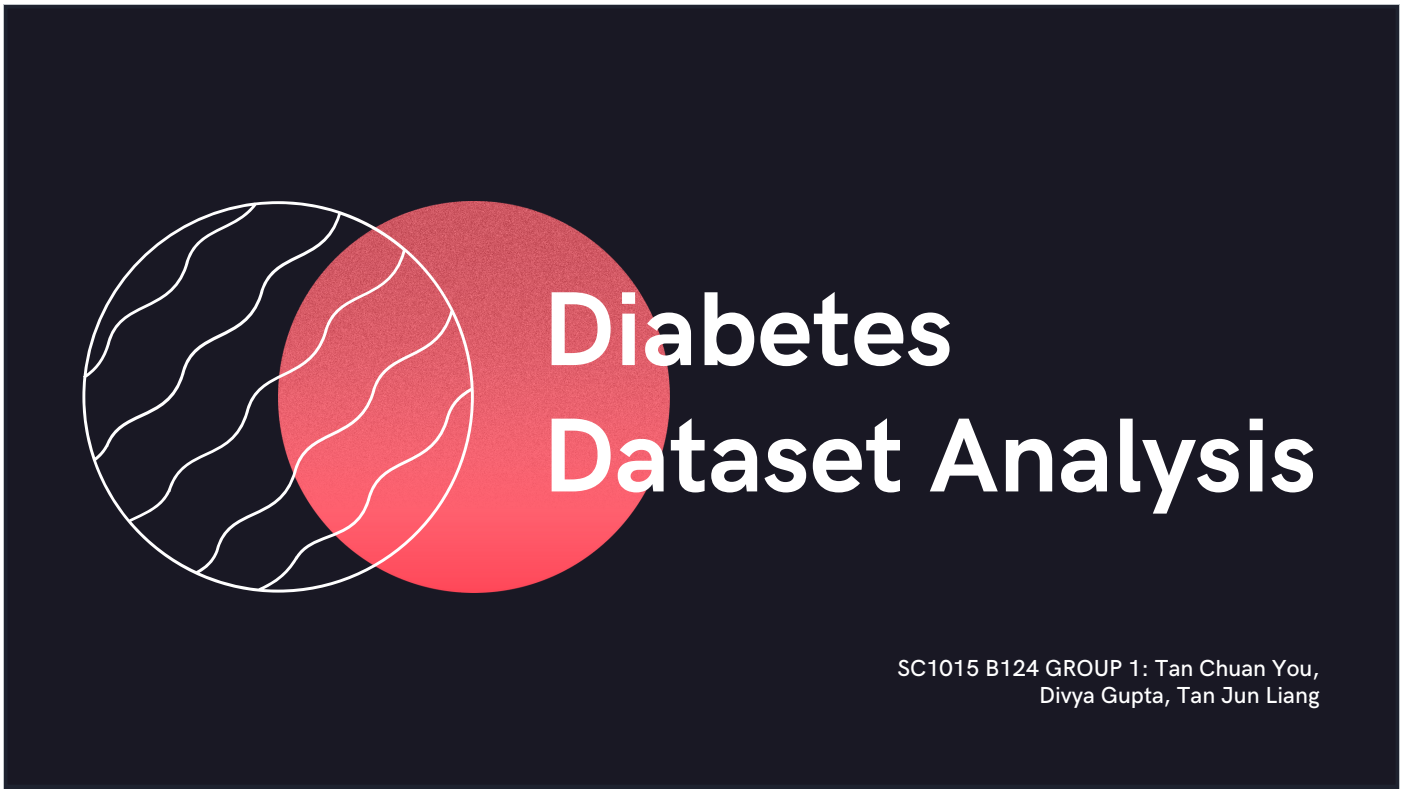
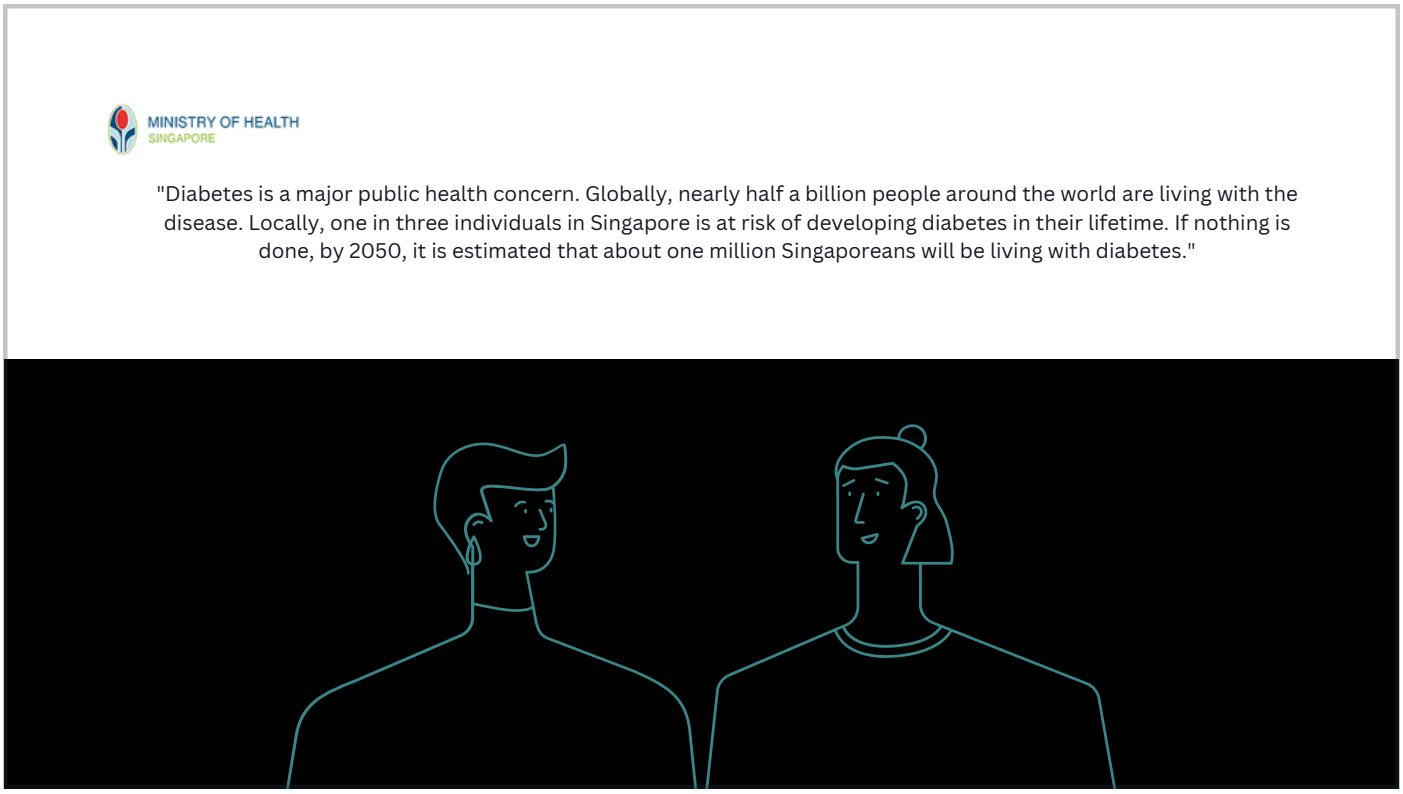
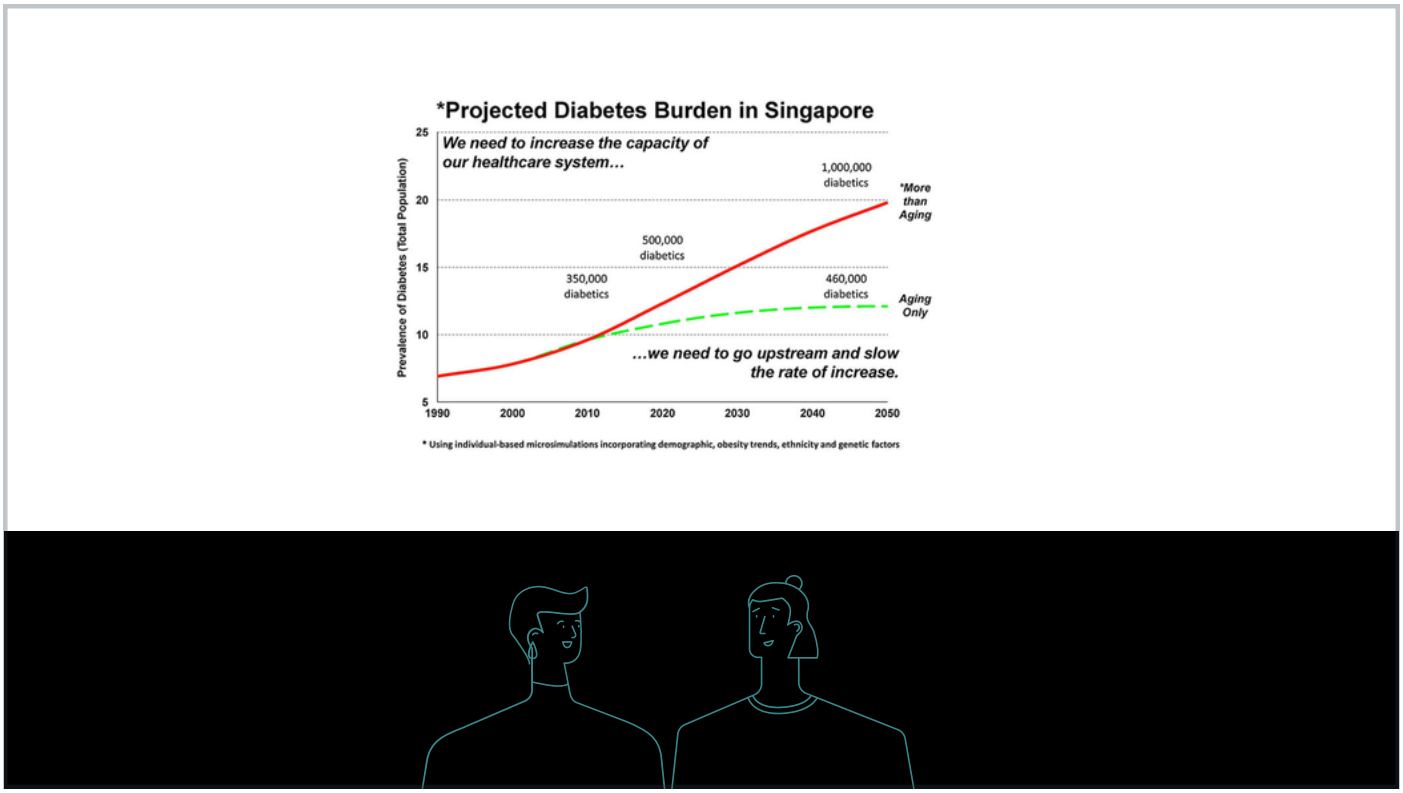


