



# DSAI Mini Project Team 8

Done by: Joshua, Harith and Rohan



# Table of contents

**01**

**Practical  
Motivation**

**02**

**Exploratory Data  
Analysis**

**03**

**Core Analysis**

**04**

**Insights**



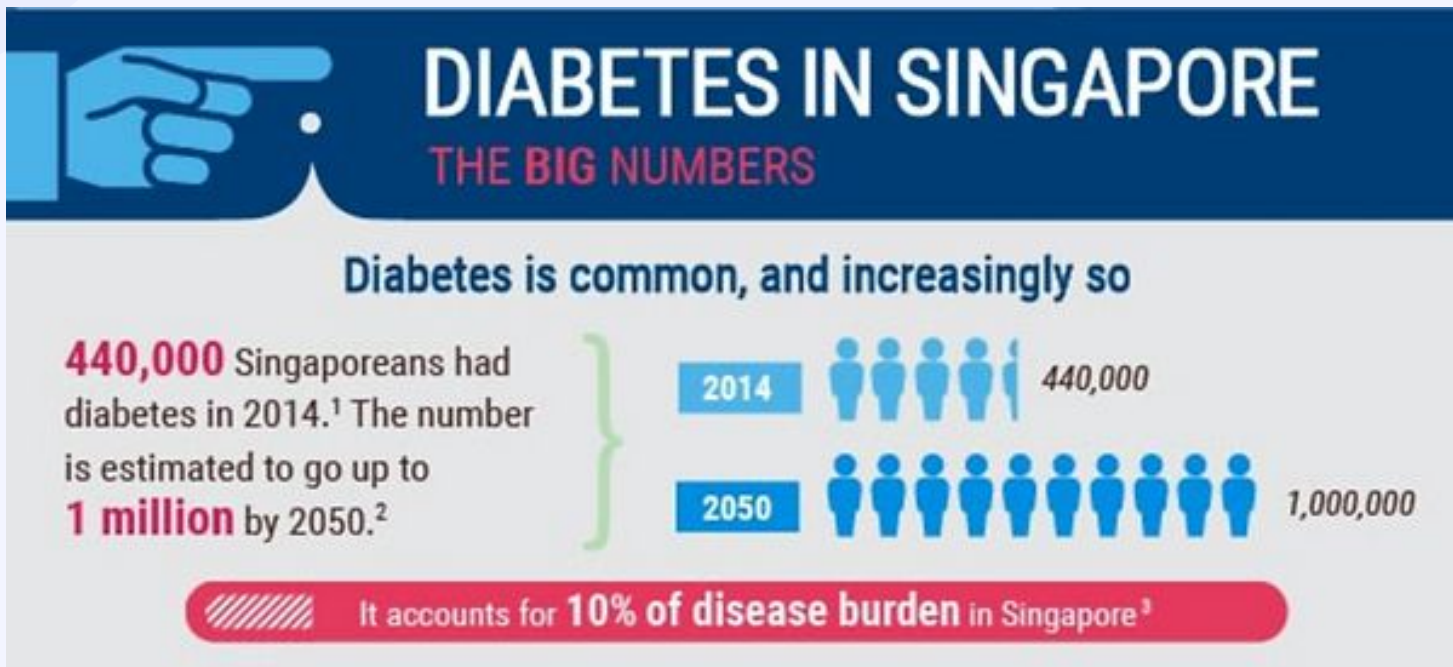
# 01 Practical Motivation

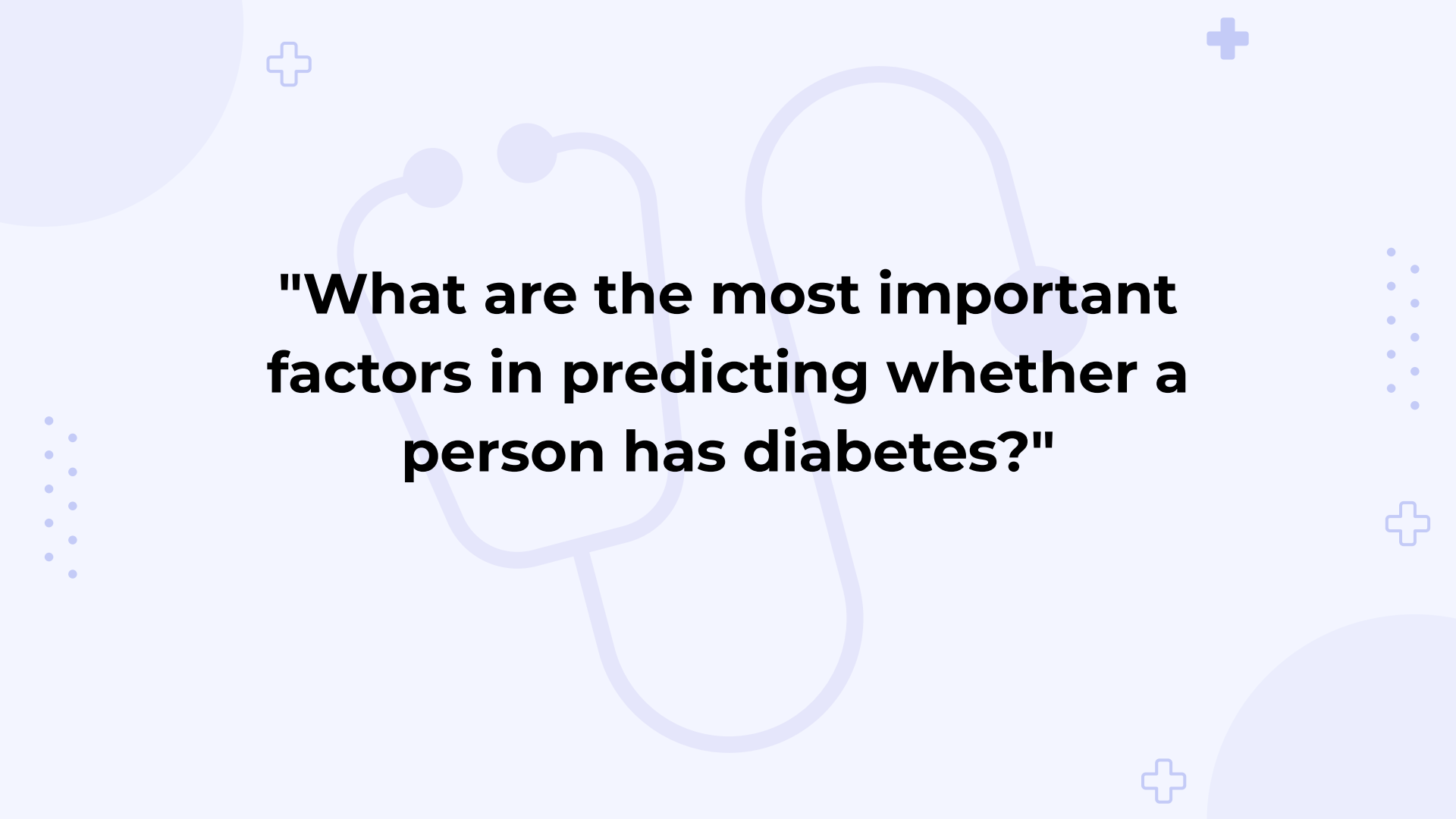




# 422 Million

Individuals has diabetes worldwide according to  
the World Health Organisation (2022)



The background is a light blue gradient. A large, faint, light blue stethoscope is centered behind the text. There are several small, light blue plus signs scattered around the edges. On the left and right sides, there are vertical dotted lines of small light blue dots. Large, faint light blue circles are also visible in the corners.

