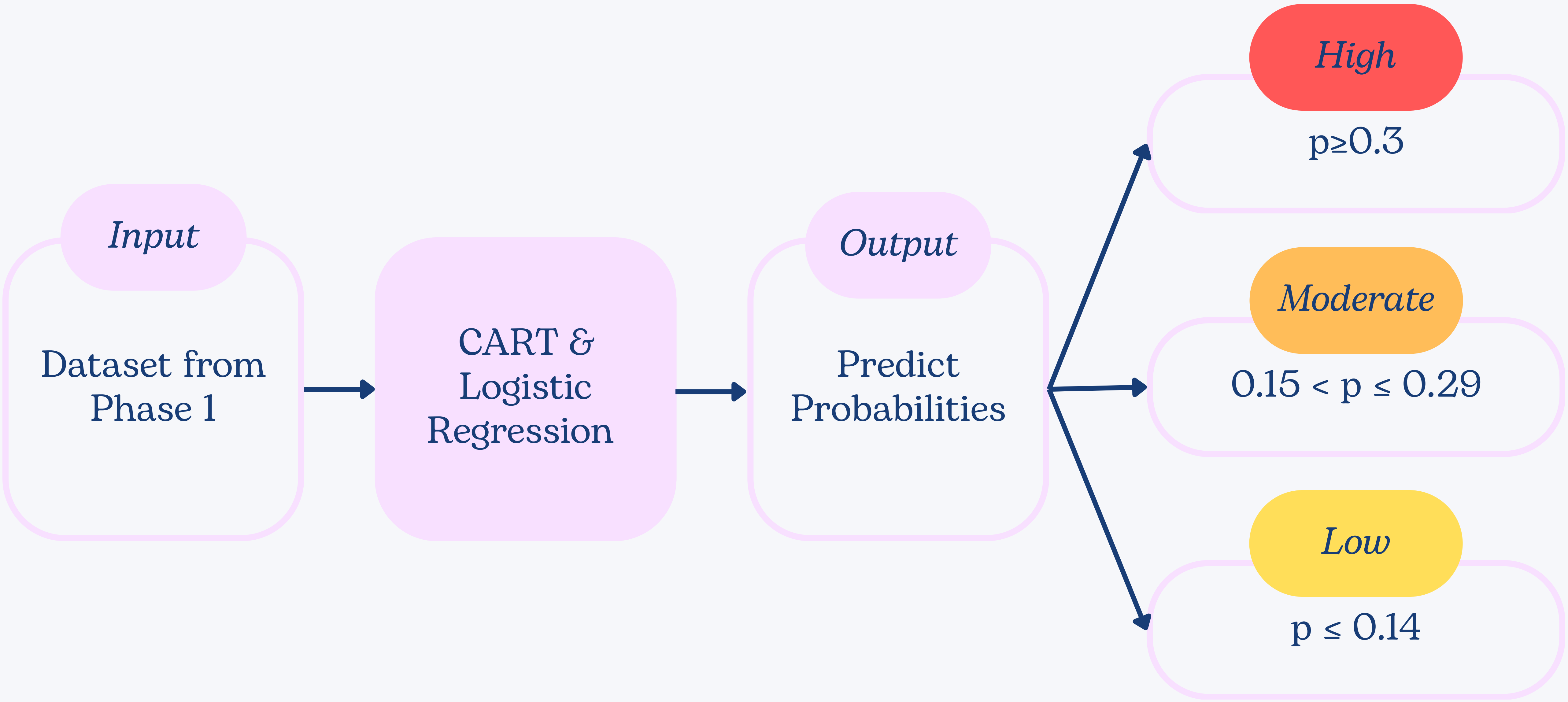
Questions for Questionnaire

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) Yes: drank more than 14 drinks per week (adult men) drank more than 7 drinks per week (adult women)	HvyAlcoholConsump
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

Phase 2

Generation of Diabetes Risk Scores

- Combine data from residents and doctors → run Logistic Regression + CART models
- Predict each individual’s likelihood of developing diabetes
- Classify results into three risk tiers for clear interpretation