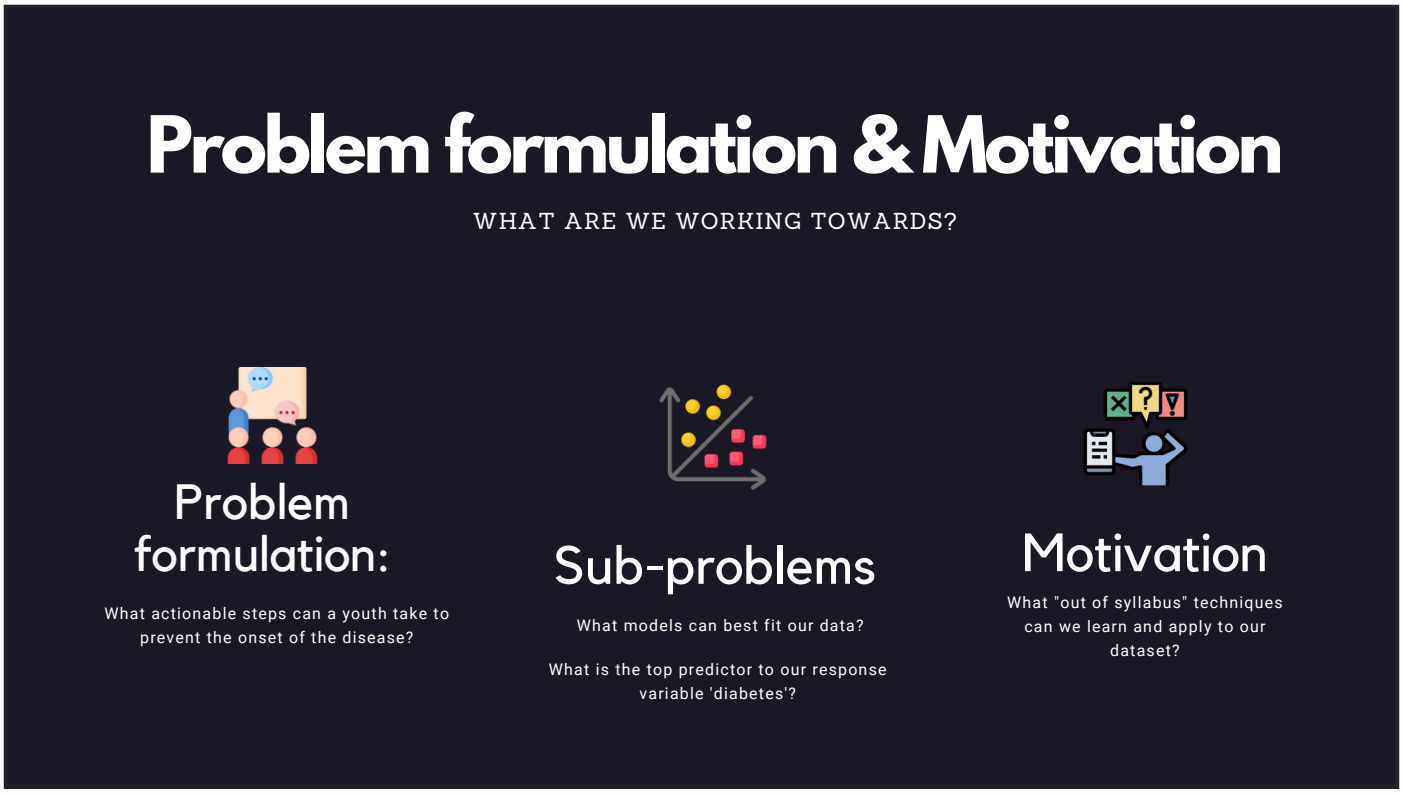
Hi everyone, today I will be presenting our analysis on a dataset related to diabetes.



Globally, diabetes has become a rising concern. As seen from the slide, one in 3 individuals in Singapore is at risk of developing diabetes in their lifetime.

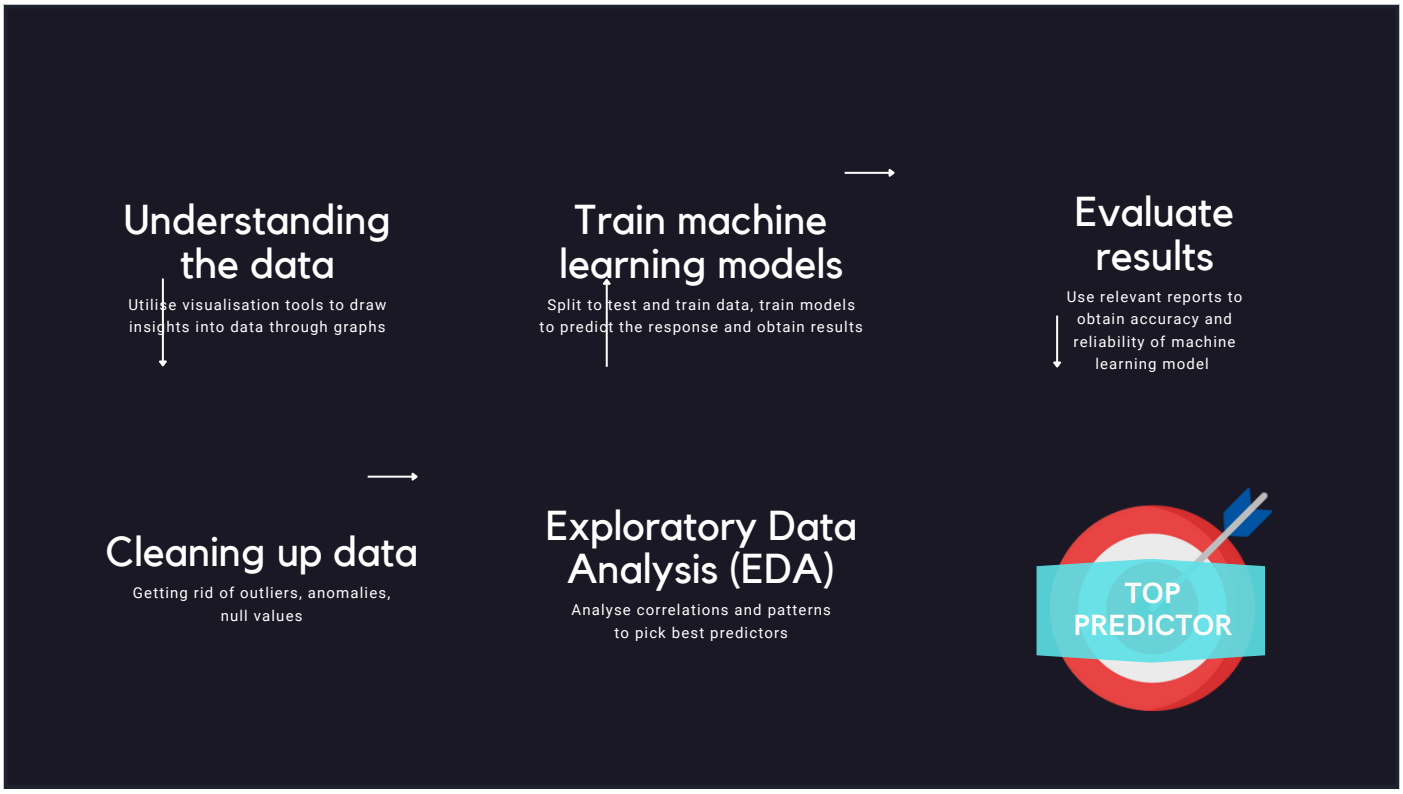


This rising concern motivated us to choose this topic to work on so that we can learn more about this disease.



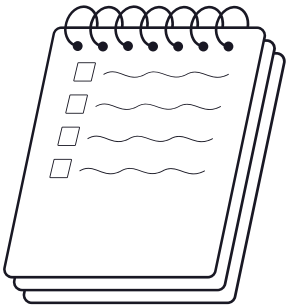
This brings us to our problem formulation. Our main question statement is to find out What actionable steps can one take to prevent the onset of diabetes? The sub-problems would be to figure out what model best fits our data and what is the top predictor that affects our outcome variable "diabetes". Learning new techniques to apply to our dataset motivates us to analyse in different manners so that we can produce a more reliable result.

To address our main question we need to attain our top predictor in response to the outcome variable and to do so we have to go through a few steps.



firstly, understanding the data by using library tools such as plotly to visualise the insights provided by the dataset.
secondly, data-cleaning to increase the usability in terms of performance accuracy.
thirdly, exploratory data analysis where we analyse trends and correlations to pick the potential predictors to work with.
Next, we train machine learning models with test and train data to predict the response variable.
Lastly, we evaluate the models' performances to obtain the most accurate and reliable learning model and thereafter we answer our question of what is the top predictor in response to whether or not one has diabetes?

DIABETES PREDICTION DATASET



Now, I will be presenting about the dataset we chose.

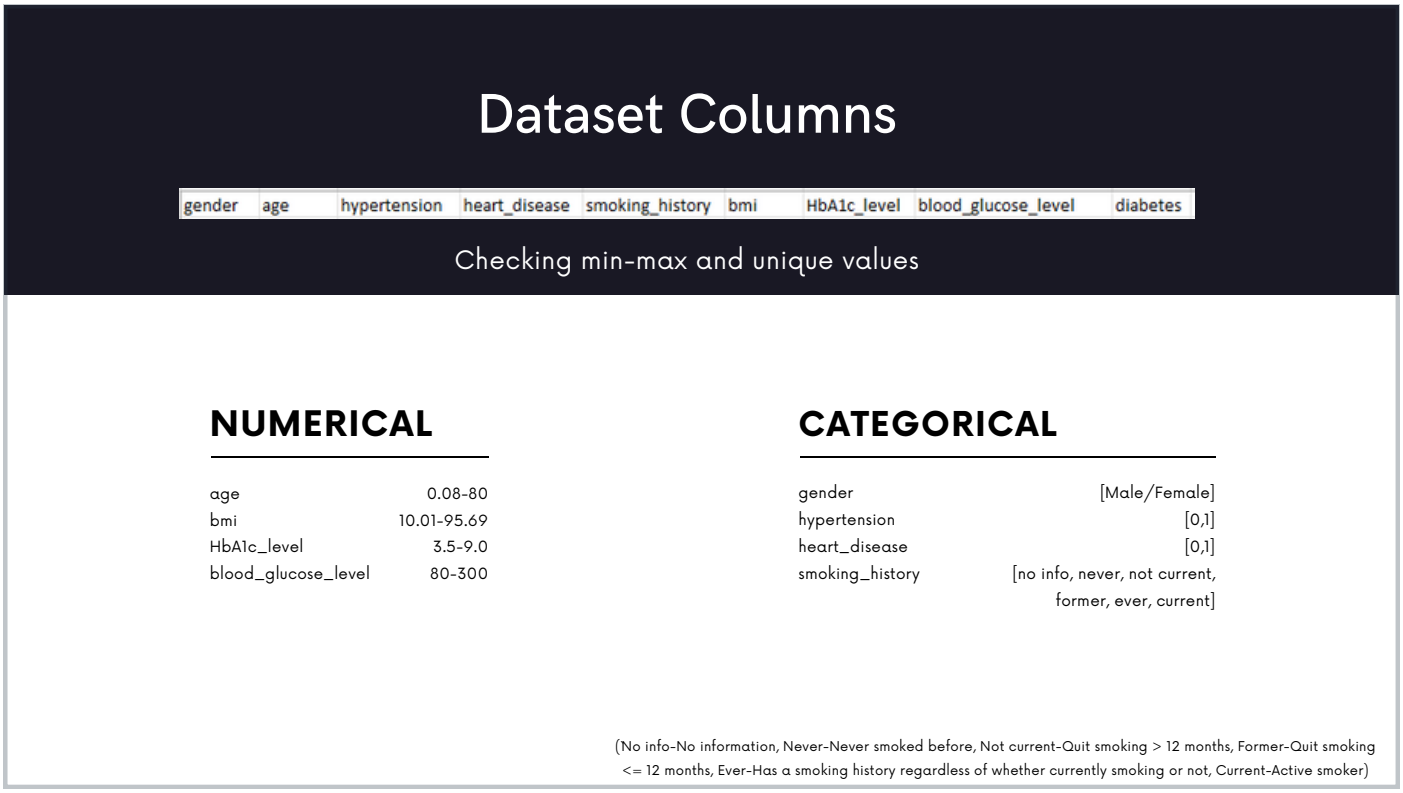
DIABETES PREDICTION DATASET

Collection of medical and demographic data from patients with their diabetes status (Positive or Negative)

