PROGRAMME	:	Master of Business Analytics
SUBJECT CODE AND TITLE		BAA5043
ASSIGNMENT TITLE		Business Intelligence Group Assignment
LECTURER		Dr. Mubbasher
ASSIGNMENT DUE DATE:		13/08/2025

SUNWAY UNIVERSITY
Sunway BUSINESS SCHOOL

ASSIGNMENT COVER SHEET

STUDENT'S DECLARATION

1. I hereby declare that this assignment is based on my own work except where acknowledgement of sources is made.
2. I also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University/College or other institutions.
   [Submit "Turn-it-in" report (please tick √): Yes √ No]

NO.	NAME	STUDENT ID NO.	SIGNATURE	DATE
1.	Nidal Bencheikh Lehocine	19097617	Nidal Bencheikh Lehocine	13/08/2025
2.	Suhail Dorasamy	20020574	Suhail Dorasamy	13/08/2025
3.	Pyus Kabala	25006586	P.K	13/08/2025
4.	Muhammed Muftih	25000225	MUF	13/08/2025

E-mail Address / Addresses (according to the order of names above):

1. 19097617@imail.sunway.edu.my	3. .25006586@imail.sunway.edu.my
2. 20020475@imail.sunway.edu.my	4. 25000225@Imail.sunway.edu.my

APPROVAL FOR LATE SUBMISSION OF ASSIGNMENT (If applicable)
IF extension is granted, what is the revised due date? _____

Signature of Lecturer: _____ Date: _____

Contents

Introduction

For the purposes of our assignment we will be utilizing business intelligence and business analytics techniques for the fictional company Glucowell Medical, a specialized diabetes clinic. Glucowell is dedicated to delivering comprehensive patient care to individuals living with diabetes. Its mission is to improve the quality of life for their patients through a combination of clinical consultations, diagnostic testing, nutritional guidance and patient education initiatives. By integrating these services together, Glucowell offers a holistic approach to diabetes management that addresses factors from both the clinical and lifestyle perspective that influence patient health.

Due to the nature of the clinic, it operates in a very data-rich environment. With each interaction a patient has with the clinic generating valuable information. The effective management and analysis of the information generated by patients is crucial for Glucowell to maintain and improve their standard of care. However, this information is not useful in its raw form, especially due to its complexity and volume, it requires business intelligence and analytics processes to develop the information into actionable insights.

To support goals of Glucowell we have applied a range of BI techniques. Some of the key objectives include:

- Data cleaning and handling

To ensure that patient records and data are accurate, consistent and ready to be analysed by our machine learning models and for use in MS PowerBI.

- Descriptive analysis

To explore trends within patient demographics to inform decisions across the entire clinic.

- Comparative machine learning models

The evaluation of predictive models to forecast likelihood of patient having or being high risk of developing diabetes.

- Clustering for risk assessment

Grouping patients based on risk profiles to enable proactive interventions and tailored treatment strategies.

- Power BI dashboard creation

Developing an interactive, real-time dashboard to provide staff with accessible insights for decision-making and monitoring.

We expect that through the implementation of these processes that Glucowell stands to gain several operational and clinical benefits. With the predictive models and clustering techniques early detection and intervention programs can be developed and supported. This would enable the clinic to act before complications become potentially life-threatening. The patient segmentation through clustering would also allow for a more personalized care program, better aligning treatment plans with individual patient needs and risk factors.

Literature Review

Glucowell Medical, like most clinics and similar healthcare facilities, operates as a service-oriented business, where in the primary objective is to deliver personalized and high-quality care instead of tangible products. Healthcare services are inherently intangible, highly variable as they depend on individual patient needs and are often inseparable in their production and consumption, with care being delivered in real-time. The success of a healthcare organisation therefore is dependent not only on medical expertise but also on its ability to build patient trust and provide consistent services.

The growing burden of chronic diseases such as diabetes has created an increasing demand for continuous monitoring and long-term patient engagement, placing pressure on providers to deliver care efficiently. Healthcare facilities are also required to maximise efficiency without compromising quality with staffing, budget and time constraints.

In this context, data-driven decision-making has become indispensable for modern healthcare. Business Intelligence (BI) enables providers to convert large volumes of clinical and operational data into actionable insights that support informed decisions, optimized workflows, and strategic planning. Healthcare systems continuously produce extensive data throughout the diagnostic, treatment, and follow-up stages, which can be utilized to enhance clinical decisions, improve efficiency, and deliver better patient outcomes (Zargoush et al., 2025).

For diabetes care in particular, BI facilitates continuous monitoring of health indicators, early identification of high-risk patients, prediction of complications, and development of personalized treatment strategies. These capabilities not only enhance patient outcomes but also improve operational efficiency through better resource allocation, streamlined scheduling, and real-time performance tracking, ultimately strengthening the clinic's ability to deliver proactive, high-quality care (Das et al., 2025).

The dataset used for this project integrates a combination of quantitative variables such as BMI, fasting plasma glucose, systolic and diastolic blood pressure, cholesterol, triglycerides, LDL, HDL, daily caloric intake, and weekly exercise minutes and qualitative variables including gender, family history of diabetes, smoking, drinking, diet type, and sleep quality. This mix of clinical, behavioural, and demographic data provides a comprehensive foundation for examining both the medical and lifestyle factors influencing diabetes management (Munjala et al, 2025).

Key indicators such as fasting plasma glucose, lipid profiles, and blood pressure are vital for diagnosing and monitoring diabetes and its complications, while lifestyle factors like diet, exercise, and sleep quality offer insights into modifiable risks. Demographic attributes, including gender and family history, enable further patient stratification and inform personalized interventions. Prior to analysis, the dataset undergoes screening for missing observations and outliers to ensure integrity and consistency, with standardized units applied across all clinical variables.

In diabetes clinic centre, the business process flow begins with collecting clinical data from patient intake forms, vital signs (e.g., BMI, blood pressure), and lab results (e.g., fasting glucose). These raw data will be extracted, transformed, and loaded (ETL) into a centralized data warehouse. This will consolidate information from disparate sources for historical and operational analytics (Gavrilov et al., 2020). Once centralized, this will robust BI functions opens many possibilities. Data mining and discovery reveal trends (such as elevated BMI correlated with diabetes prevalence).

Interactive data visualization and reporting provides real-time dashboards to the clinics and administrators for proper analysis and performance monitoring (KMS Healthcare, 2024). This integrated BI workflow enables targeted interventions by identifying at-risk patient groups and supporting data-driven decision-making in diabetes management.

In this project, the dependent variable is the binary classification of diabetes status, coded as “0” for non-diabetic and “1” for diabetic patients. This outcome variable is the focal point of analysis, as it reflects the condition we aim to explain or predict. Its variation is examined in relation to multiple independent variables in the dataset, including demographic attributes (e.g., age, gender), physiological measurements (e.g., BMI, blood pressure), and biochemical indicators (e.g., fasting plasma glucose, cholesterol levels).

This aligns with the definition of a dependent variable in analytics, where it is the measurable result, whose changes are influenced by explanatory variables. In predictive modelling and regression analysis, understanding how independent variables relate to this dependent outcome allows decision-makers to identify high-risk groups, target interventions, and optimize care strategies in diabetes management (Cote, 2021).