**"What are the most important factors in predicting whether a person has diabetes?"**

# 02 Exploratory Data Analysis



# Initial Data Insights

Categorical Variable	Relationship with Diabetes (response variable)
Blood Glucose Levels	Proportion increases with blood glucose levels
Age	Proportion increases with age
Smoking History	Proportion may increase slightly with smoking history





# Data Collection/Curation/Cleaning

In preparing the data for our diabetes prediction problem, we followed a systematic approach to ensure data quality and relevance. Here's how we collected, curated, cleaned, and prepared the data:

## Data Collection:

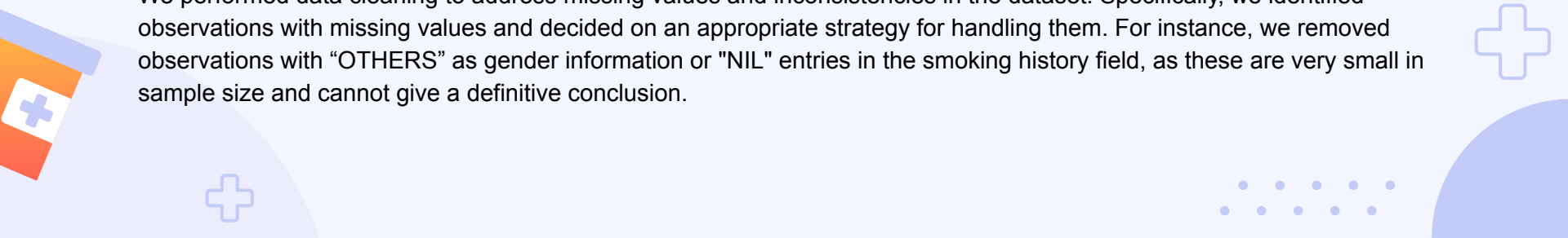
We obtained the diabetes prediction dataset from a reputable source or database. The dataset includes information on various demographic and clinical factors such as gender, age, blood glucose level, smoking history, HbA1c level, and diabetes status.