SIZE OF DATASET: 100K

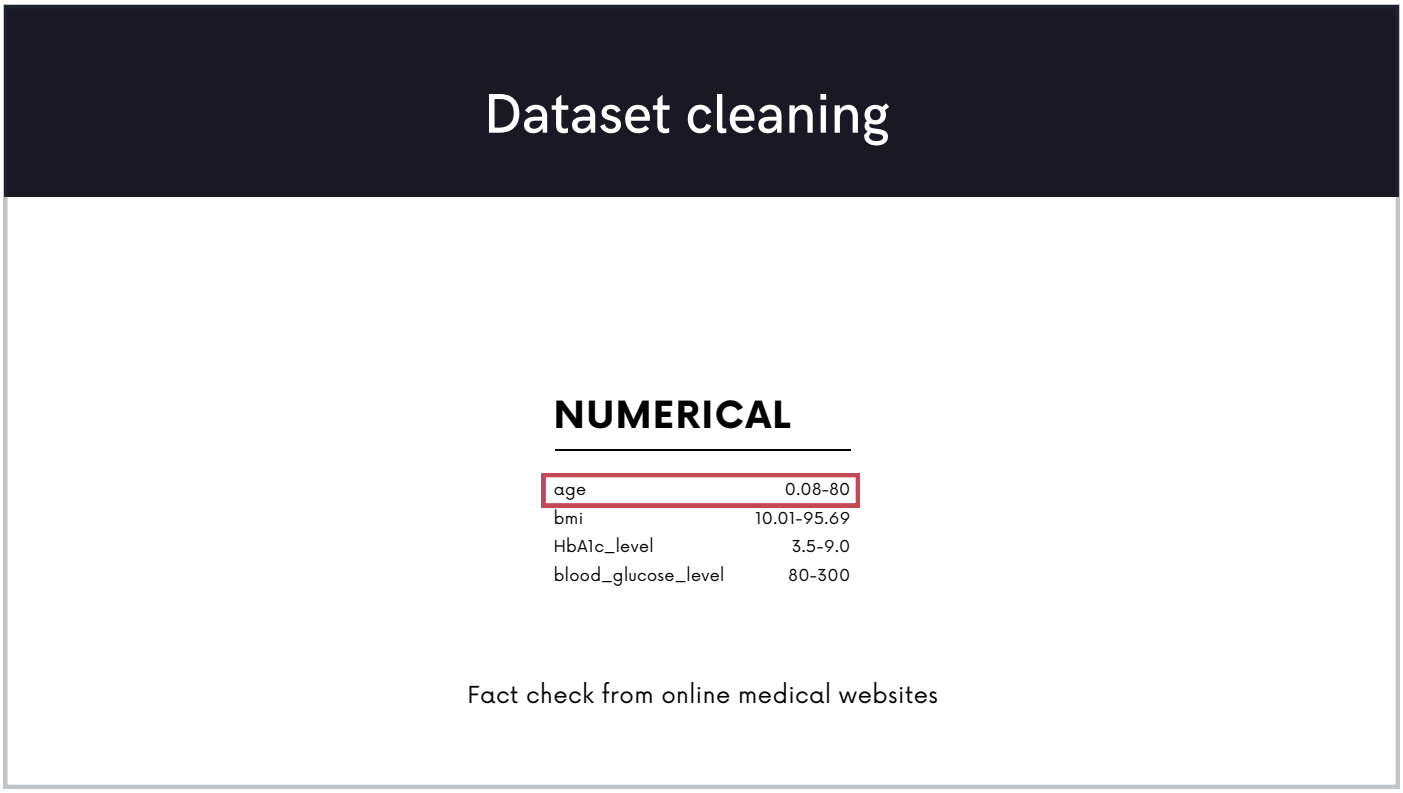
NUMBER OF PREDICTORS:
4 NUMERICAL
4 CATEGORICAL

This diabetes prediction dataset presents data collected from 100,000 patients of various medical backgrounds and demographics with their diabetes status, positive or negative. There is a total number of 9 variables in this dataset including the response variable.



To visualise our data, we checked the range of the predictors and checked for any possible anomalies.

Dataset cleaning

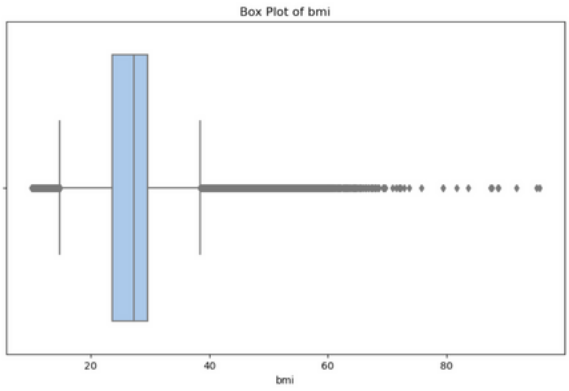


One value struck out which was the smallest value in age, 0.08 years, a 4 week newborn. However upon further online research we found out that the values fit the range of the other variables and the data is valid.

Dataset cleaning

NUMERICAL

age	0.08-80
bmi	10.01-95.69
HbA1c_level	3.5-9.0
blood_glucose_level	80-300

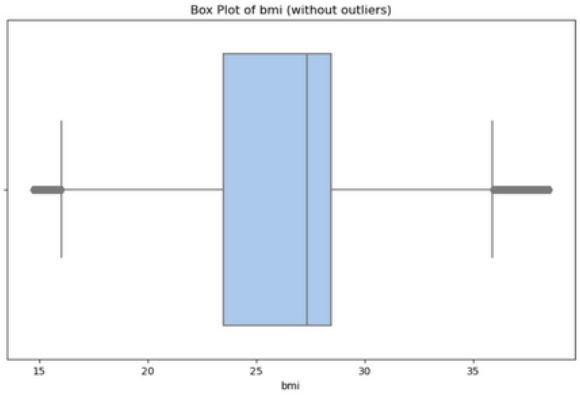


Next, we used box plots to visualise our data and the outliers that come with it.

Dataset cleaning

NUMERICAL

age	0.08-80
bmi	10.01-95.69
HbA1c_level	3.5-9.0
blood_glucose_level	80-300

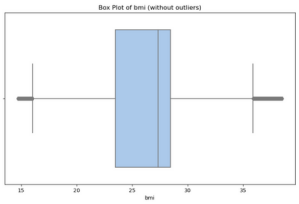


We removed the outliers as part of data cleaning to make the data more accurate in terms of representation.

Dataset cleaning

NUMERICAL

age	0.08-80
bmi	10.01-95.69
HbA1c_level	3.5-9.0
blood_glucose_level	80-300

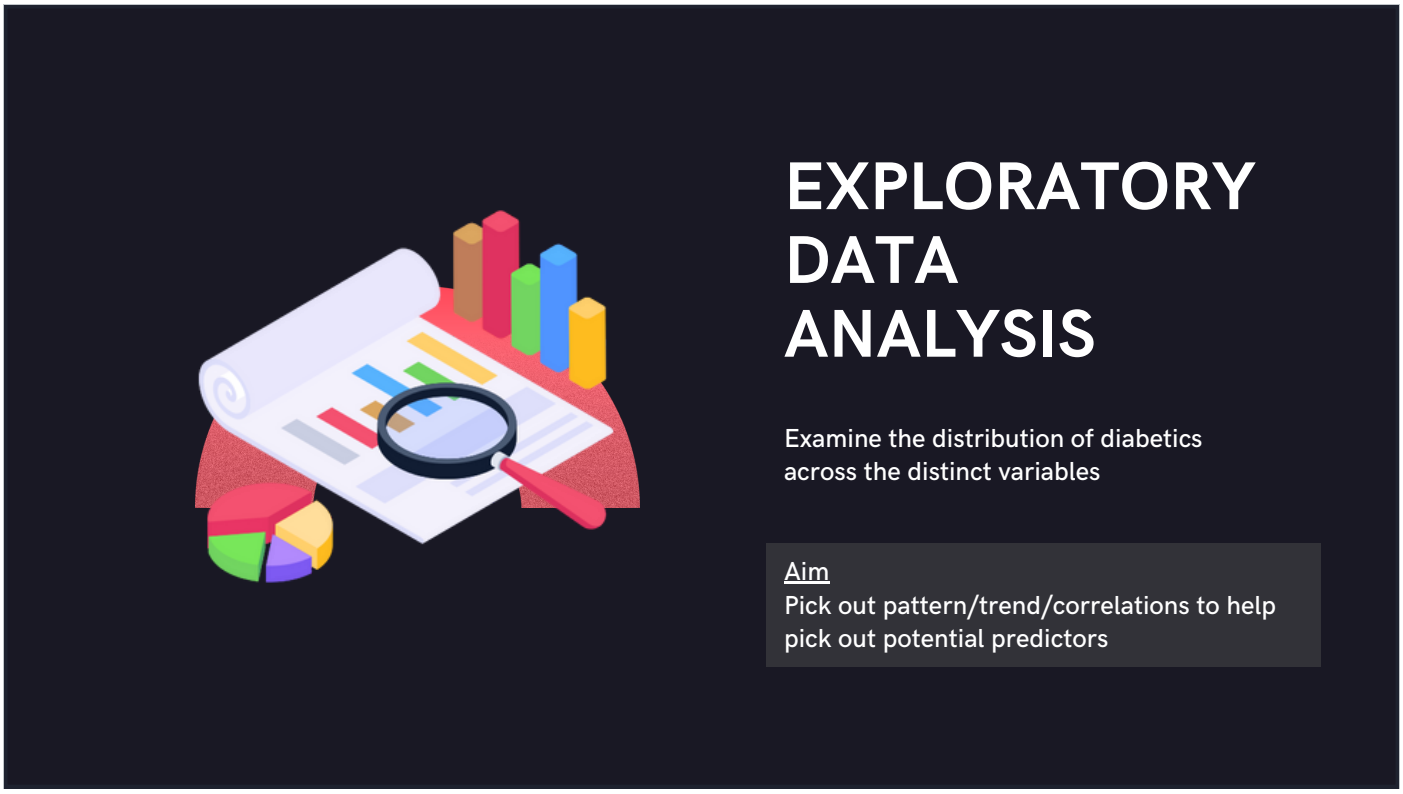



```
null_columns = dData.columns[dData.isnull().any()]
print(dData[null_columns].isnull().sum())
```

```
Series([], dtype: float64)
```

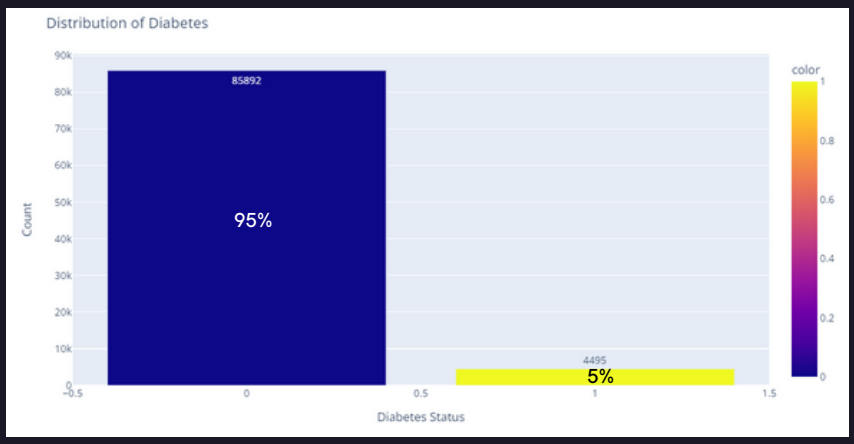
We noticed that there were no null data that was to be cleaned or replaced

Lastly, we checked for any null values and found out there was no null data we had to work with or clean.