This research together with the existing research leverage logistic regression for diabetes prediction because of its interpretability and clinical relevance (Dekamin, A. et al.,2022).Common biomarkers like Glucose levels and the BMI create a foundation for selecting the features for this project (Kavakiotis et al., 2017) and various research papers have prioritized recall and F!-score to handle the class imbalances within the datasets (Tripathi.D, 2022).Additionally, the clustering techniques are used to identify patient subgroups for risk stratification (Yang.,2025).

Prior works rely mostly on public datasets with limited clinical variables (Dekamin, A. et al.,2022), whereas our work combines real time EHR data from Glucowell Medical that include verified clinician features such as medication adherence. Contrarily to studies treating model development as an academic exercise, we integrate SHAP-driven explainability into the clinical workflows, providing risk alerts in EHRs. During the system development, we additionally work in a close partnership with clinicians for the customization of model outputs a gap in almost all existing literature regarding model implementations (Kavakiotis et al., 2017).

Z-score normalization is a technique that transforms data in which each variable has a mean of 0 and a standard deviation of 1. For Glucowell's diabetes dataset where variables like glucose, BMI, age, and blood pressure range differently. This standardization ensures one variable doesn't dominate predictive models simply because of its scale.

After implementing z-score normalization, all variables align on a common scale, making comparisons fair and allowing the model to weigh each factor accurately when predicting diabetes status (Firmansyah et al, 2024). This process removes scale bias, improves the stability and performance of models like logistic regression, and ensures that health indicators are evaluated based on their true predictive power rather than their numerical magnitude (Han et al., 2012).

KNN (K-Nearest Neighbors) imputation fills in missing values by finding similar records and using their values as estimates. In Glucowell's dataset, variables such as insulin or skin thickness might be missing. KNN identifies records with similar glucose, BMI, and blood pressure values and uses their average insulin value to fill any gaps.

It keeps the dataset intact by avoiding row deletions, produces realistic estimations, and maintains data structure integrity. This leads to more reliable model predictions and better-informed clinical decisions in diabetes risk management (Hu et al., 2024).

Logistic regression was employed to predict the likelihood of important outcomes, such as whether a patient might fall into a higher-risk category based on their profile and historical patterns. This allowed the clinic to estimate probabilities for individual patients and prioritize targeted care strategies. Linear regression was used to model continuous outcomes, such as projected changes in health indicators or expected variations in service utilization over time. These insights enabled Glucowell Medical to quantify the potential impact of interventions, allocate resources more effectively, and forecast future needs, ultimately supporting the clinic's mission to improve long-term patient outcomes while maintaining operational efficiency.

The Elbow Method was used to evaluate how many patient clusters best described the data. The idea was to run K-Means clustering for a range of k values and keep on calculating Within-Cluster Sum of Squares (WCSS). The optimal number of clusters k can be defined as the "elbow" point since WCSS decreases considerably at that point when we plot WCSS against the number of clusters (Lugner, M.,2021).

This method balances within-cluster compactness and between-cluster separation, making it less prone to overfitting to highly substructured or outliers. In this project, the method allowed us to explore clinically meaningful patient strata in a similar manner for example dividing the patients based on BMI and fasting plasma glucose levels. Additional validation methods were used for cases where the elbow cannot be clearly observed such as silhouette scores and were considered to ensure robust cluster selection.

After determining the optimal number of clusters, K-means clustering was used to group patients with the similar characteristics into the distinct categories. This technique groups data by minimizing the sum of squared Euclidean distances between points and their cluster centroids, gradually adjusting the position of the centroids step by step up to when there is no change (Alam,.2019). For this project, the K-means proved to be very effective in large datasets of our clinic's patients helping to discover hidden relationships between features like glucose, BMI and so on that are critical to the diabetes risk assessment. It’s speed and efficiency in computations aligned with our projects' need for faster and real time insights.

However, there are limitations such as sensitivity to the fast centroids' placement and reduced performance with non- spherical cluster shapes. These limitations were resolved by running the algorithm many times with different initialization and check the results for any consistency to ensure that the clusters produce both statistical and clinical insights from the dataset.

Methodology

To meet our objectives for this project, the methodology is split into 6 milestones, Data collection, Descriptive analysis, Data cleaning, Comparative ML models for estimating diabetes, K-Means clustering to identify risk levels and Power BI dashboard. The methodology is visualized in the following figure.

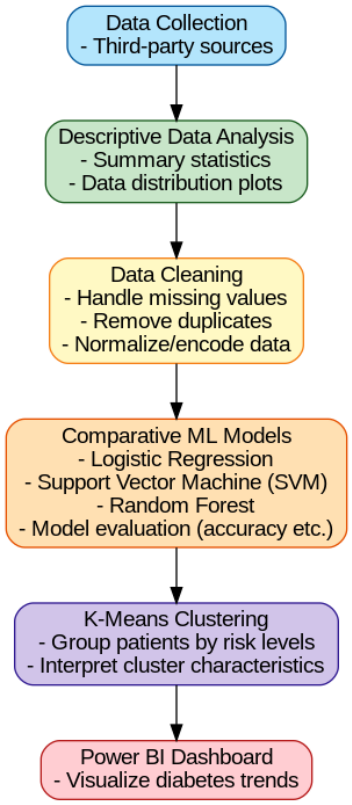

Figure 1 – Methodology Flowchart

Phase 1: Data collection

To meet our objective, we looked for data that contains important health readings that contribute to the development of diabetes, such as BMI, Age, Family history, Drinking history etc. Our data was collected from the Data dryad organization, specifically from a research paper written by authors affiliated with the, Department of Endocrinology, Metabolism Department of Health Examination and, Nantong Center for Chronic and Noncommunicable Disease Control and Prevention. This ensures that our data is reliable and accurate as it is collected by official authorities.

Phase 2: Data exploration

To further understand the data and its story line, data exploration must be done. This can be done through visualizations or statistical description, we must also ensure the quality of the data by identifying, null values, duplicates, imbalances and data distributions. Another important task is figuring out which features will be most relevant for out models, though models can still learn from all the features it is better to provide it with meaningful features to produce the most accurate results. Once exploration is done, the next phase begins, which is data cleaning.

Phase 3: Data cleaning

This phase is where we act on the insights discovered in the exploration phase, duplicates need to be deleted as they can cause biases during training, null values need to be handled by either deletion or imputation as they can negatively impact our training, imbalanced features need to be balanced as they can cause biases towards majority classes which is detrimental when estimating diabetes and finally, the distribution needs to be normalized to maintain a standardized shape across our features. During this phase we will be utilizing the following techniques:

1. KNN imputation
2. Z-score normalization
3. SMOT balancing technique
4. Python

In data analysis an imbalanced dataset the minority class has far less samples than the majority class, which can lead to issues when performing model training as models trained on imbalanced data may develop tendencies to ignore the minority class because correctly predicting the majority class will, most of the time, lead to high accuracy.

SMOT or synthetic minority oversampling does not solve this problem by simply duplicating the samples from the minority class, it generates new synthetic samples by interpolating between existing samples within the majority class. Essentially the process begins by picking a sample within the minority class, find it's k nearest neighbours, randomly choosing one of those neighbours and creating a synthetic sample by "drawing a line" between the two neighbours and picking a random point along that line to generate the new synthetic sample.