## Data Curation:

Upon obtaining the dataset, we carefully reviewed its structure and content to ensure accuracy and completeness. We checked for any inconsistencies or anomalies in the data that could affect our analysis.

## Data Cleaning:

We performed data cleaning to address missing values and inconsistencies in the dataset. Specifically, we identified observations with missing values and decided on an appropriate strategy for handling them. For instance, we removed observations with "OTHERS" as gender information or "NIL" entries in the smoking history field, as these are very small in sample size and cannot give a definitive conclusion.

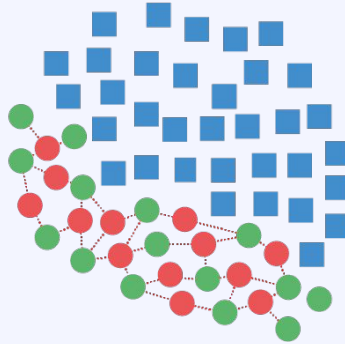


# What else did we learn?

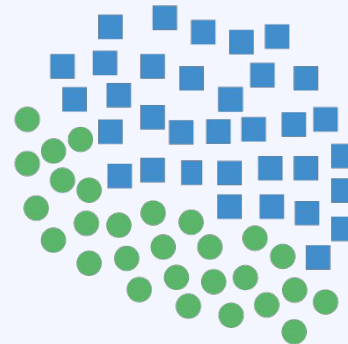
## Synthetic Minority Oversampling Technique (SMOTE)



Original Dataset



Generating Samples



Resampled Dataset

# What else did we learn?

## One Hot Encoding

smoking_history
No Info
No Info
former
never
current

# 03 Core Analysis



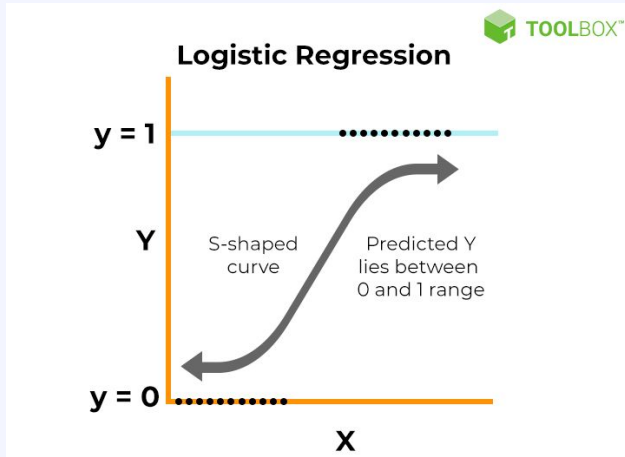


# Objective

Predict **diabetes risk** based on available features in the dataset, using logistic regression, random forest tree classifier, naive Bayes, and neural network analysis.

# How do they help?

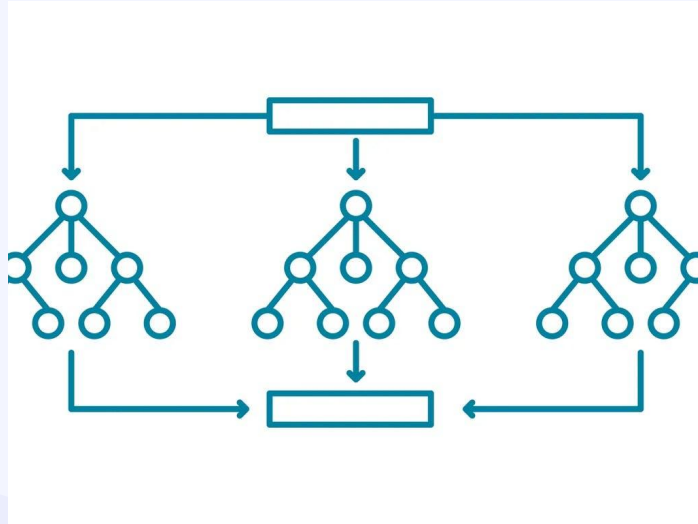
## Logistic Regression



- Linear model used for binary classification tasks
- Models the probability that an instance belongs to a specific class
- Provides interpretable coefficients for ease of understanding