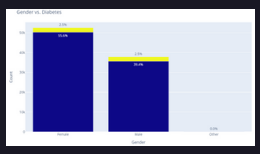
Moving on to the EDA. Our aim for this section was to identify any patterns, trends or correlations to help us pick out potential predictors to work with.

DISTRIBUTION OF DIABETICS

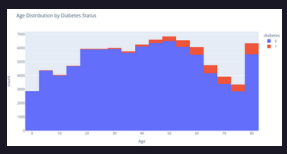


We used plotly as our visualisation tool, something out of our syllabus. As you can see, only 5% of the dataset have a positive status for diabetes making the data disproportionate and imbalanced.

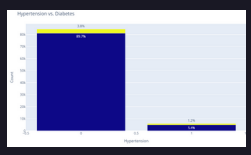
VISUALISE



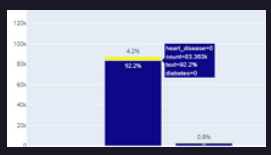
gender



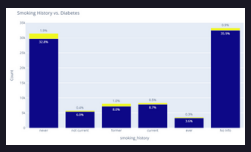
age



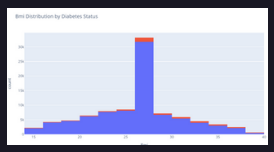
hypertension



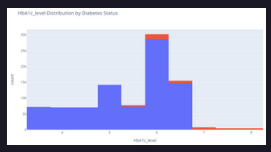
heart_disease



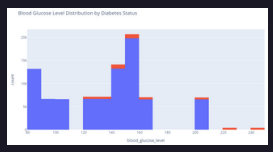
smoking_history



bmi



HbA1c_level

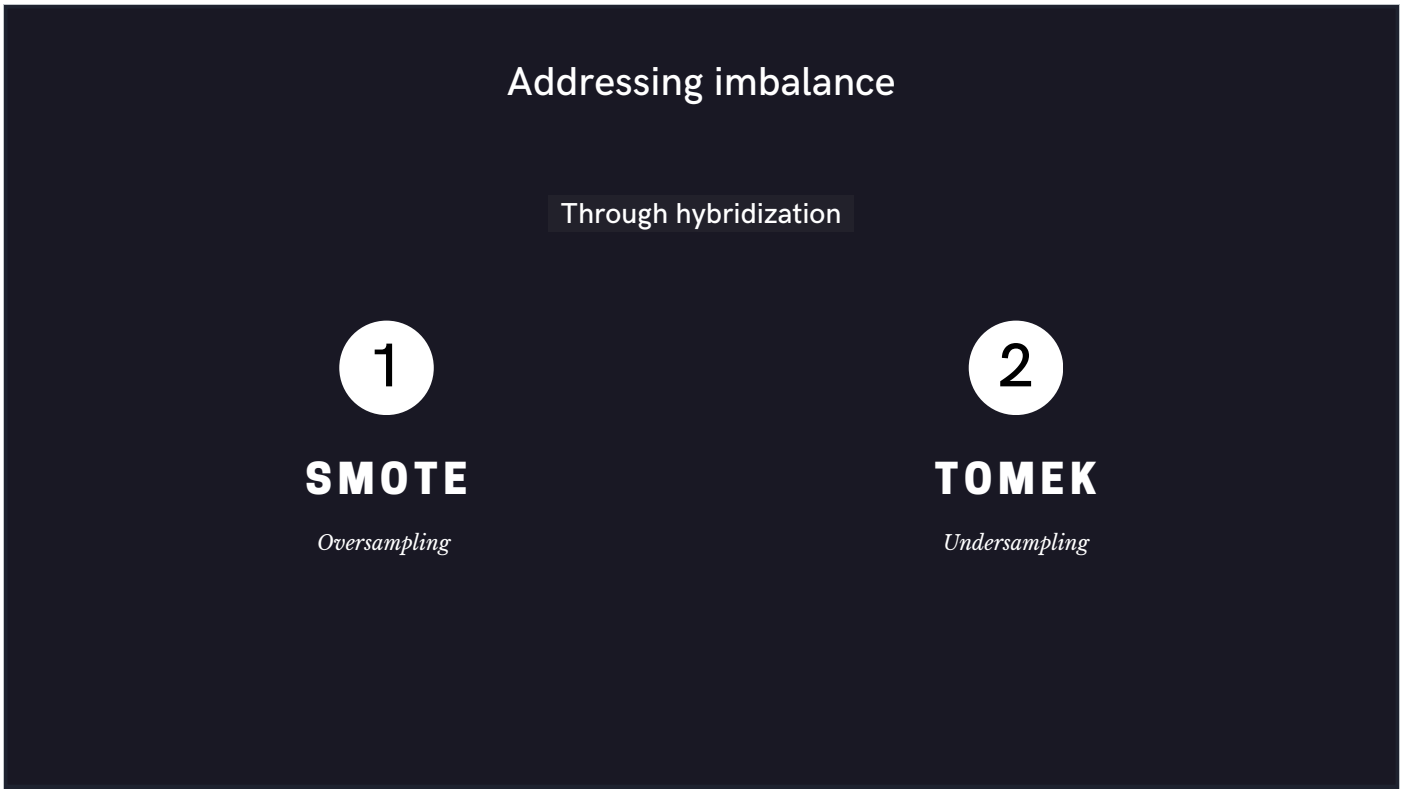


blood_glucose_level

We plotted similar graphs to visualise the distribution of the individuals across the various predictors and we see there is a similar imbalance between the diabetics and non-diabetics.

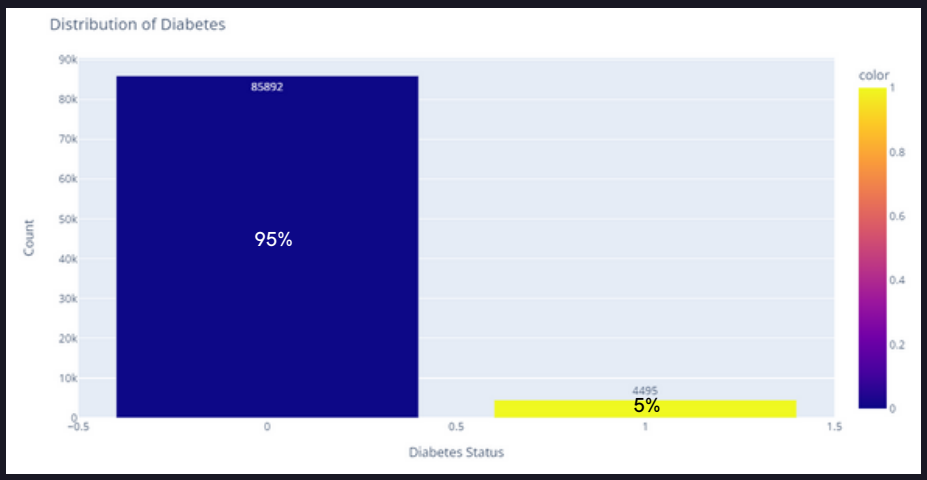
Addressing imbalance

To address this issue, we implemented the following technique to resample our data



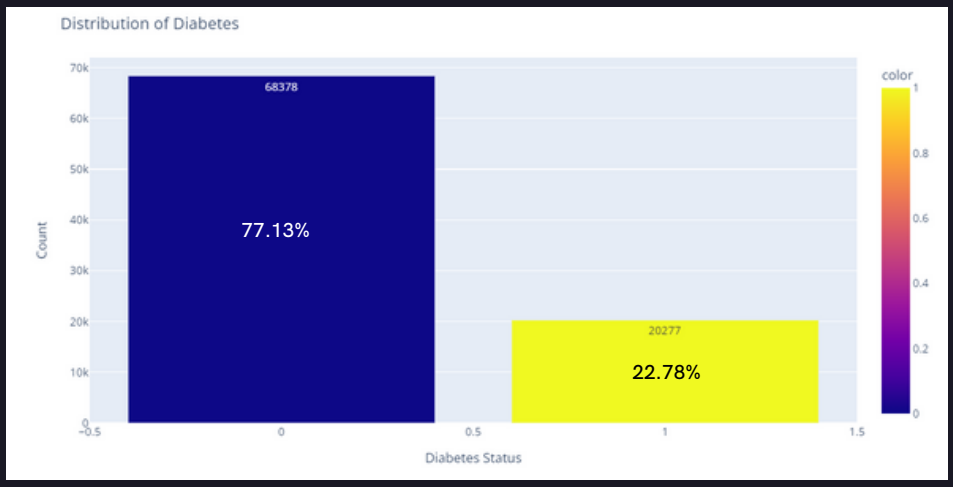
Our technique utilises a combination of SMOTE and Tomek.
first of all SMOTE, is an oversampling technique used especially when the minority class is underpresented. The method generates synthetic samples by interpolating new points between existing minority class samples. However, it is sensitive to noise in the data and if the minority class is too small, this technique may not be efficient and might duplicate results.
On the other hand, TOMEK is an undersampling technique that addresses class imbalance where the majority class is overrepresented. It works by identifying pairs of samples from different classes that are very close to each other in feature space. It then removes samples belonging to the majority class.
The hybrid technique which uses SMOTE and Tomek together works in this manner:
SMOTE creates synthetic minority class samples while Tomek removes samples that may be wrongly classified as majority class samples. This results in a better balance of class distribution and a reduction in overfitting.

DISTRIBUTION OF DIABETICS



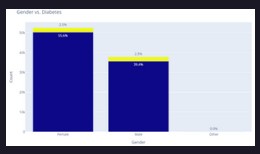
After visualising the resampled data we see an improvement in the data distribution. This is the original data distribution.

DISTRIBUTION OF DIABETICS

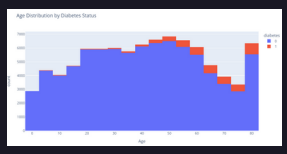


and this is the resampled data on the train dataset. There is a better distribution as seen from a 17.78% increase in the data with positive diabetes status. Across the different predictors, we see a similar improvement in distribution across the graphs as well.