The benefit of using SMOT compared to other methods like random sampling is the reduced overfitting, this is due to SMOT generating new samples and not just duplicating existing samples. For these reasons SMOT was utilized to address imbalances in feature that would be indicative of whether a patient was high risk or low risk. By generating realistic synthetic data for underrepresented groups, we ensured that the predictive models could learn from a balanced dataset, improving the clinic's ability to identify patient risk groups easily and early.

Phase 4: Model development

After our data is cleaned the heavy work is done and model development can start, Our aim is to train several models to find out which model is more suitable to estimate diabetes. We chose to train 3 separate models Logistic Regression (LR), Support Machine Vector (SVM) and Random Forest Classifier (RF). The models will be run on default for the first stage and in stage 2 the hyper parameters will be tuned to further enhance the model's performance. Three performance metrics will be utilized, Recall, Precision and F1-Score. Recall will be the most important as we need to ensure our model will not falsely identify diabetes (Type II error). Our fourth model is K-Means clustering, we will create clusters based on the most relevant features to identify which patients fall in high-risk zones and low-risk zones, using the cluster we will create a risk assessment score for our patients.

Logistic regression is a statistical model that is used for classification tasks, where the goal is to predict the probability of a categorical outcome. Compared to linear regression which produces continuous numerical predictions, logistic regression outputs probabilities between 0 and 1 using the sigmoid function. Using a threshold, most commonly 0.5, these probabilities are converted to class predictions.

By minimizing log loss, also known as cross-entropy, logistic regression can estimate coefficients that best fit the training data and thus prevent overfitting. Through these coefficients the influence of each feature on the predicted probabilities can be interpreted, this makes logistic regression a popular choice when interpretability is crucial.

Phase 5: PowerBI dashboard

After developing our models, we will use the dataset to create a PowerBI dashboard to visualize the trends amongst our data and extract insights that we act on to further develop our business strategy.

Result and Analysis

This section presents the findings of the data analysis phase to address the objectives of this estimating diabetes occurrence and risk. This dataset contained clinical and demographics variables related to diabetes occurrence. Analysis was carried out to extract insights from the dataset, identify correlation between variables and prepare training and testing sets for machine learning models.

The analysis consists of the following stages:

1. Descriptive analysis
2. Comparative ML models
3. Clustering to establish risk score
4. PowerBI dashboard

Descriptive Analysis:

The analysis begins with data exploration, it is important that we fully understand our data before we start cleaning; this includes knowing its shape, no.of features, duplicates, null values and data distribution. Our chosen method of exploration is python, with the use of the Pandas and NumPy libraries we were able to easily handle our data. Our data's shape was (211833, 25) this means we had more than 211000 rows of data entries and 25 features, this amount of data is ideal for model training as it offers a variety of inputs for our model to learn from~~.~~, However, this will depend on the quality of said data~~, a~~. In the following figure we can see the count of missing values in all the columns, it ranges from 2 null values to then thousands and then hundreds of thousands. Having missing values in the thousands range can still be handled without needing to delete them especially when the dataset is big enough, however when it reaches the hundreds of thousands as we can see in the smoking and drinking history.

```
id                                                                  0
Age (y)                                                             0
Gender(1, male; 2, female)                                          0
site                                                                0
height(cm)                                                          2
weight(kg)                                                          0
BMI(kg/m2)                                                          0
SBP(mmHg)                                                          23
DBP(mmHg)                                                          24
FPG (mmol/L)                                                        0
Cholesterol(mmol/L)                                              4854
Triglyceride(mmol/L)                                             4887
HDL-c(mmol/L)                                                   94562
LDL(mmol/L)                                                     93421
ALT(U/L)                                                         1782
AST(U/L)                                                       123290
BUN(mmol/L)                                                     21551
CCR(umol/L)                                                     11175
FPG of final visit(mmol/L)                                         19
Diabetes diagnosed during followup (1,Yes)                     210529
censor of diabetes at followup(1, Yes; 0, No)                       0
year of followup                                                    0
smoking status(1,current smoker;2, ever smoker;3,never smoker)  151603
drinking status(1,current drinker;2, ever drinker;3,never drinker)  151603
family histroy of diabetes(1,Yes;0,No)                              0
dtype: int64
```

Figure 2 - null value count

In figure 3 we can a statistical summary of our features, this includes the central tendency, dispersion and distribution of the shape of our dataset. From the summary we can conclude several facts, the mean age of our sample is 42 years with 50% of patients being below 39 years of age, a mean BMI of 23.2 which falls in the normal range with an outlier of 52.7 which falls in the morbidly obese category which indicates that this individual has a higher chance of developing diabetes. Diabetes incidence comes at 1.97% this indicates a serious imbalance in the diabetes feature, this would need to be addressed in the data cleaning stage. Overall, this tell us that the sample consisted of middle-aged individuals with relatively healthy markers with several outliers that need to be addressed in the next stage.

	count	mean	std	min	25%	50%	75%	max
id	211833.0	343114.275594	197768.164898	1.000000	171837.000000	343120.000000	514482.000000	685286.000000
Age (y)	211833.0	42.097567	12.649956	20.000000	32.000000	39.000000	50.000000	99.000000
Gender(1, male; 2, female)	211833.0	1.451818	0.497674	1.000000	1.000000	1.000000	2.000000	2.000000
site	211833.0	5.846596	3.068889	0.000000	3.000000	5.000000	8.000000	16.000000
height(cm)	211831.0	166.432212	8.326936	106.500000	160.000000	166.400000	172.500000	198.500000
weight(kg)	211833.0	64.677388	12.223239	32.000000	55.200000	63.500000	72.800000	157.000000
BMI(kg/m2)	211833.0	23.235742	3.342934	15.000000	20.800000	23.000000	25.380000	52.700000
SBP(mmHg)	211810.0	119.063467	16.379803	59.000000	107.000000	118.000000	129.000000	222.000000
DBP(mmHg)	211809.0	74.177178	10.813487	38.000000	66.000000	73.000000	81.000000	164.000000
FPG (mmol/L)	211833.0	4.915240	0.611792	0.590000	4.530000	4.900000	5.300000	6.990000
Cholesterol(mmol/L)	206979.0	4.706962	0.900628	0.020000	4.080000	4.620000	5.230000	17.840000
Triglyceride(mmol/L)	206946.0	1.339086	1.032368	0.000000	0.730000	1.070000	1.610000	32.640000
HDL-c(mmol/L)	117271.0	1.372051	0.307458	0.000000	1.160000	1.350000	1.560000	10.400000
LDL(mmol/L)	118412.0	2.768231	0.680866	0.000000	2.290000	2.700000	3.160000	10.070000
ALT(U/L)	210051.0	23.954027	22.127355	0.000000	13.000000	18.000000	27.500000	1508.400000
AST(U/L)	88543.0	24.084889	12.362870	0.000000	18.600000	22.000000	26.700000	1026.200000
BUN(mmol/L)	190282.0	4.657683	1.186493	0.100000	3.810000	4.530000	5.370000	29.400000
CCR(umol/L)	200658.0	70.065176	15.802667	13.000000	58.000000	69.300000	80.900000	1116.600000
FPG of final visit(mmol/L)	211814.0	5.133327	0.687021	3.100000	4.730000	5.060000	5.400000	29.700000
Diabetes diagnosed during followup (1,Yes)	1304.0	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
censor of diabetes at followup(1, Yes; 0, No)	211833.0	0.019704	0.138982	0.000000	0.000000	0.000000	0.000000	1.000000
year of followup	211833.0	3.122593	0.938764	2.001369	2.162902	2.989733	3.945243	7.564682
smoking status(1,current smoker;2, ever smoker;3,never smoker)	60230.0	2.556550	0.804845	1.000000	3.000000	3.000000	3.000000	3.000000
drinking status(1,current drinker;2, ever drinker;3,never drinker)	60230.0	2.806442	0.448283	1.000000	3.000000	3.000000	3.000000	3.000000
family histroy of diabetes(1,Yes;0,No)	211833.0	0.020507	0.141726	0.000000	0.000000	0.000000	0.000000	1.000000

Figure 3 – Statistical Summary

Next up is a skewness and kurtosis test, in figure 4 we can see the results for the skewness and kurtosis test for all the features, from this tests we concluded that some variables such as the Trigliceride, ALT, AST, FFPG had extreme skewness and kurtosus/leptokurtosis, this indicates the presence of outliers and/or imbalances.

