Hello everyone. We are Group 10 from Z139. Today we will be presenting on our Diabetes Prediction System for our SC1015 Mini-Project.
Our group members consist of myself Nanditha, Madhumitha and Praveena.
Over the next 10 minutes, we will guide you through the various components of our diabetes prediction system project.
This will be the flow of our presentation.

Before we jump straight into the details of our diabetes prediction system, let me explain why we decided on this problem in the first place.
Diabetes is a serious global concern affecting many countries worldwide with the World Health Organization identifying it as a global epidemic. They projected that by 2030, diabetes will be the 7th leading cause of death, highlighting its severity and fatality. Hence it is crucial to take proactive measures in preventing this chronic illness.
Therefore, the objective of our project is to develop and determine the best machine learning model that accurately predicts the likelihood of an individual developing diabetes based on various input factors such as glucose and BMI. This prediction can then guide further medical evaluation and treatment.
Next, let's take a look at the data collection process.

We obtained three diabetes datasets from Kaggle.
From the Diabetes Health Indicators Dataset, we concatenated two sub-datasets to create the Big Dataset. Big Dataset has a sample size of approximately 320,000 data points and 22 variables, which are mostly categorical.
Next, from the Diabetes Data Set, which we named Small Dataset has 2000 data points and 9 variables, which are mostly numerical.
Both Big and Small datasets have a binary response variable indicating the presence of diabetes, and we chose to work with both datasets to include a combination of numerical and categorical data in our prediction system.

Moving on to our data preparation and cleaning process,
We began by dropping irrelevant variables from the Big Dataset and removing any duplicate data points from both datasets. Following this, we removed data points with values of 0 for numerical variables such as Glucose, as they are unfeasible.
Next, we randomly sampled 700 data points from the Big Dataset and 300 data points from the Small Dataset and appended them to create the Joint Data with 1000 data points and 18 variables.
Due to the merger of multiple datasets, there were several null values present in the data. Therefore, we utilised KNN imputation, which imputes missing values by replacing them with the mean or median value of its K nearest neighbours, to replace the null values in our data.
This is just an overview of our finalised variables and data points after cleaning.

One point to be noted is that since our project uses random sampling of data points, there can be slight differences in results and analysis as data points differ for every run. To address this, the code was run multiple times and the analysis is based on the consistency of results across different runs.

Moving on, we proceeded to do some exploratory data analysis and data visualisation on our cleaned dataset in order to identify the variables that show the highest association with our target_response variable, diabates_binary, for subsequent analysis.

In the Univariate Data Exploration section, we used various techniques to explore the characteristics of each variable individually. For numerical variables, we used summary statistics, box-plots, histograms, and violin plots to understand their central tendency, dispersion, and distribution shape. For categorical variables, we used count plots to show the frequency or proportion of each category.

In the Bi-Variate Data Exploration section, we analysed the relationship between each variable and diabetes_binary. For numerical variables, we used swarm plots, joint plots and histograms, and observed any patterns or trends in the data points. For categorical variables, we used count plots, heatmaps, point plots, and chi-square tests to compare the frequency or proportion of each category for different levels of the diabetes_binary.

For Multivariate Data Exploration, we used a correlation heatmap to explore the relationships between all possible pairs of numeric variables. We focused on the correlation strength and direction between the different numeric variables and diabetes_binary, allowing us to identify the extent of association between those variables.

Based on our EDA, we identified the variables that show the highest association with diabetes_binary.
For numerical variables, the correlation heatmap revealed that Glucose, BMI, and SkinThickness have the highest correlation with diabetes_binary.
For categorical variables, the chi-square test showed that DiffWalk (which stands for difficulty walking), HeartDiseaseorAttack, and HighChol (which stands for high cholesterol) have the highest association with diabetes_binary. These three variables have the highest chi-square statistic and lowest p-values as compared to the other categorical predictor variables.
This suggests that these six variables may be strong predictors of diabetes and thus, we selected these variables for our machine learning models.