Lim Li Xuan Valencia

Phase 2

Generation of Diabetes Risk Scores

Usage of Predictive Score in Solution

Tier	Probability Threshold	Measures
High Risk	≥ 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	≤0.14	• Maintain lifestyle plan • Receive standard preventive reminders

Lim Li Xuan Valencia

Phase 3

Dynamic Updates and Automation

Behavioural updates

Users see tangible progress after active participation and meal logging improve scores over time

e.g. “If maintained for 8 weeks, your risk may drop by 5%.”

Clinical updates

Annual check-up or lab results automatically refresh risk scores

Doctor dashboard

Alerts clinicians when patients’ risk levels increase for timely follow-up

Lim Li Xuan Valencia

Phase 4

Clinical Regulation

MOC

Establish
Model Oversight Committee:
clinicians + data scientists

→

- Review model outputs quarterly for accuracy and fairness
- Reduce false negative cases to avoid missing high risk individuals
- Ralse positive cases to manage clinical workload
- Track key performance metrics:
  - accuracy
  - sensitivity
  - specificity
  - precision
  - screening uptake
  - doctor engagement

Ngiam Kai Her Darryn

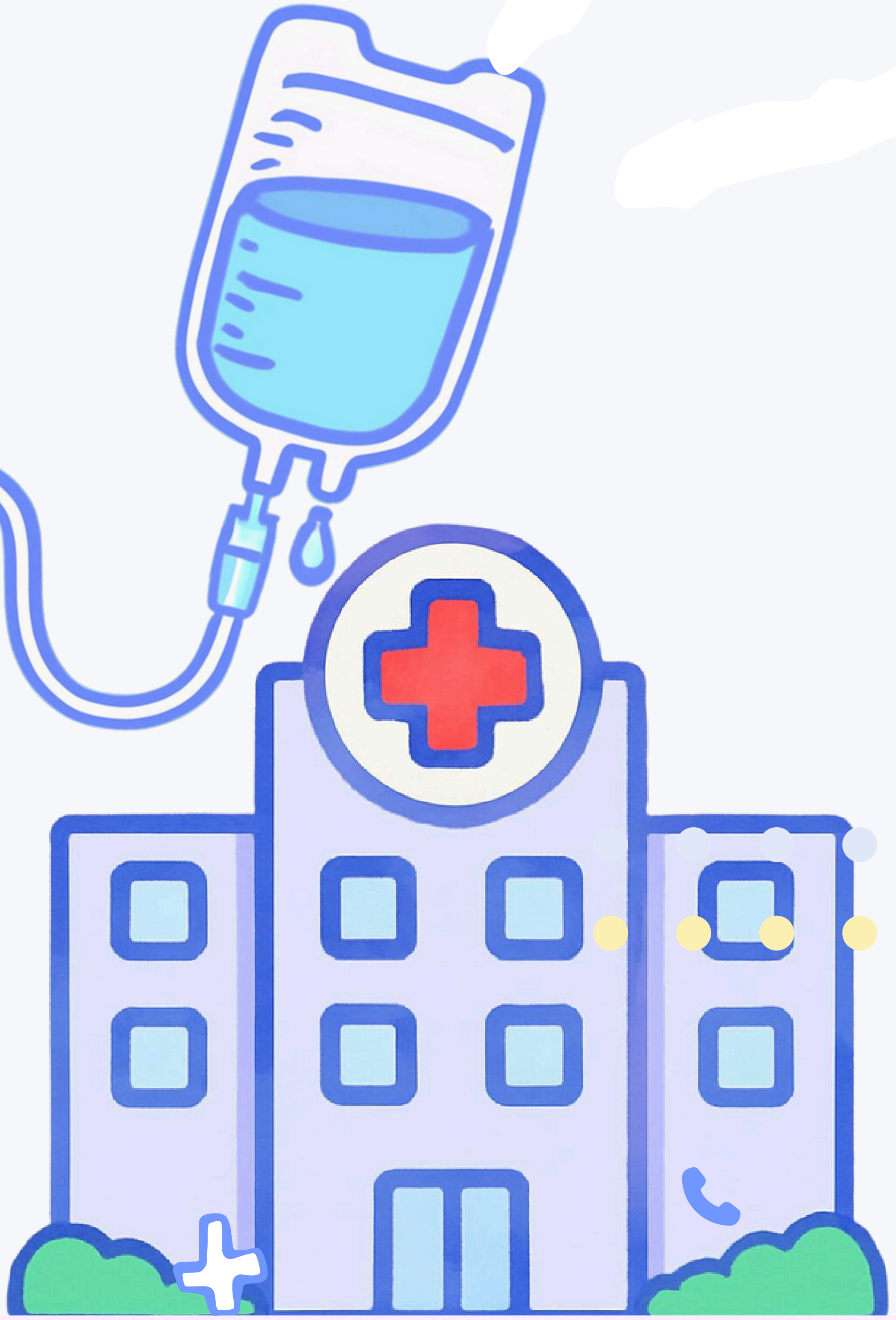
Strengths

- Reactive and data-driven
- Support for both patients and doctors
- Transitioning the healthcare model from reactive to proactive

Ngiam Kai Her Darryn

Strengths:

Digital Patient Tools



- Predictive analytics
- Personalised predictive risk alerts
- Recommends risk-targetted engagement to allow the correct interventions

Ngiam Kai Her Darryn

Strengths:

Primary Care Management

- Allows for efficient allocation of resources
- Doctors can easily identify and prioritise high risk patients
- Comprehensive, real-time data view by integrating patient-generated app data

Ngiam Kai Her Darryn

Strengths:

Social Prescriptions

- Using analytics to decide the most relevant community program based on the user's specific risk profile and preferences.
- Efficient usage of community resources
- Increased likelihood of patient adherence

Ngiam Kai Her Darryn

Limitations