Skewness and Kurtosis:	skew	kurtosis
id	-0.000701	-1.200925
Age (y)	0.921103	0.253666
Gender(1, male; 2, female)	0.193630	-1.962526
site	0.707517	-0.188526
height(cm)	0.055236	-0.411905
weight(kg)	0.587487	0.368618
BMI(kg/m2)	0.547506	0.554012
SBP(mmHg)	0.657875	0.865872
DBP(mmHg)	0.576120	0.821224
FPG (mmol/L)	0.078520	0.596464
Cholesterol(mmol/L)	0.688767	1.755458
Triglyceride(mmol/L)	4.602340	48.957267
HDL-c(mmol/L)	0.688774	7.431497
LDL(mmol/L)	0.754213	1.910380
ALT(U/L)	10.718198	366.872570
AST(U/L)	20.531661	1094.833297
BUN(mmol/L)	0.819465	3.281956
CCR(umol/L)	3.419559	166.963514
FPG of final visit(mmol/L)	4.111924	50.795839
Diabetes diagnosed during followup (1,Yes)	0.000000	0.000000
censor of diabetes at followup(1, Yes; 0, No)	6.911684	45.771808
year of followup	0.533124	-0.753679
smoking status(1,current smoker;2, ever smoker;...	-1.337687	-0.115695
drinking status(1,current drinker;2, ever drink...	-2.266847	4.501503
family histroy of diabetes(1,Yes;0,No)	6.766542	43.786501

Figure 4 – Skewness and Kurtosis

After conducting the skewness and kurtopsis test we found out that some features do have several outliers, so next we decided to count the number of outliers in each feature which will help us grasp a better understanding of it. In figure 5 we can see the count of outliers in each column, most notably Triglyceride, FFPG, ALT and AST.

```
Number of outliers per column: Age (y)
site                                                                247
height(cm)                                                          202
weight(kg)                                                         1584
BMI(kg/m2)                                                         2061
SBP(mmHg)                                                          2829
DBP(mmHg)                                                          2236
FPG (mmol/L)                                                       4489
Cholesterol(mmol/L)                                                3306
Triglyceride(mmol/L)                                              11563
HDL-c(mmol/L)                                                      1576
LDL(mmol/L)                                                        2001
ALT(U/L)                                                          15309
AST(U/L)                                                           4619
BUN(mmol/L)                                                        2547
CCR(umol/L)                                                         678
FPG of final visit(mmol/L)                                         6730
censor of diabetes at followup(1, Yes; 0, No)                      4174
year of followup                                                      1
smoking status(1,current smoker;2, ever smoker;3,never smoker)    14634
drinking status(1,current drinker;2, ever drinker;3,never drinker)10307
family histroy of diabetes(1,Yes;0,No)                             4344
```

Figure 5 – Outlier Count

We can better see the outlier distribution of the ALT and triglycerides in figure 6, where we plotted boxplots to better demonstrate the severity of these outliers. We can clearly see the amount of outliers is extreme and would require serios handling before model training begins.

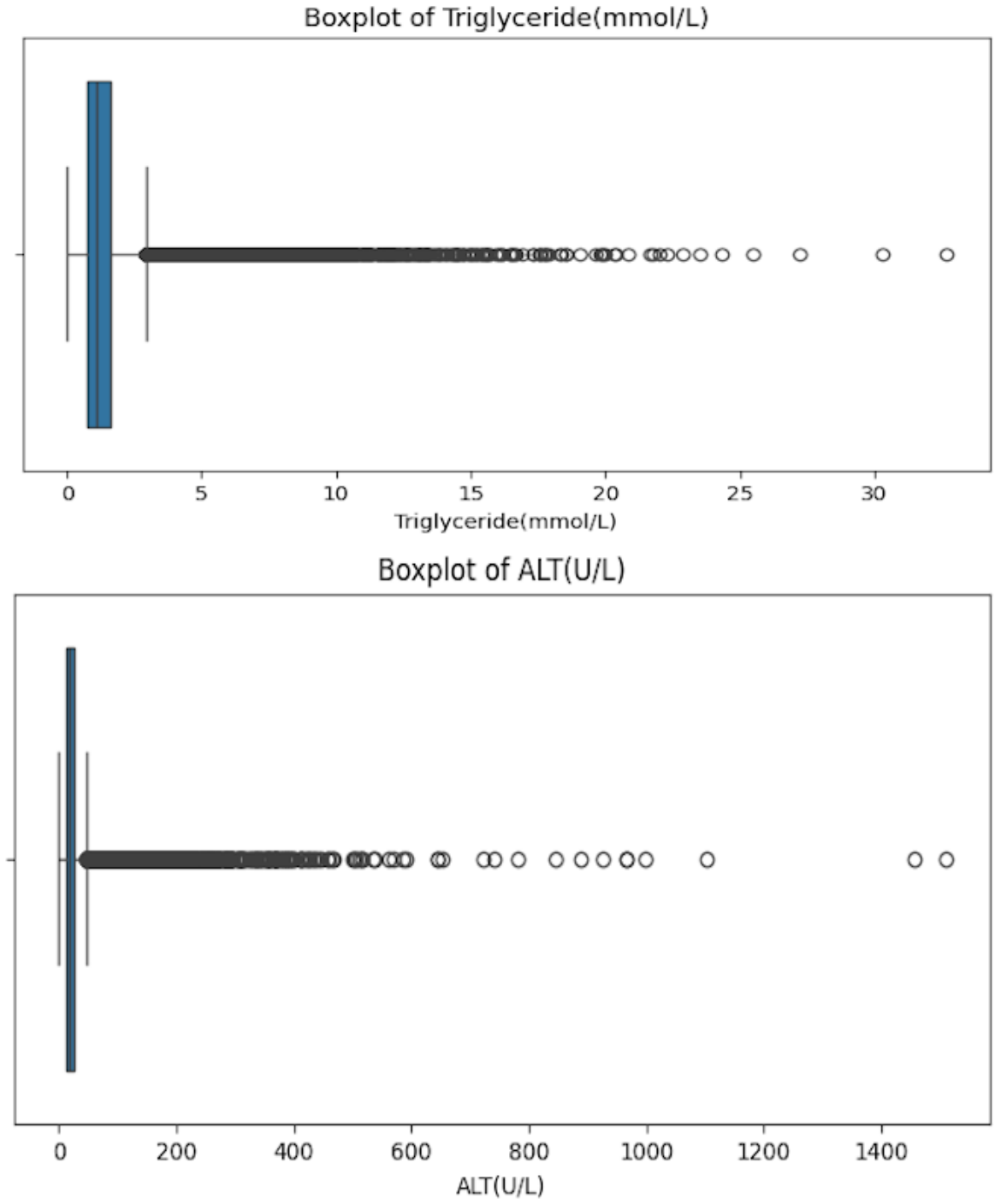

Figure 6 – ALT and Triglyceride outlier boxplots

Our dataset also had 0 duplicates which is a good sign as it indicates that our dataset is authentic, the last step of exploration is to figure out which of the 24 variables is truly correlated to diabetes occurrence, by having cutting down on the number of features it makes our model training less intensive and more efficient. To achieve this we used a correlation heatmap to see how all the features correlate to diabetes. This is visualized in figure 7. The legend on the right side tells us which colour corresponds to which level of correlation. In the next section data cleaning we will discuss how we used these insights to clean the data.