Now, let's move on to discuss our machine learning process.
Based on our selected predictor variables, we decided to use logistic regression model, random forest, and support vector machine model (which is also known as SVM) for our diabetes prediction system.
We chose these three models as they can handle the selected predictor variables which includes categorical binary variables and numerical variables. They can also handle the binary response variable.

After selecting the models, we then trained and evaluated them to determine which one performs the best in predicting diabetes.
Firstly we used a logistic regression model. Logistic regression estimates the probability of an event using a logistic function and is commonly used for binary classification. It uses an optimization algorithm to fit the model parameters and can predict the probability of having diabetes based on predictor variables.

This slide shows the image of the logistic regression model that we built.In our case, we used all six of our predictor variables to build a multivariate logistic regression model that estimates the probability of having diabetes.

Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to create a powerful predictive model. It uses subsets of data and features to create decision trees and aggregates their predictions for the final outcome. It is also suitable for classification and regression tasks.
This is the random forest tree generated for our numerical and binary predictor variables. A more clearer and detailed view of the tree can be seen in our code. For numerical variables the model splits data into intervals and creates decision trees based on the intervals. For binary variables, a threshold of 0.5 is used to decide sample class membership.

SVM is an algorithm that finds a hyperplane that maximises the margin between data points of different classes, which is the distance between the hyperplane and the closest data points of each class.
Once the SVM model is trained, when we input new data, it calculates the position of the data point relative to the hyperplane. Based on which side of the hyperplane the data point lies, the SVM model makes a classification decision on whether the person has diabetes or not.

After developing our models, we tested them with a random set of values for the predictor variables and all of our three models showed consistent results as seen in the slide.

After training the three models, we evaluated the accuracy and effectiveness of each of our models in predicting diabetes, mainly using confusion matrices for test and train data.

From this slide we can see that the logistic regression model has a higher accuracy on the train data compared to the test data, indicating the need for improvement in the model's performance on the test data.
Detailed evaluations using other methods such as computation of f1 score and deviations can be seen in our code
Random Forest model has a TPR of 0.907 and an FNR of 0.093, indicating that there is still room for improvement in predicting positive samples. The model performs well on the train data but has a lower accuracy and TPR on the test data..
Additionally, classification reports were also created for both train and test data, which are available in our code.
The SVM model has low accuracies and extremely low TPR and high FNR on both the train and test data, making it the least suitable model for diabetes prediction.

To choose the best model for our diabetes prediction system, we have chosen to evaluate and compare the accuracy and False Negative rate (FNR) across the three models.
Accuracy is the proportion of correct predictions to the total number of predictions, and is crucial for predicting a chronic condition like diabetes.
FNR is important in diabetes prediction, as it indicates how many individuals with diabetes are being missed by the prediction system which may lead to dire consequences.

Comparing all three machine learning models based on these two important factors, we

have concluded that the logistic regression model shows the best prediction for diabetes as it shows the highest accuracy score, and a generally lower FNR for train and test data,

While the logistic regression model may be accurate in predicting negative samples, this may not be applicable for positive samples. Based on our evaluations, we believe that the model has overfitted to the data, resulting in poorer performance on the test data.
There are various ways to improve the performance of the logistic regression model. Some recommendations are feature selection, ensembling, regularisation and data augmentation.

In conclusion, for this mini-project, we were able to achieve the objective of our project by determining and developing a machine learning model that could accurately discern the possibility of a person developing diabetes.
Through this project, we explored new ways of visualising and predicting data, as shown on the screen here, including implementing new machine learning techniques like Random Forest.
The necessity of model evaluation was also highlighted to us during this project. From our model evaluations, it can be seen that although the Logistics Regression model is a better prediction system than the other two, all three models were able to output the same correct prediction. Therefore, we learnt that understanding the process of getting accurate predictions is more important than just getting the right outcome.
With that, we hope you enjoyed our presentation on our diabetes prediction system. Thank you for listening!