# How do they help?

## Random Forest Classifier



- Builds multiple decision trees during training
- Handles both numerical and categorical features effectively
- Captures non-linear relationships and interactions between features



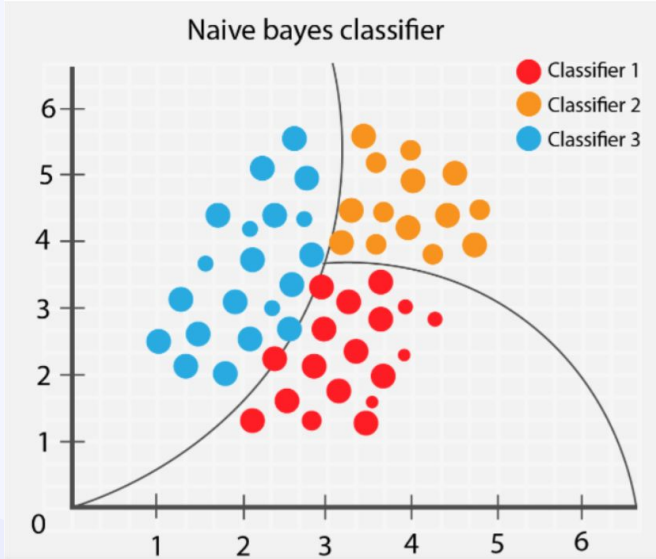
# What else did we learn?

What new techniques did we learn about and implement?



# How do they help?

## Naive Bayes Classifier



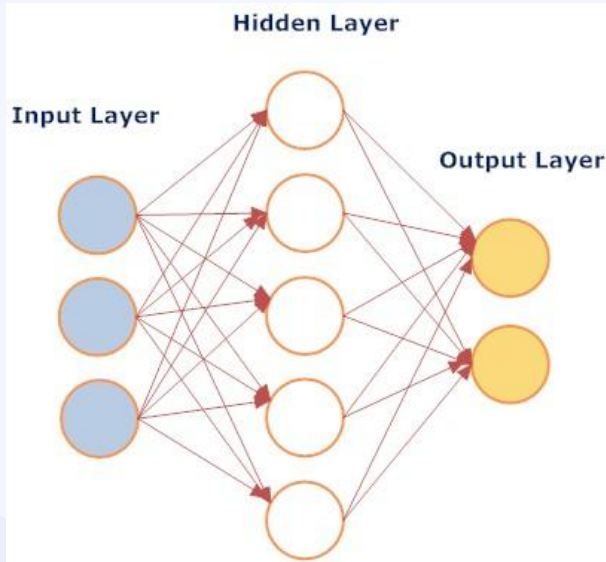
- Probabilistic classifier based on Bayes' theorem with the assumption of feature independence

- Performs well in practice, especially with high-dimensional data

- It is computationally efficient and requires minimal training data

# How do they help?

## Neural Network Analysis



- Highly flexible and capable of learning intricate patterns in data
- Consist of multiple layers of interconnected nodes that can capture complex relationships
- Excel at handling large and complex datasets



# How do they help?


Logistic Regression is a linear model used for binary classification tasks, making it suitable for predicting binary outcomes like diabetes. It models the probability that a given instance belongs to a particular class (e.g., diabetic or non-diabetic) based on the linear combination of input features. Logistic regression provides interpretable coefficients, allowing us to identify the relative importance of each feature in predicting diabetes risk.

Random Forest Tree Classifier is an ensemble learning method that builds multiple decision trees during training and combines their predictions through voting or averaging.

It is robust against overfitting and can handle both numerical and categorical features effectively.

Random forest captures non-linear relationships and interactions between features, making it suitable for capturing complex patterns in the dataset.

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of feature independence. Despite its simple assumption, Naive Bayes can perform well in practice, especially with high-dimensional data like the diabetes dataset. It is computationally efficient and requires fewer training data compared to other algorithms, making it suitable for quick predictions.



Neural networks are highly flexible and capable of learning intricate patterns in data. They consist of multiple layers of interconnected nodes (neurons) that can capture complex relationships between features. Neural networks excel at handling large and complex datasets, automatically learning hierarchical representations of the data.