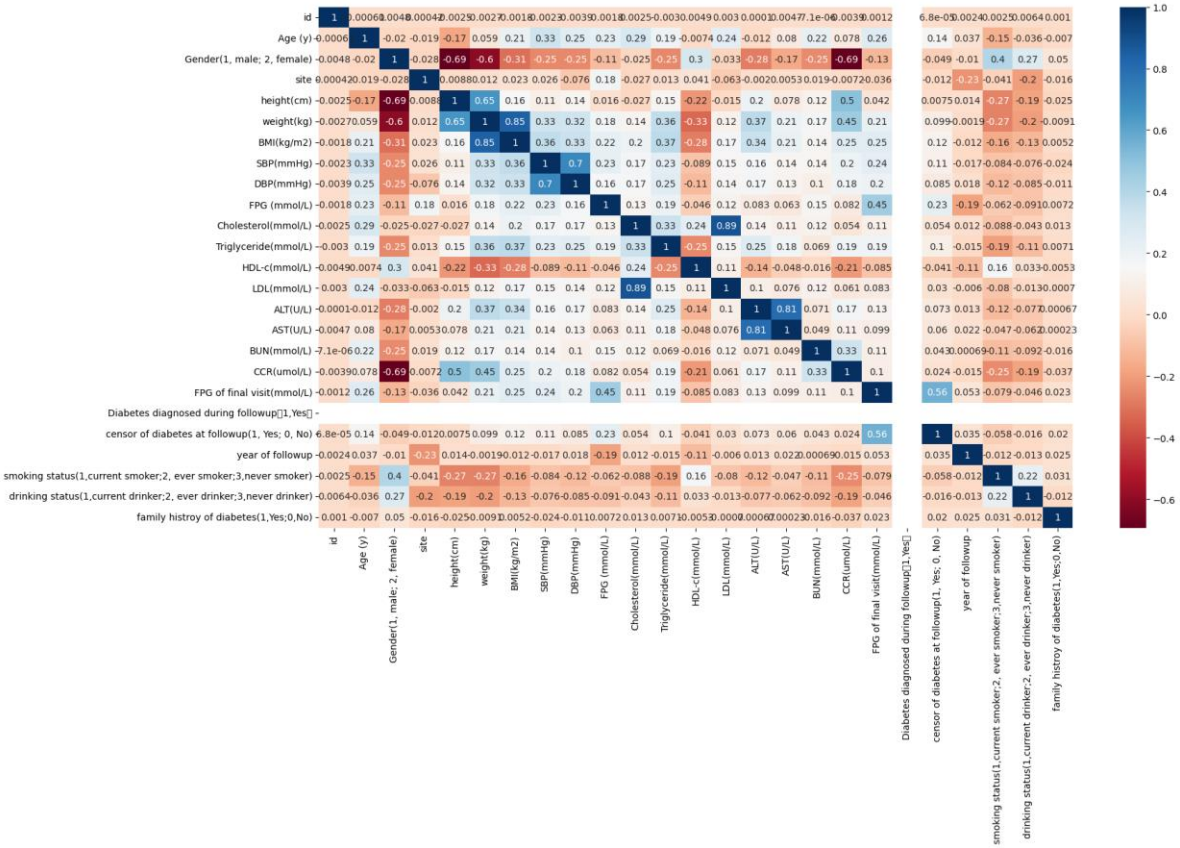

Figure 7 – Correlation heatmap

Comparative ML models:

In this section we will discuss the results of model training, as discussed earlier 3 ML models were trained on our cleaned data. As previously mentioned in our methodology section, the 3 models we have chosen to do a comparison of are a logistic regression (LR), Support Machine Vector (SVM) and Random Forest Classifier (RF) models.

The end goal of performing this comparison is to assess each model's ability to accurately estimate whether a patient has diabetes, which would have direct implications for potential improvements to patient care, operational efficiency and decision making for Glucowell Clinic. Through this evaluation we can identify not only the most technically effective model but also the one that best aligns with the clinic's business objectives.

To objectively evaluate the performance of our models we chose to use 3 key metrics, being precision, recall and F1-score, we've also created a confusion matrix for each model to better display its performance. To briefly summarise what each of the performance metrics mean:

- Precision is the proportion of samples that were estimated to be positive and were positive. Calculated as the number of true positive predictions (TP) divided by total positive predictions (TP + FP).
- Recall is the proportion of correctly identified positive samples, calculated as TP divided by TP + FN
- F1-Score is the harmonic mean of precision and recall, calculated as precision multiplied by recall with the result divided by precision + recall and the result after division is multiplied by 2.

The logistic regression model achieved the following results:

- Precision: 0.244
- Recall: 0.931
- F1-Score: 0.387

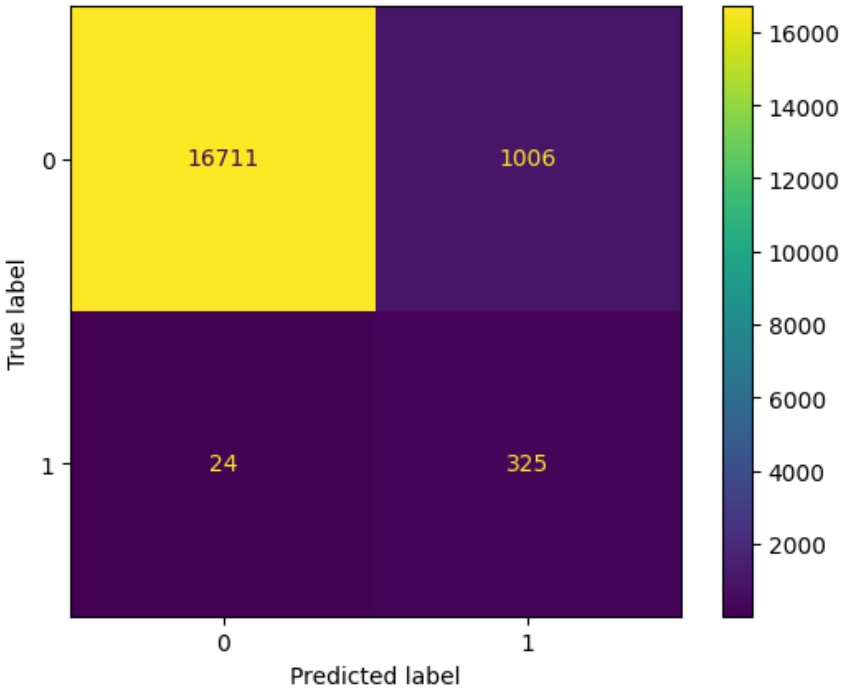

Figure 8: Confusion Matrix for Logistic Regression Model

Based on these results we can see that the model achieved exceptionally high recall, which can be interpreted as it being able to successfully identify many patients who have diabetes. However, its low precision means that there is a high amount of false positive cases.