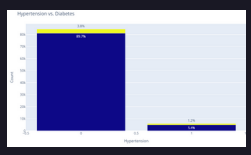
VISUALISE



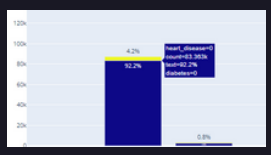
gender



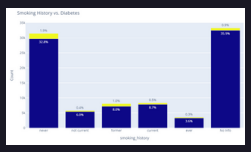
age



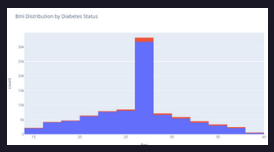
hypertension



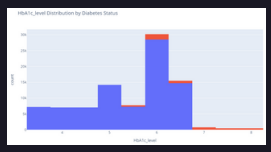
heart_disease



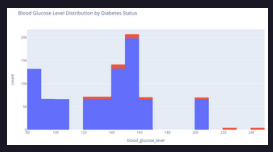
smoking_history



bmi



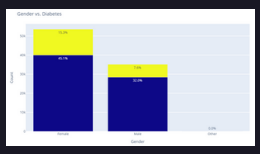
HbA1c_level



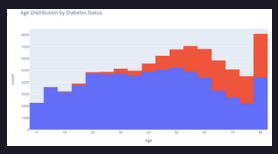
blood_glucose_level

This is the before

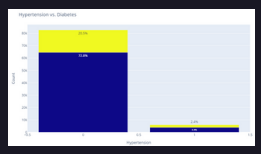
VISUALISE



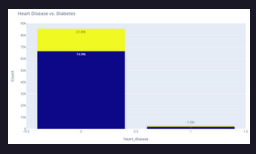
gender



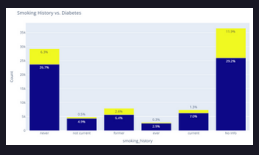
age



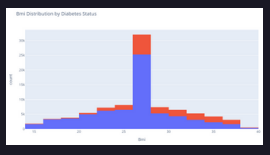
hypertension



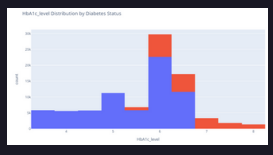
heart_disease



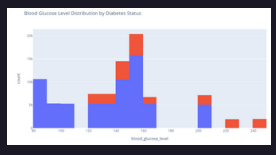
smoking_history



bmi

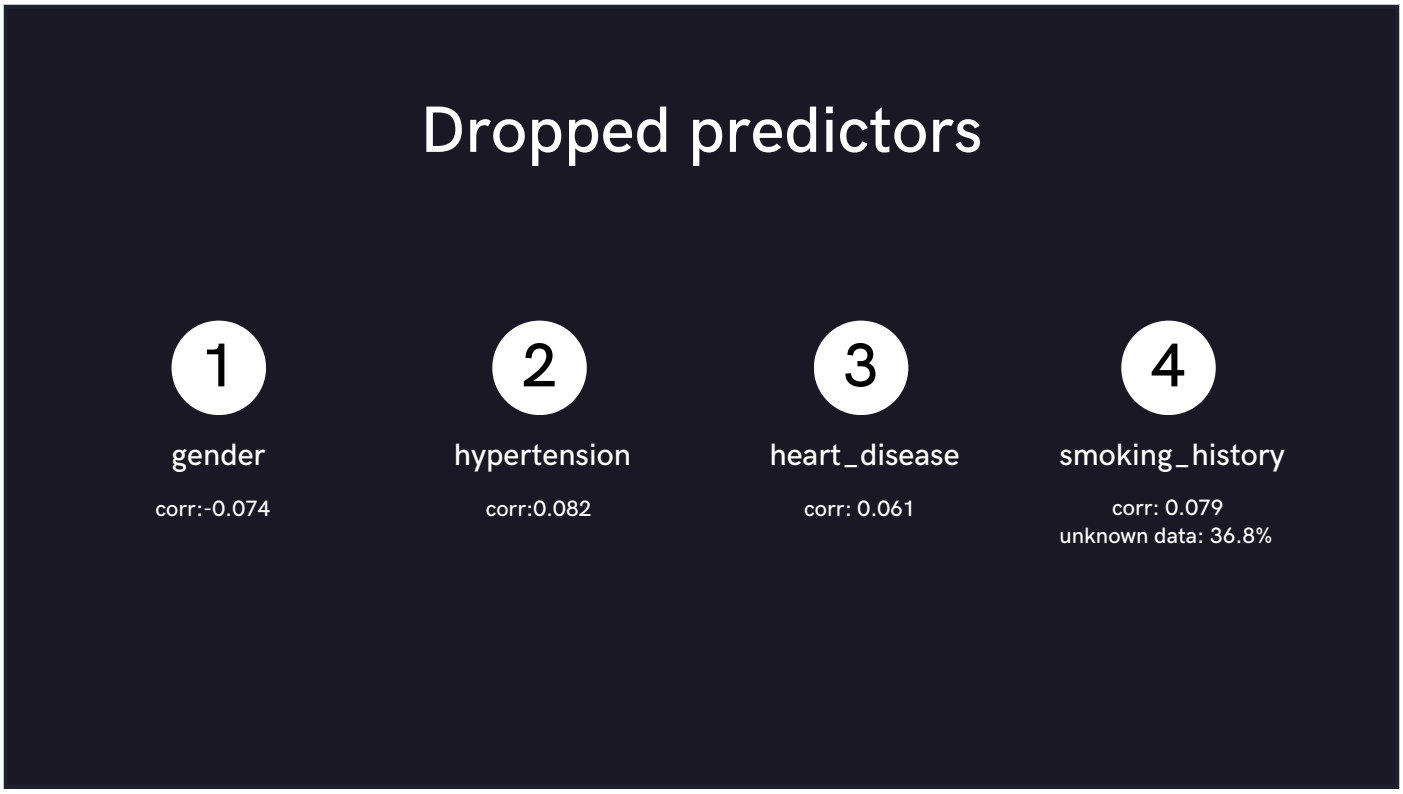


HbA1c_level

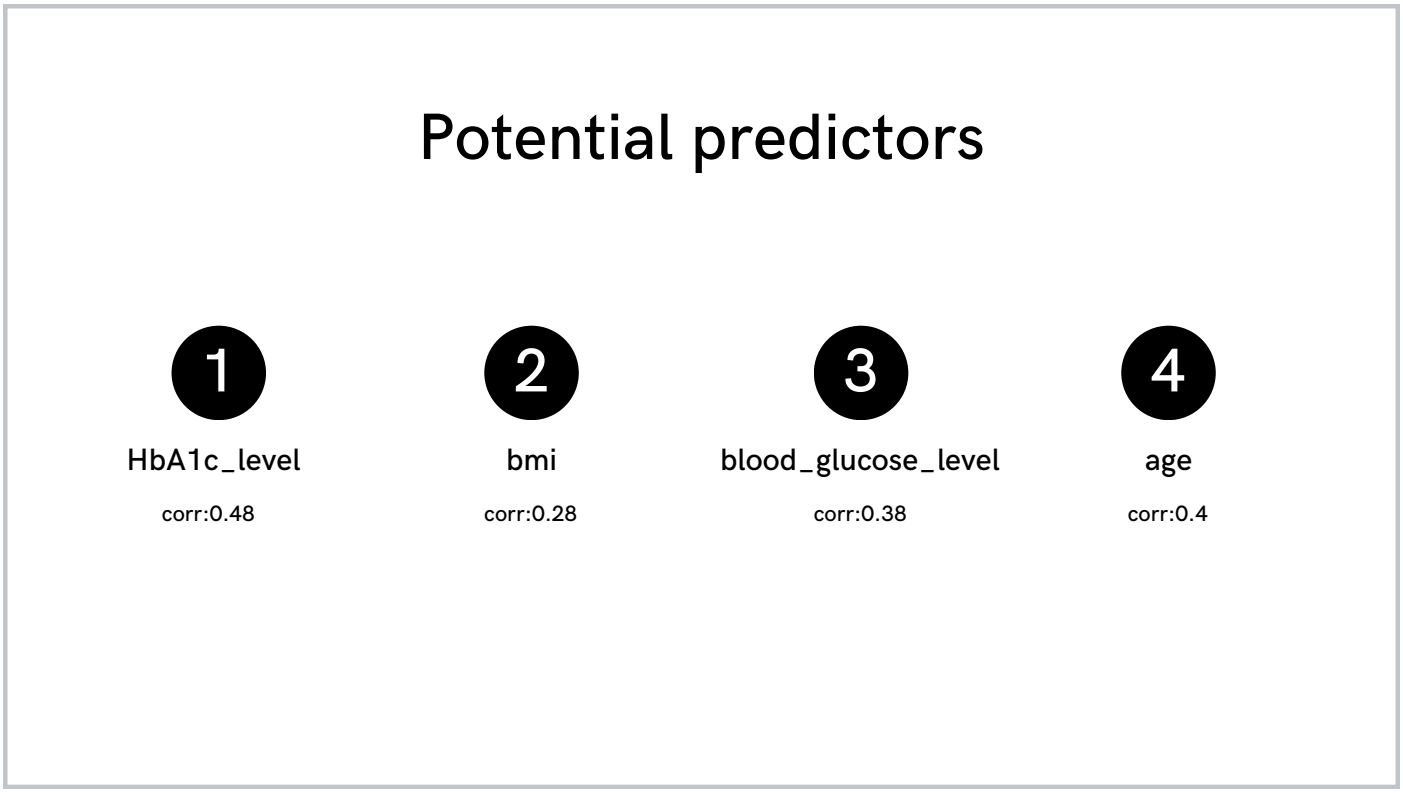


blood_glucose_level

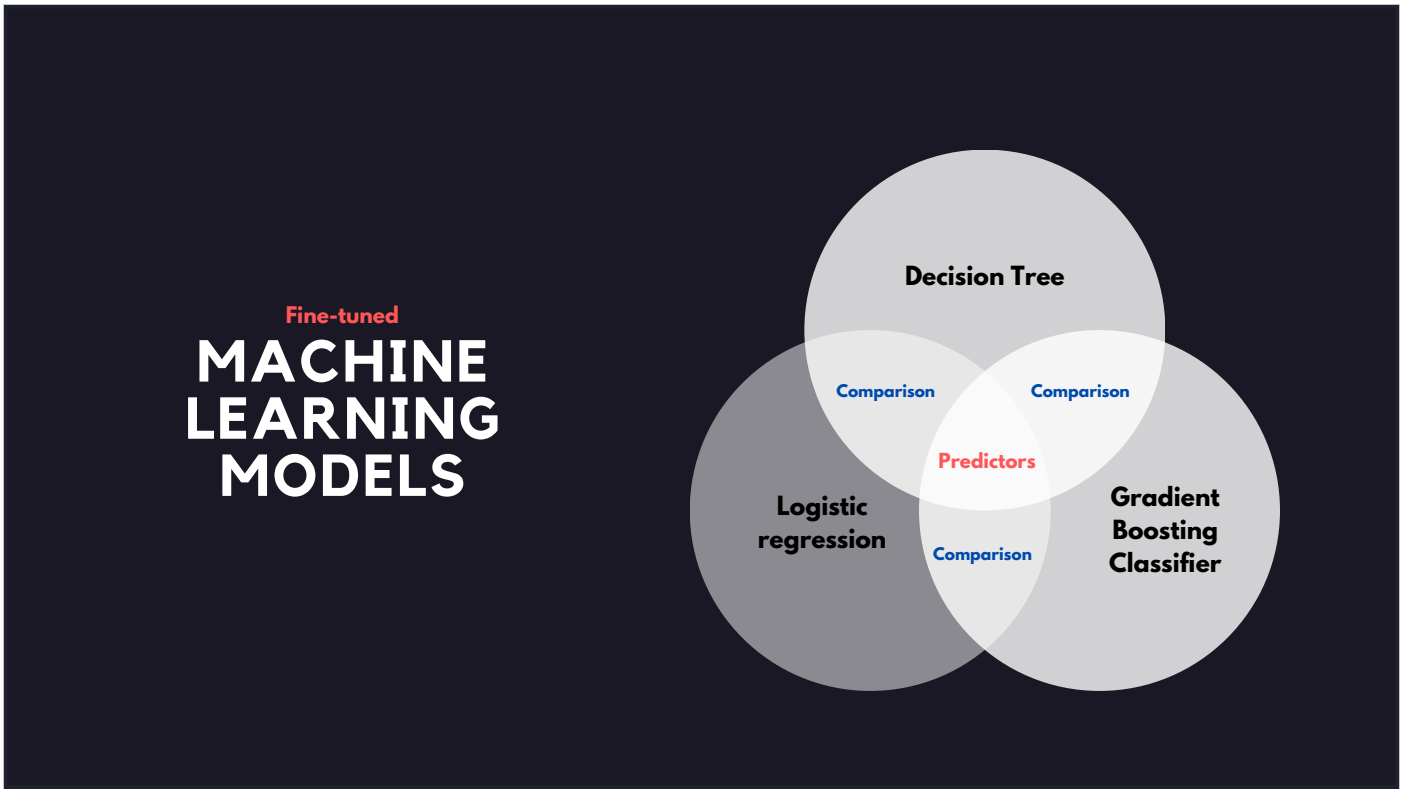
And this is the after. Clearly, this technique helped show us the trends and patterns better.