By leveraging these techniques, we aim to explore the diabetes dataset comprehensively, uncovering predictive features



# How did we apply ML to solve

After analyzing all the possible variables, we concluded that HbA1c and Blood Glucose have the highest correlation with diabetes. Naturally, we tried each of the machine learning techniques to further investigate the relationship between both measurement variables, HbA1c and Blood Glucose, and the nominal variable, diabetes. Random forest tree classification proved to have the highest accuracy of 0.964, so we then used it to explore the relationship between the variables singularly, meaning HbA1c against Diabetes and Blood Glucose against Diabetes. Though the accuracy was lowered to 0.938 and 0.932 respectively, they are still generally high values for accuracy, thus concluding that each of them have a very strong relationship with diabetes.

We then used random tree classification to investigate the relationship between both measurement variables, HbA1c and Blood Glucose, and other variables with high correlation with diabetes. From our earlier visualisations, we selected heart disease and hypertension. HbA1c and Blood Glucose proved to have a strong relationship against both heart disease and hypertension.

#There is high accuracy in detecting heart disease and hypertension with the same variables

#High blood glucose and Hb1Ac levels affects these factors as well

#Explains why many who experience these also experience diabetes

# What else did we learn?

Though we used course content such as logistic regression and random forest tree classification, we also employed new tools. We incorporated Naive Bayes and neural network analysis into our diabetes prediction model introduces exciting opportunities to leverage their unique capabilities. Naive Bayes, with its assumption of feature independence, can provide a simple yet effective approach for classification. On the other hand, neural network analysis, with its ability to learn complex patterns in data, offers a more sophisticated modeling technique.

# 04 Insights





# Outcome

Machine Learning models:

**Random Tree Classifier >**

**Accuracy: 0.837**

**Logistic Regression**

**Naive Bayes**

**Neural Network**



# Outcome

## Random Tree Classifier:

**High Accuracy**

**High F1-Score**

**Low FP + FN**

Accuracy: 0.8369436786136274

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.88	0.85	11553
1	0.87	0.79	0.83	11298
accuracy			0.84	22851
macro avg	0.84	0.84	0.84	22851
weighted avg	0.84	0.84	0.84	22851

Confusion Matrix:

```
[[10209 1344]  
 [ 2382 8916]]
```



# Outcome

**Hb1Ac level and  
Blood Glucose Level  
have Relatively  
Strong Relationship  
with Diabetes**

Accuracy: 0.8369436786136274

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.88	0.85	11553
1	0.87	0.79	0.83	11298
accuracy			0.84	22851
macro avg	0.84	0.84	0.84	22851
weighted avg	0.84	0.84	0.84	22851

Confusion Matrix:

```
[[10209 1344]  
 [ 2382 8916]]
```

Mean Squared Error: 0.16305632138637258

# Outcome

2 Variables -> Room for Error

More Variables improve Accuracy

**All Variables:**

**Accuracy - 0.976**

**F1 Score - 0.98**

Accuracy: 0.9757559844208131

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	11553
1	0.99	0.96	0.98	11298
accuracy			0.98	22851
macro avg	0.98	0.98	0.98	22851
weighted avg	0.98	0.98	0.98	22851

Confusion Matrix:

```
[[11448  105]
 [  449 10849]]
```

# Outcome

## 4 Variables Comparison

Accuracy: 0.8850816156842152

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.91	0.89	11553
1	0.90	0.86	0.88	11298
accuracy			0.89	22851
macro avg	0.89	0.88	0.88	22851
weighted avg	0.89	0.89	0.89	22851

Confusion Matrix:

```
[[10466 1087]
 [ 1539 9759]]
```

**Acc: 0.886**

Accuracy: 0.9700231937333158

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	11553
1	0.98	0.96	0.97	11298
accuracy			0.97	22851
macro avg	0.97	0.97	0.97	22851
weighted avg	0.97	0.97	0.97	22851

Confusion Matrix:

```
[[11286 267]
 [ 418 10880]]
```

**Acc: 0.970**

Accuracy: 0.8754540282700976

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.96	0.89	11553
1	0.95	0.79	0.86	11298
accuracy			0.88	22851
macro avg	0.89	0.87	0.87	22851
weighted avg	0.89	0.88	0.87	22851

Confusion Matrix:

```
[[11040 513]
 [ 2333 8965]]
```

**Acc: 0.875**





**Gender &  
Smoking History**

**BMI & Age**

**Hypertension &  
Heart Disease**



# Insights

- Older generation and those of higher BMI are at risk of Diabetes
  - High glucose and Hb1Ac are strong indicators for Diabetes
- 
- 
- 
- 

# Recommendation

- Avoid consuming sugary food
- Exercise regularly
- Regular check ups





# Thank You!

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#) and infographics & images by [Freepik](#)

Please keep this slide for attribution