The support vector machine model achieved the following results:

- Precision: 0.284
- Recall: 0.923
- F1-Score: 0.435

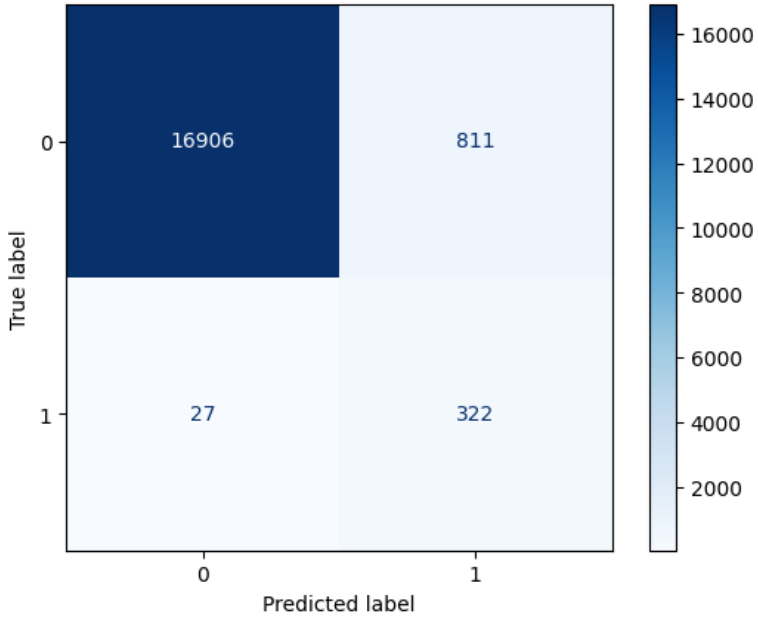

Figure 9: Confusion Matrix for Support Vector Machine model

As these results show, the SVM model offered slightly better precision compared to our logistic regression model, while maintaining similarly high recall. While this balance does indicate a slight improvement to the proportion of correctly estimated positive samples, the false positive rate is still concern.

Lastly of our machine learning models, the random forest model achieved the following results:

- Precision: 0.883
- Recall: 0.822
- F1-Score: 0.852

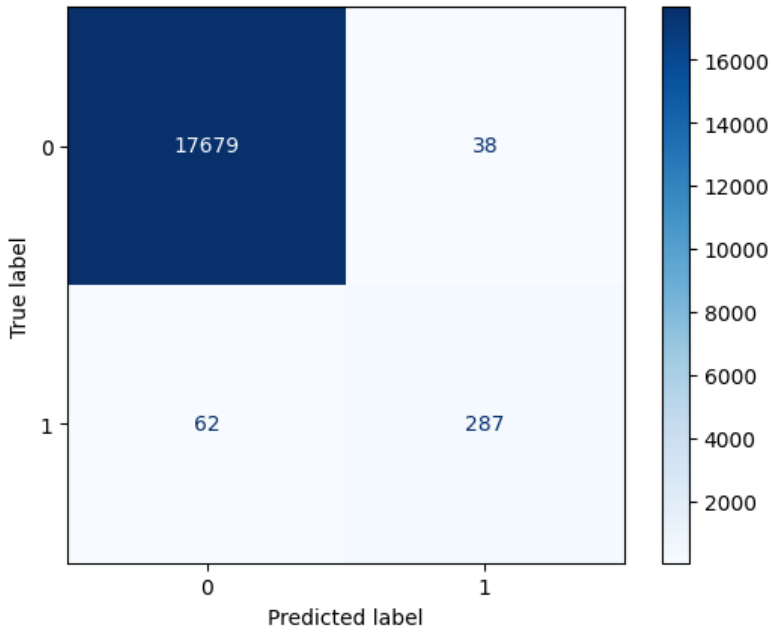

Figure 10: Confusion Matrix for Random Forest Model

The random forest model was able to achieve the highest performance overall with good precision and recall. This informs us that not only was the model capable of correctly diagnosing patients with diabetes, but it also did not misdiagnose the majority of patients who did not have diabetes.

Based on these results, we can say that the Random Forest model performed the best and was the most balanced in terms of correctly estimating a patients diabetes status. This does not mean however that the other models are completely useless to the clinic, depending on the application if the clinic priorities the ability to catch the greatest number of diabetes cases at the expense of a potentially high number of false negatives, the SVM model may be preferred. This situation is unlikely and the Random Forest should be chosen in the majority of applications.

Choosing the right predictive model is more than a technical decision, it will directly influence the diabetes clinic's ability to deliver effective patient care and manage resources efficiently. Assisting with direct patient diagnosis is not the full extent of where these models could be implemented either, they can be integrated into broader business analytic workflows. One such case could be the forecasting of patient inflow, allowing the clinic to properly plan for staffing needs and potential appointment scheduling. Aligning model selection with operational priorities ensures that the predictive analytics support clinical decisions and drive strategy efficiency.

Clustering for Risk Assessment

Moving onto the next section, we will discuss the results of our clustering for risk assessment, as mentioned previously we have used K-Means clustering for this assessment. In the case of Glucowell clinic, K-Means was utilized to segment patients into high and low risk categories for developing or worsening diabetes. These clusters allow the clinic to create a risk assessment score for each patient, enabling more targeted interventions.

The selection of features for clustering was guided by both domain knowledge from diabetes research and EDA, variables with strong clinical relevance and statistical correlation to diabetes risk were prioritized.

The K-Means algorithm was applied to selected patient features to group individuals into distinct clusters based on health characteristics. Based on our results from the Elbow Method, we determined that K=2 would provide us with a meaningful segmentation, allowing us to group patients into high and low risk groups.

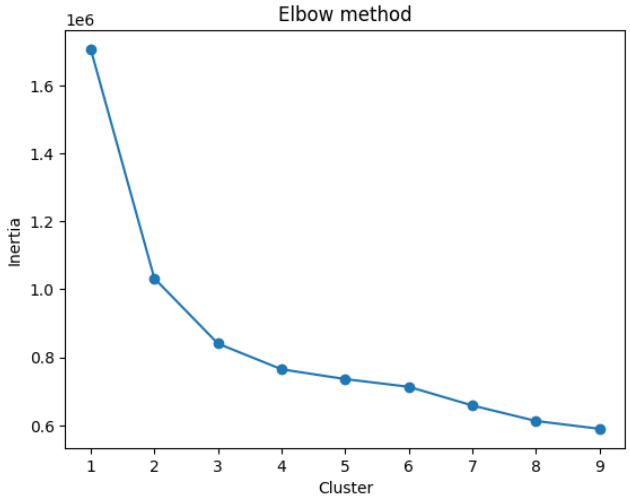

Figure 11: Elbow method graph, indicating ideal number of clusters

The cluster with the higher mean value was designated as the high-risk group, while the other was labelled as low risk. This interpretation confirmed that the algorithm had effectively segmented patients by their underlying diabetes risk profile. Cluster quality was further evaluated using the Silhouette Score, which provided a quantitative measure of how well each patient fit within their assigned group compared to other clusters. The resulting score suggested a reasonable degree of separation, supporting the validity of the clustering results.