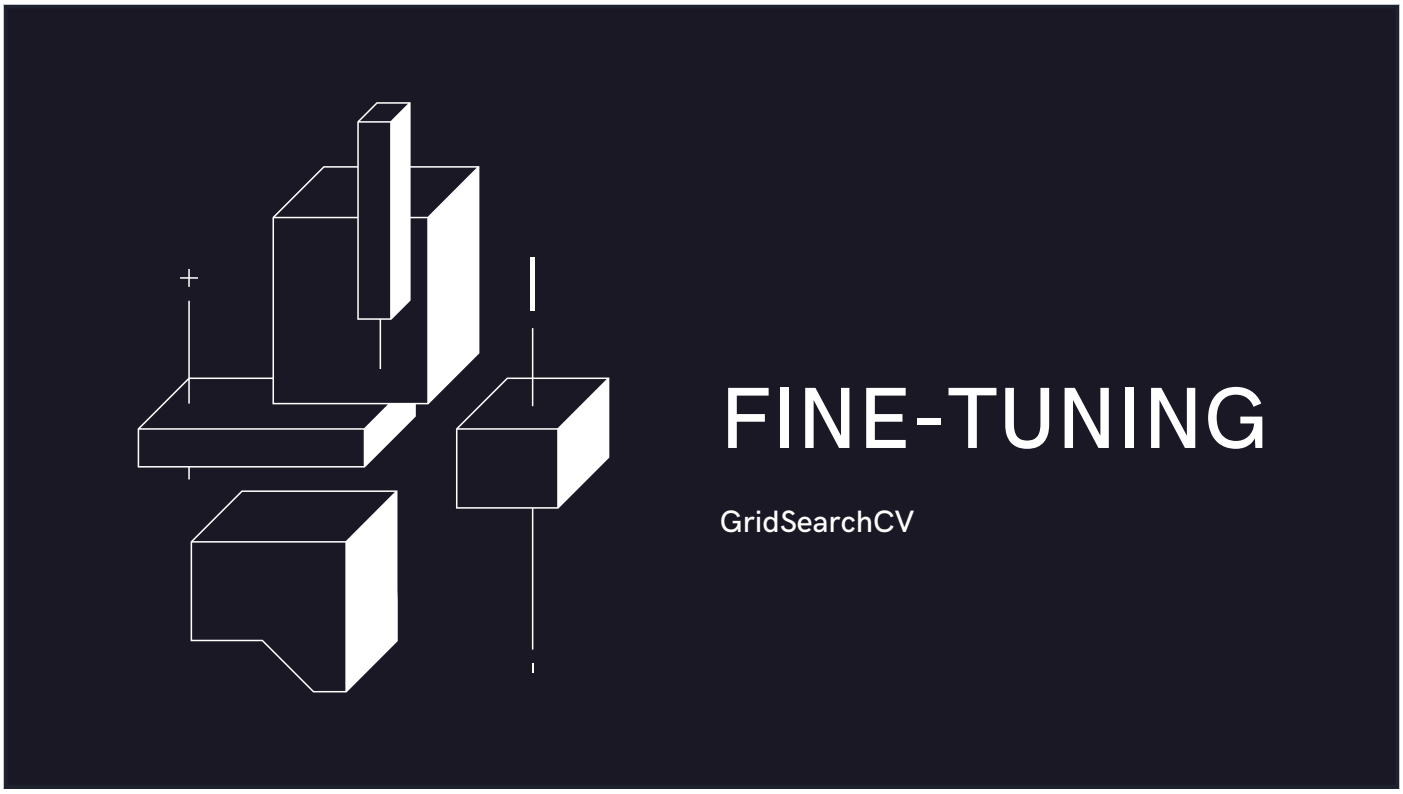
Upon further analysis of the predictors, we chose to drop the following 4 predictors due very low correlation between them and the response variable. We dropped smoking history due to the high percentage of unknown data as well.



And these are our resulting potential predictors we will be working with due to their relatively high correlation values. Seeing that they are all numerical data, we would take note of this while choosing the machine learning models.

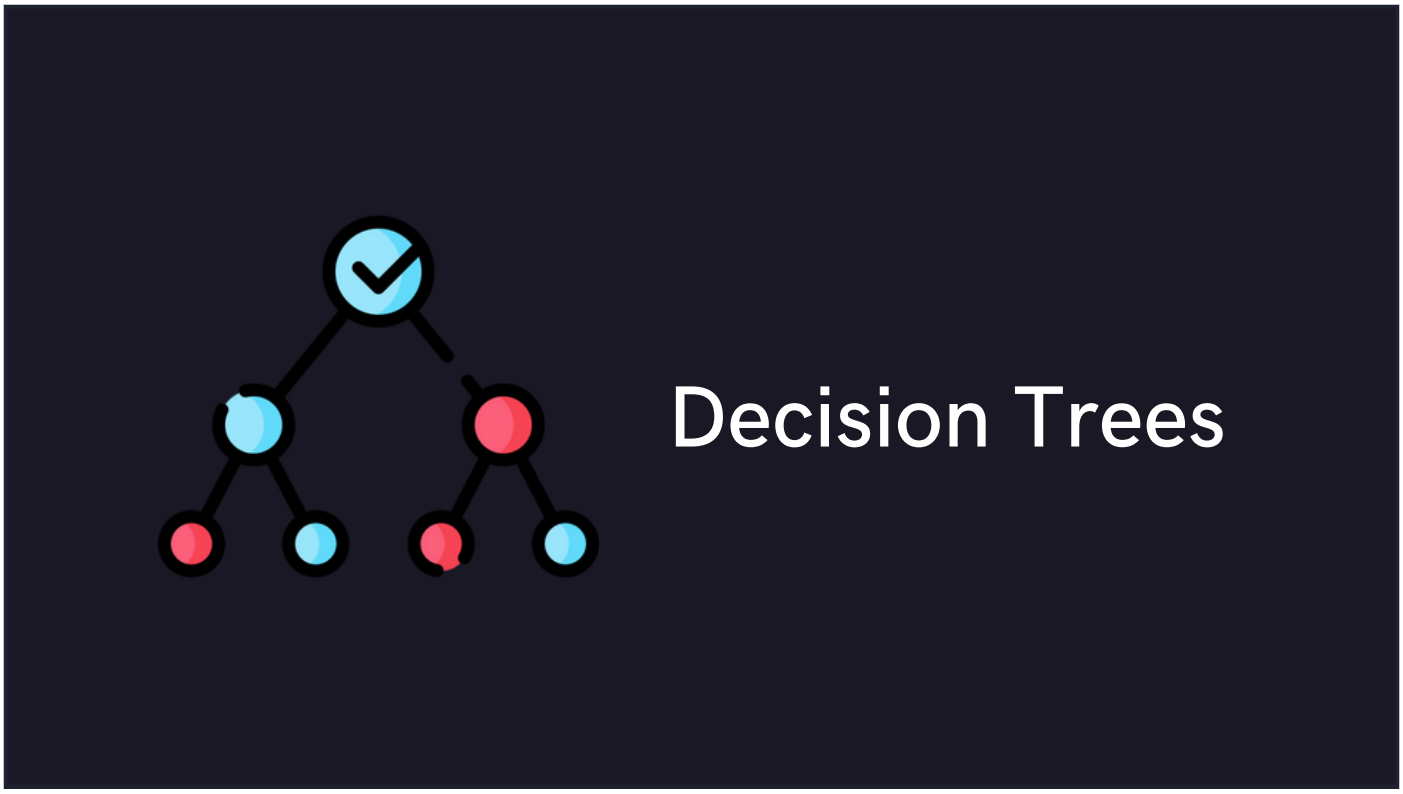


In general, we trained these three machine learning models after fine-tuning them and used performance metrics along with ROC curve to evaluate the models.



Before we move on to the respective machine learning models, we researched on the fine-tuning tools and decided to work with GridSearchCV. This technique exhaustively searches for the best combination of hyperparameters for a machine learning model by testing all possible parameter combinations. We incorporated this into all our machine learning models to increase the accuracy of our results.

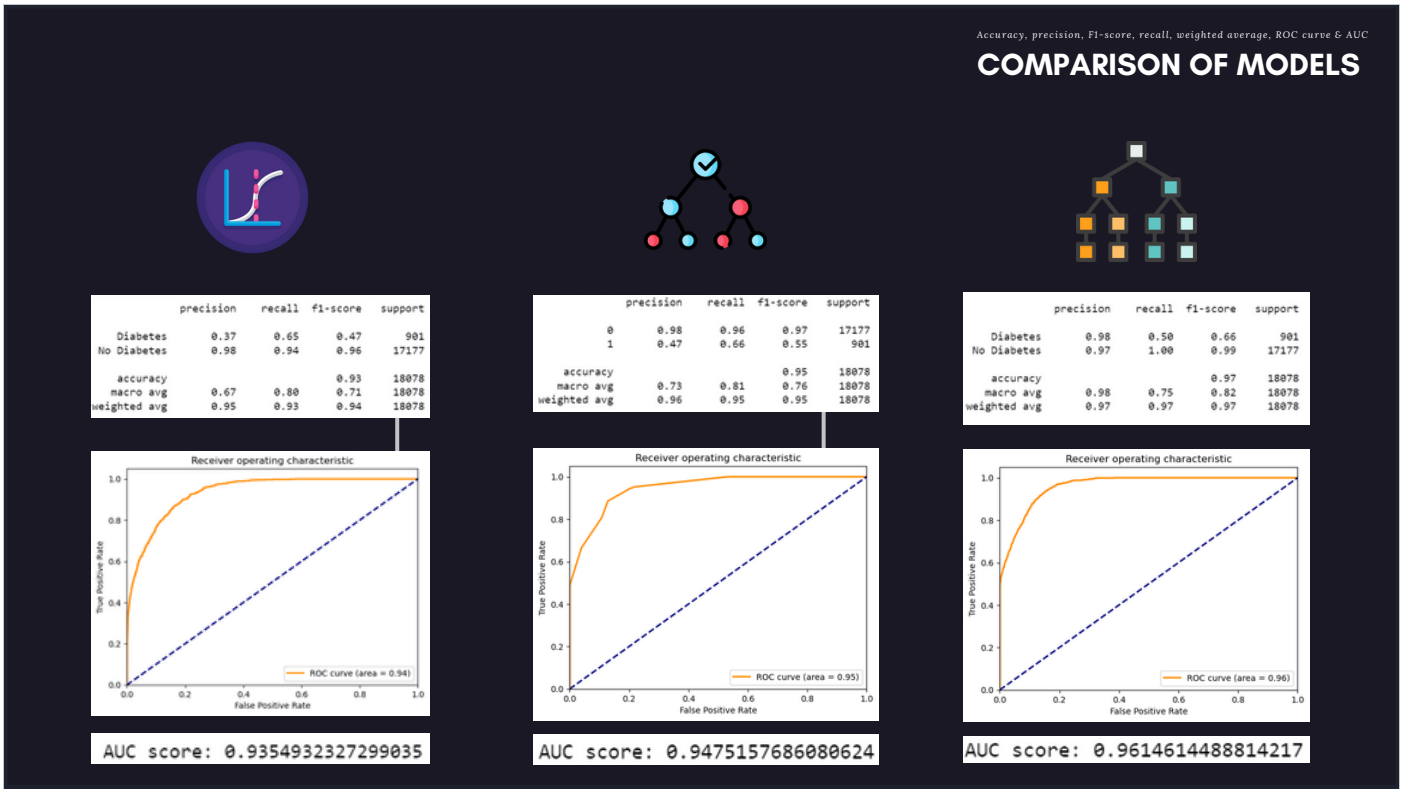
The first model we used is logistic regression with the reason being that it evaluates binary outcomes, which fits our binary response variable 'diabetes'. It also has a performance evaluation metric that allows us to account for the imbalance in the data.