- Unrepresentative datasets
- Narrow scope
- User uptake

Ngiam Kai Her Darryn

Limitations:

Unrepresentative Datasets

- DiaScope was built on publicly available datasets, not from Singapore

- Not representative of Singapore’s genetic diverse makeup

Limitations:

Ngiam Kai Her Darryn

Narrow Scope

- DiaScope only focuses on Diabetes
- Dashboard might be too simplistic
- Narrow range of applications

Ngiam Kai Her Darryn

User Uptake

- Patients may not be motivated to download or use the app.

- DiaScope cannot function properly without the active usage of the app.

Future Improvements

Enhancing Model Accuracy

Key Issue: Current models built on international datasets, not representative of Singapore's diverse population

Improvement:

- Validate and retrain models using **local Singapore datasets**
- Improves **accuracy, classification,** and **relevance** to local health needs
- Enables DiaScope to deliver **better personalised predictions** and insights

Future Improvements

Expanding the Scope

Current Limitation: Focuses mainly on diabetes

Improvement:

- Expand DiaScope to include **other chronic diseases**
- Creates a **comprehensive preventive health platform**
- Allows **wider population impact** and long-term **adaptability**

Rae Ng

Conclusion!

Conclusion: Toward Proactive Healthcare

Core Idea: Predictive analytics can **strengthen Healthier SG** through early detection and prevention

Model Insights:

- CART → better for symptom-based, public-facing tools
- Logistic Regression → clearer for clinical decision support

Conclusion: Toward Proactive Healthcare

System Impact:

- DiaScope integrates into HealthHub & Healthy365 → personalised health plans + data-driven dashboards
- Predictive Clinical Dashboard also helps doctors prioritise high-risk patients and act early
- Supports Singapore's shift from **reactive to proactive care,** empowering individuals and reducing hospital burden

Next Step:

- Validate locally and expand to other chronic diseases to make preventive care **targeted, efficient, and sustainable.**

Thank You!