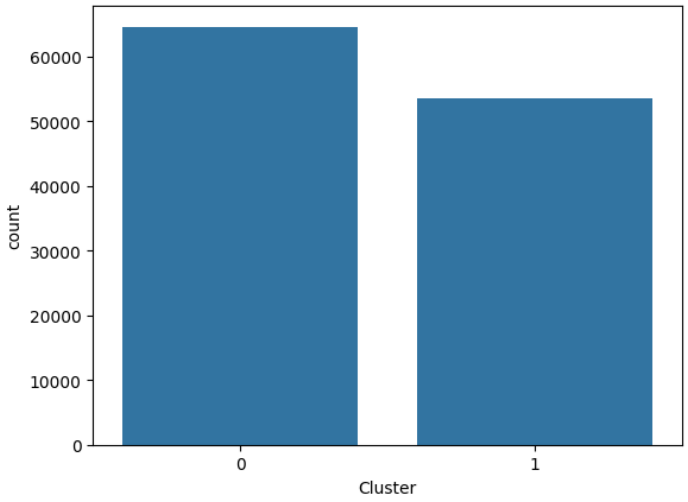

Figure 12: Patient distribution between clusters

```
Cluster
0    0.105749
1    0.976004
```

Figure 13: Probability of patient falling into either cluster

Cluster quality was further evaluated using the Silhouette Score, which provided a quantitative measure of how well each patient fit within their assigned group compared to other clusters. The resulting score suggested a reasonable degree of separation, supporting the validity of the clustering results.

```
Silhouette Score: 0.3571
Cluster risk means:
Cluster
0    0.105749
1    0.976004
```

Figure 14: Silhouette Score evaluation of clusters

Now that we have created the clusters and the patient risk groupings, the clinic will be able to develop a practical risk assessment score to use a guide for clinical decision making. Patients assigned to the high-risk cluster received a higher risk score, while those in the low-risk cluster were assigned a lower score. The integration of a risk assessment score into the clinic's patient management system would allow for improvements in proactive follow-up appointment scheduling, preventative screenings and enable the clinic to initiate early interventions for high-

risk patients. Over time, these scores can be further improved with new patient data, ensuring that the assessment remains dynamic and up to date.

Power BI dashboard analysis:

This dashboard presents a structured overview of key diabetes metrics which includes, BMI, fasting glucose levels, lipid profiles and lifestyle statuses among Chinese patients. Firs the datasets where uploaded into PowerBI and a relationship was established using the primary key PatientID, as demonstrated in figure 15. Several new columns were calculated using DAX functions to extract more insights out of the data and further enhancing the dashboard's usefulness. In this section we will analyse the dashboard to identify trends and possible areas for clinical interventions. Figure 16 is a screenshot of the dashboard created using Power BI.