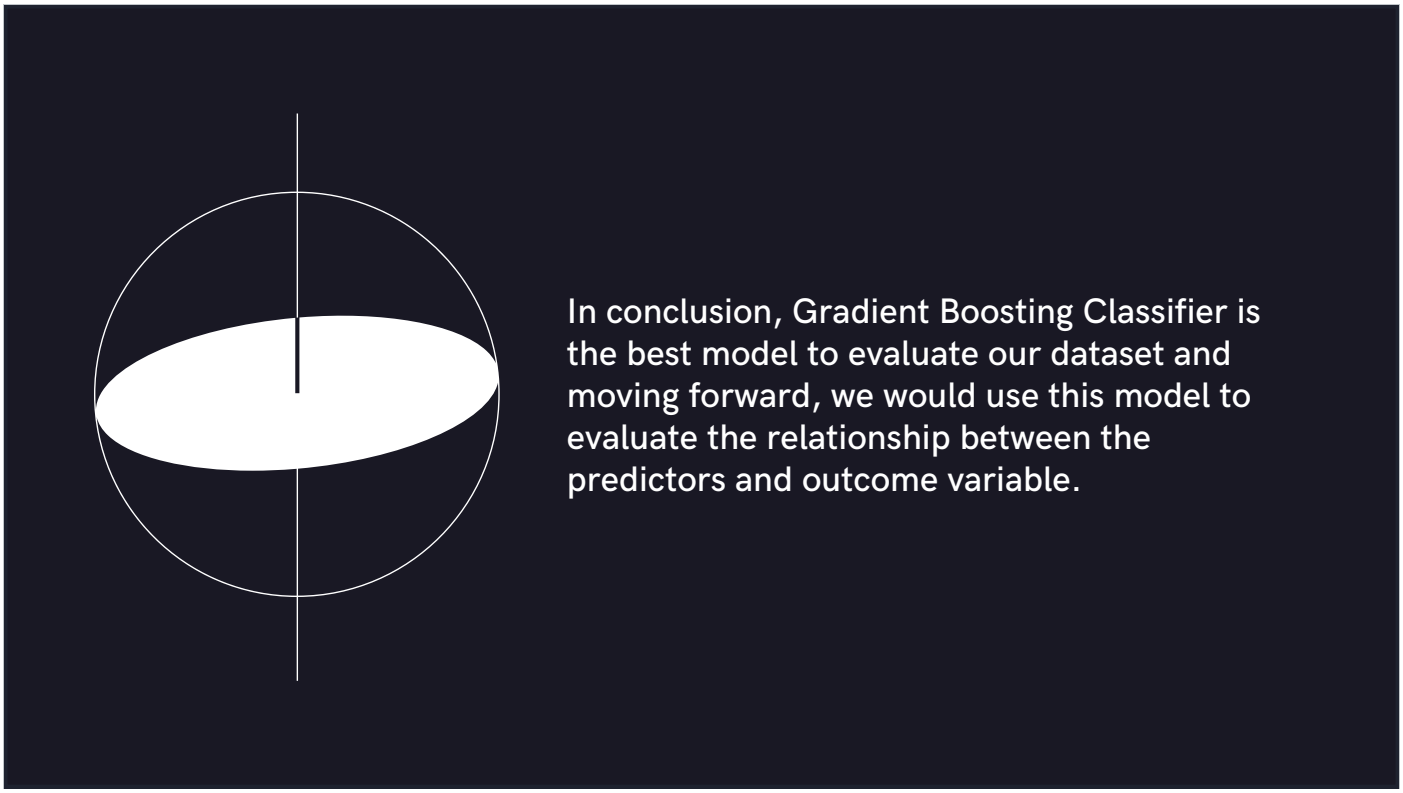
The next model we chose was decision trees as it can identify complex nonlinear relationships between the predictors and outcome variable. This model is also trained relatively quickly and is computationally effective. Therefore, with our huge data of 124000 rows, the decision tree is suitable as it can be scaled to handle larger data if need be.

Since the accuracy scores of the decision tree was higher than that of the logistic regression, we chose a model with a similar foundation to the decision tree model, Gradient Boosting Classifier. This model is able to handle complex relationships, high-dimensional data, class imbalance and provide interpretable results. We will refer to this model as GBC

Next comparing the performances of these models.

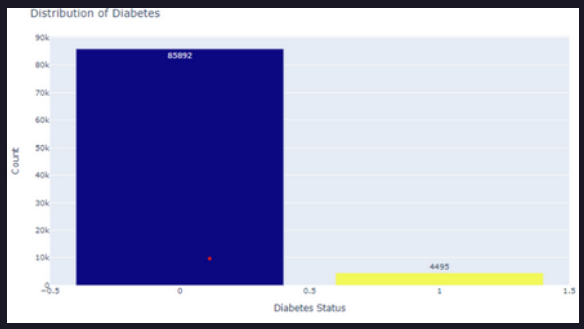


Firstly comparing the score values from the classification report we see that the highest accuracy belongs to the GBC with a value of 0.97. Comparing the weighted average score which takes into account the imbalanced class distribution while calculating the average of precision, recall and F1-score, we see that the GBC has the highest weighted average score of 0.97 as well. therefore, its performance is considered the best among the three models. Next comparing using the ROC curve, we see that the AUC score of the GBC is the highest indicating the best overall performance in terms of distinguishing between the two classes. Also, GBC's curve is the closest to the top left corner indicating the highest true positive rate for a low false positive rate.



Hence we deem Gradient Boosting Classifier as the best model to evaluate our dataset.

Drawbacks & Discoveries

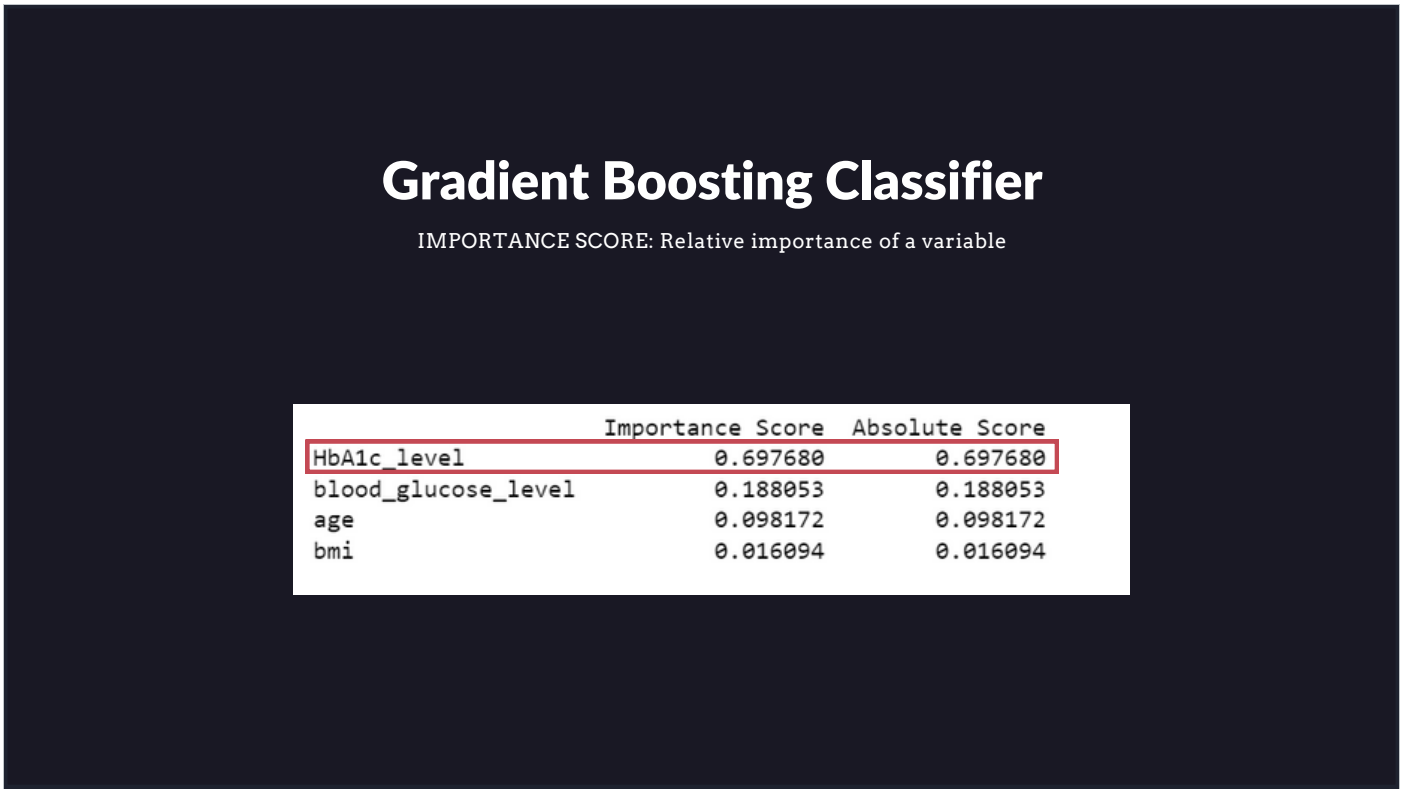


Before we move on to talk about our top predictor, we would like to point out some discoveries and drawbacks of our project.

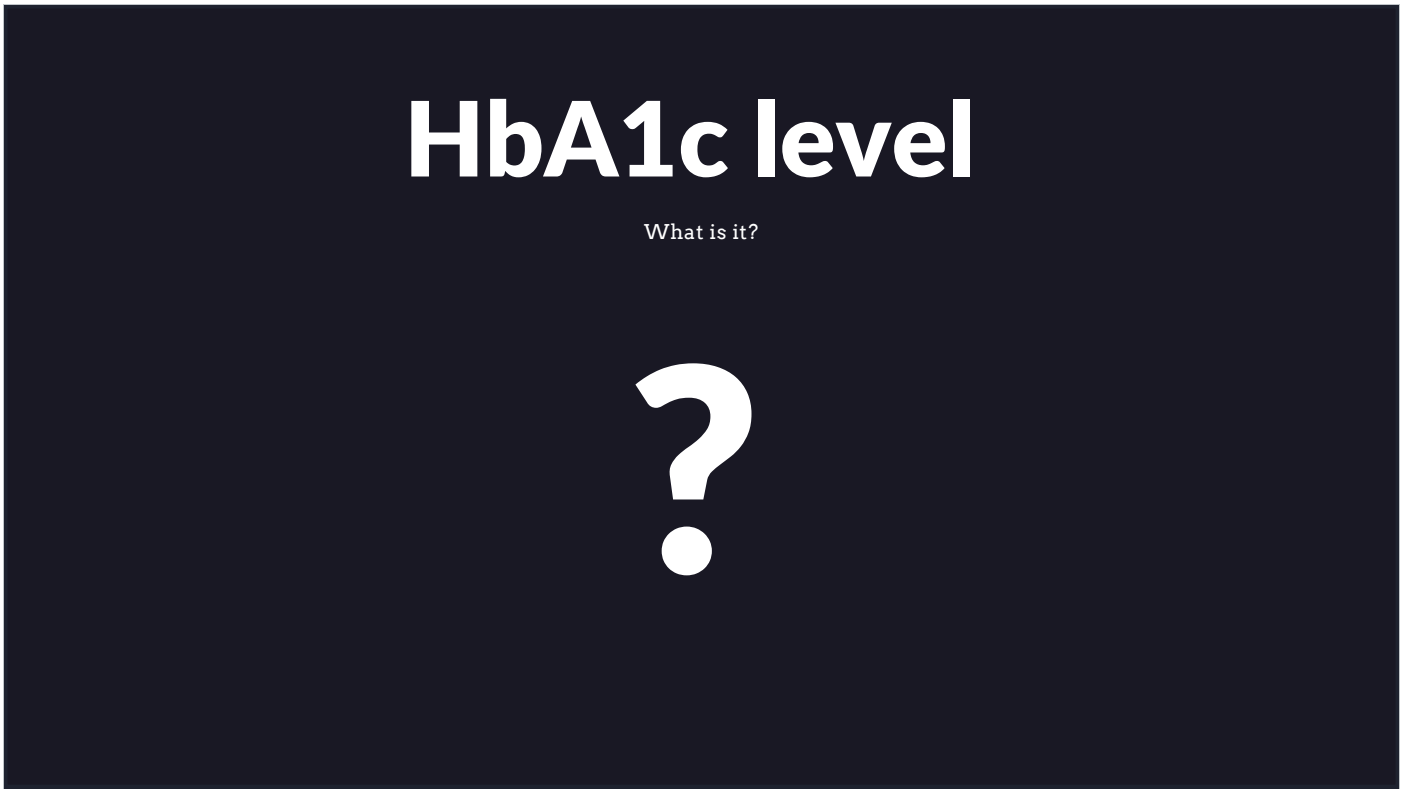
As our data was significantly imbalanced, we had to resort to SMOTE and TOMEK to spread out the distribution. Although it is an appropriate technique to deal with imbalanced data, overfitting may still occur.

To get better accuracy in the real world, datasets should consider the response variable and strive to even out the percentage of groups.

In addition, an interesting observation we made was that pre-existing medical conditions like hypertension and heart disease have low correlations to the onset of diabetes!
Now, lets take a look at which predictor is the most important to our response variable.



Importance score is a measure of the relative importance of a variable in a non-linear model, in this case, GBC. Ranking the variables by their importance score, we see that HbA1c_level has the highest value indicating that it has the highest impact on the response variable, followed by blood_glucose_level, age and lastly BMI. In addition, the other models gave the same ranking to the variables, making it undisputable for HbA1c_level to take the number one spot.



So, what is HbA1c?

HbA1c level

What is it?

HbA1c (also called glycated hemoglobin or hemoglobin A1c)

Blood test that measures the average blood sugar level over the past 2-3 months.

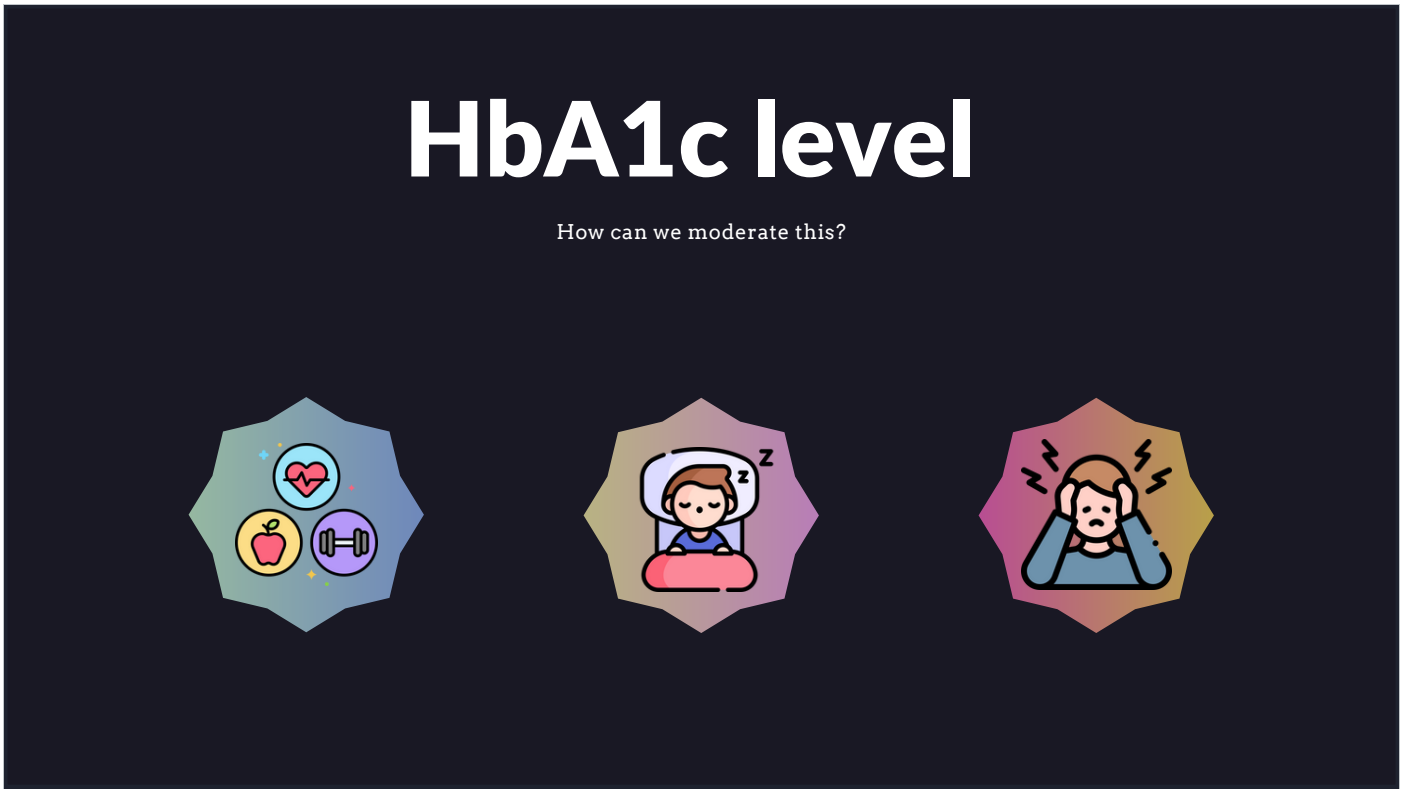
In people without diabetes, HbA1c levels should be kept below 5.7%.

In people with diabetes, the targeted HbA1c level should be below 7%.

It is a blood test that measures the average blood sugar level over the past 2-3 months. It is used as a diagnostic tool to determine if a person has diabetes or to monitor the blood sugar control in people with diabetes.

Safe levels of HbA1c for non-diabetics are typically below 5.7% while the target level for diabetics is generally below 7%. However, the optimal range may vary depending on individual circumstances, such as age and other health conditions.

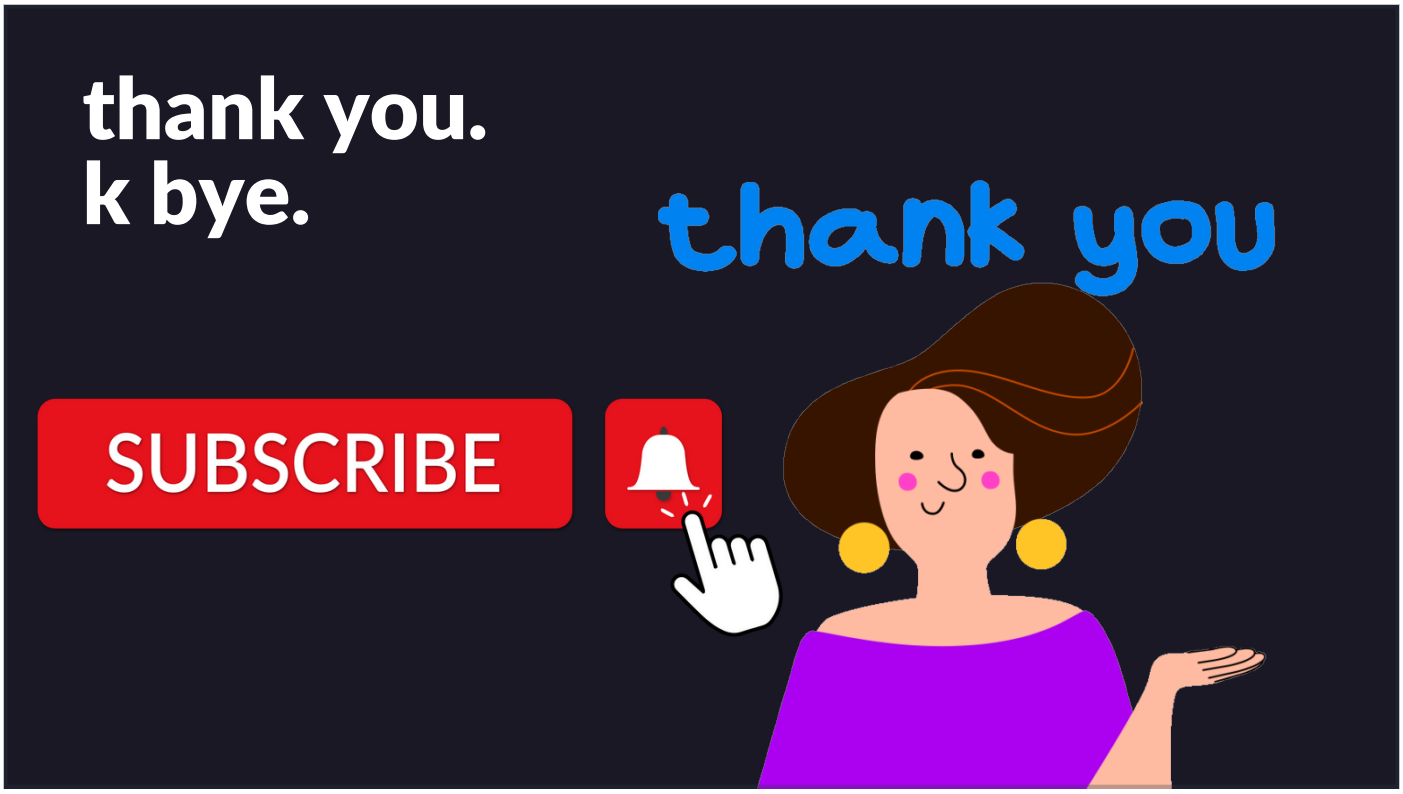
So what should youths do to moderate their HbA1c levels when diabetes is a growing global issue and concern?



For youths, it's important to adopt healthy habits early on to prevent or delay the onset of diabetes. This includes eating a balanced diet having a good fiber and carbohydrate intake, engaging in regular physical activity and avoiding sugary and high-calorie foods and drinks.

The lack of sleep is a contributor as well. Lastly, stress can cause HbA1c levels to rise through a variety of physiological and behavioral mechanisms so it is important to manage stress through techniques such as deep breathing, meditation, exercises or yoga. Nevertheless, this list is not exhaustive but are the basic actionable steps one should take notice of.

As diabetes can bring about a range of effects such as cardiovascular diseases, nerve, kidney and eye damage along with depression and cognitive decline including dementia, it is of great importance for an individual to be educated around the topic. With the growing unhealthy eating habits and increasingly stressful lives, it is important for an individual to take up ownership in both checking and maintaining their HbA1c levels to prevent the development of diabetes itself. Hence we hope that we have brought forward the importance of our insights to you and how you should tackle the global growing issue called diabetes.



That brings us to the end of our presentation . Thank you