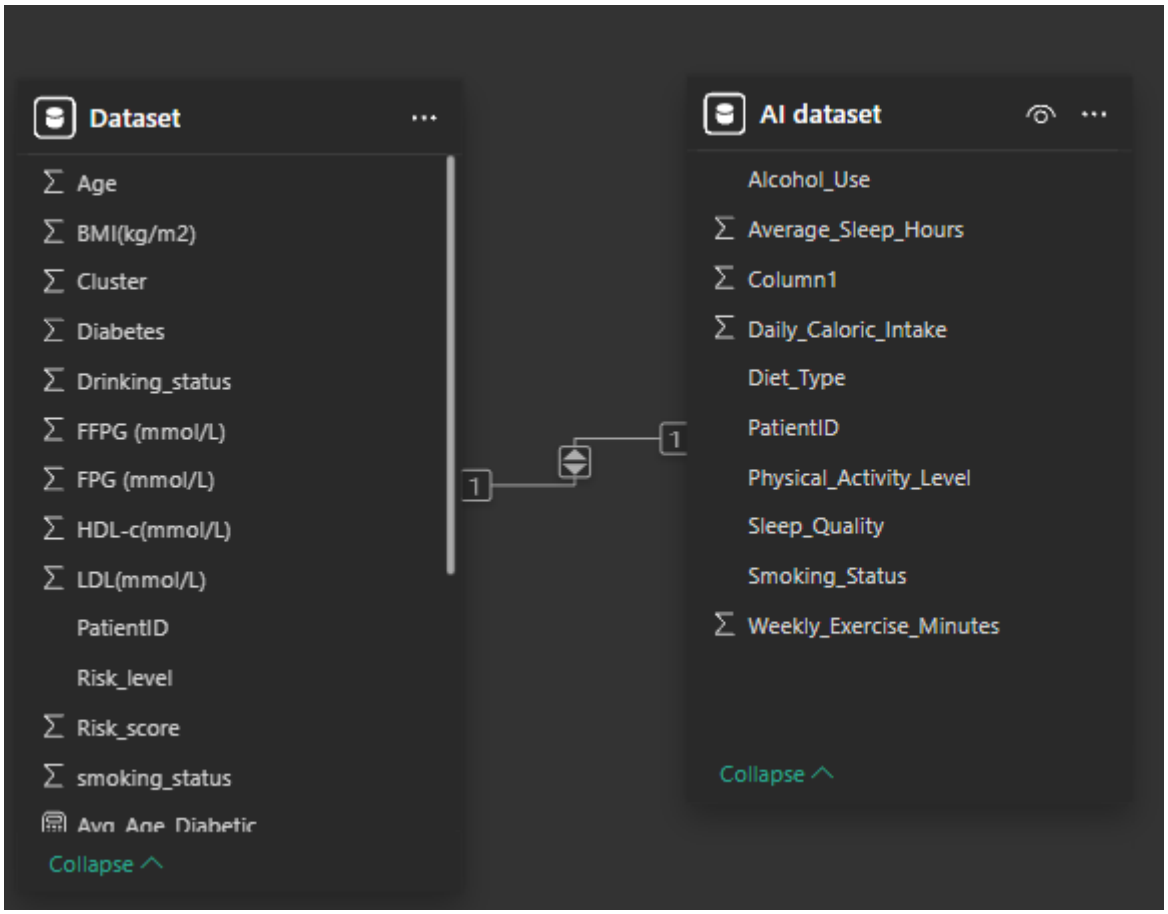

Figure 15 one to one relationship using PatientID

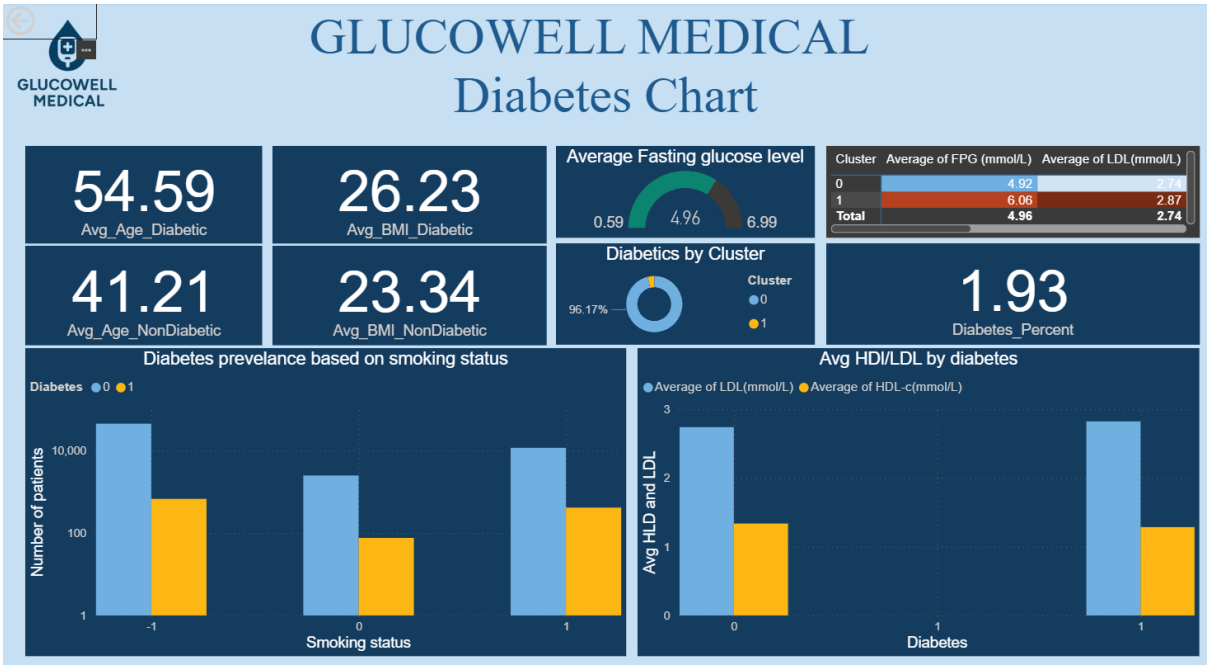

Figure 16 – Glucowell medical dashboard

Key findings include:

1. Diabetic patients had a significantly higher mean age in comparison to non-diabetic patients, with a 13.38 age difference, this indicates that Diabetes in more prevalent in individuals' with advanced age.
2. Diabetic individuals have an elevated level of fasting glucose 1.14 more than the non-diabetics, making fasting glucose level a strong indicator of diabetes.
3. The average BMI for diabetic patients is elevated by around 3 $kg/m^{2,}$ this aligns with the association between elevated BMI and increased diabetes risk, with the diabetic individuals falling into the overweight range.
4. Diabetic patients showed slightly lower readings in HDL levels and slightly higher readings in the LDL levels, this indicates a less favourable lipid profile which can be correlated to elevated diabetes risk.
5. Across the smoking categories, (-1 being non-smoker, 0 being ever smoked, and 1 being active smoker) there is not a significant difference but the current and former smoker groups slightly outnumber the non-smoker groups, identifying a potential relationship between smoking and diabetes risk.

Conclusion

Glucowell medical's shift is mainly to understand and manage diabetes, in which it's heavily focused on data driven decision making. The findings from this study demonstrated that the combination of advanced analytical methods and visual analytics can transform thousands of data into meaningful insights. Among the machine learning models the random forest achieved the most consistent results with a good balance between precision and recall (0.883 and 0.822 respectively) making it the most suitable choice for Glucowell's business operation. While the SVM and logistic regression models offered high recall their low precision performance makes them unsuitable for our diagnostic operations. The clustering technique effectively segmented the patients into high- and low- risk profiles, based on clinically relevant features, this clustering is the backbone of our dynamic risk assessment score that can aid in early intervention. Insights from the dashboard reinforced our findings, highlighting that diabetes is more prevalent in older individuals, individuals with less ideal lipid profiles, smoker groups and higher fasting glucose levels. Ultimately these findings underline the potential benefits of integrating AI modelling and dashboards into out the clinic's workflow to improve patient's care and enable data driven decision making.

References

Cote, C. (14 december, 2021). *What Is Regression Analysis in Business Analytics?* Retrieved from HBS Online: https://online.hbs.edu/blog/post/what-is-regression-analysis

Das, R., Suneela, B., Akula, D., Saurav, S., Tyagi, S., & Rani, D. E. (2025). Integrating AI-Driven data analytics into healthcare business Models: A Multi-Disciplinary approach. *Journal of Neonatal Surgery*, *14*(5S), 761–773. https://doi.org/10.52783/jns.v14.2126

Dekamin, A. (2022). Management of Type 2 Diabetes--Applications of Machine Learning and Electronic Medical Records-based Analytics (Doctoral dissertation, Toronto Metropolitan University).

D. Tripathi, S. k. Biswas, S. Reshmi, A. N. Boruah and B. Purkayastha, "Diabetes Prediction Using Machine Learning Analytics: Ensemble Learning Techniques," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-7, doi: 10.1109/ASIANCON55314.2022.9908975.

Lugner, M., Gudbjörnsdottir, S., Sattar, N., Svensson, A. M., Miftaraj, M., Eeg-Olofsson, K., ... & Franzén, S. (2021). Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study. *Diabetologia*, *64*(9), 1973-1981.

Gavrilov, G., Vlahu-Gjorgievska, E., & Trajkovik, V. (12 september , 2020). *Healthcare data warehouse system supporting cross-border interoperability.* Retrieved from Health Informatics Informatics JournaL (SAGE Journals): https://journals.sagepub.com/doi/10.1177/1460458219876793

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

KMS Healthcare. (10 january , 2024). *Data integration in healthcare: Guide & best practices.* Retrieved from KMS Healthcare: https://kms-healthcare.com/blog/data-integration-in-healthcare/

Munjala, M. B. (2025). Harnessing the Power of Data Analytics and Business Intelligence to Drive Innovation in Biotechnology and Healthcare: Transforming Patient Outcomes through Predictive Analytics, Genomic Research, and Personalized Medicine. *Cuestiones de Fisioterapia*, *54*(3), 2222–2236.

Siow, S. J. (2023). *Business Intelligence Data Visualization for Diabetes Health Prediction. International Journal of Advanced Computer Science and Applications, 14(1).* Retrieved from Business Intelligence Data Visualization for Diabetes Health Prediction. International Journal of Advanced Computer Science and Applications, 14(1).: https://thesai.org/Publications/ViewPaper?Volume=14&Issue=1&Code=IJACSA&SerialNo=90

World Health Organization. (2023). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes

Yang, X., & Li, J. (2025). A clustering-based federated deep learning approach for enhancing diabetes management with privacy-preserving edge artificial intelligence. Healthcare Analytics, 7, 100392.

Zargoush, M., Ghazalbash, S., Hosseini, M. M., Alemi, F., & Perri, D. (2025). Machine learning driven diabetes care using predictive-prescriptive analytics for personalized medication prescription. *Scientific Reports*, *15*(1).

Firmansyah, M. R., & Astuti, Y. P. (2024). Stroke classification comparison with KNN through standardization and normalization techniques. Advance Sustainable Science, Engineering and Technology (ASSET), 6(1). https://doi.org/10.26877/asset.v6i1.17685

Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.

Hu, YH., Wu, RY., Lin, YC. et al. A novel MissForest-based missing values imputation approach with recursive feature elimination in medical applications. BMC Med Res Methodol ***24****, 269 (2024). https://doi.org/10.1186/s12874-024-02392-2*