SC1015
Mini-Project

Diabetes Prediction System

Z139, Group 10
Nanditha Kumar (U2221998B)
Kalidoss Madhumitha (U2222332A)
Praveena Vijayan (U2222549G)

Table of Contents

01 Problem Formulation

02 Data Collection

03 Data Preparation

04 Exploratory Data Analysis

05 Machine Learning

06 Conclusion

01

Problem Formulation

Diabetes

- Global epidemic
- 7th leading cause of death by 2030
- Crucial to take proactive measures In preventing this chronic illness

Our Objective

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

02

Data Collection

Data Collection

Diabetes Health Indicators Dataset

- Big Dataset comprising of two sub datasets
- Total of ~320, 000 Data Points
- Mostly categorical data
- 21 predictor variables and 1 response variable

Diabetes Data Set

Predict a Model to detect Person has Diabetes or Not

- Small Dataset
- 2000 Data Points
- Mostly numerical data
- 8 predictor variables and 1 response variable

03

Data Preparation

Data Cleaning Process

Drop unnecessary variables

- From Big Dataset
- Examples of variables dropped: Fruits, Veggies, AnyHealthcare

Drop duplicates & 0s

- Dropped duplicates from big & small data.
- Filtered data points with 0s for variables that possibly cannot be zero and dropped those rows.

Impute null values

- 700 sampled points from big data.
- 300 sampled points from small data.
- KNN Imputer

Before Imputation

	index	Diabetes_binary	HighChol	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	HvyAlcoholConsump
0	140577	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0
1	224647	0.0	1.0	27.0	1.0	0.0	0.0	0.0	0.0
2	37634	1.0	1.0	22.0	0.0	0.0	0.0	1.0	0.0
3	30040	0.0	0.0	23.0	1.0	0.0	0.0	1.0	0.0
4	69582	0.0	0.0	23.0	1.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...
995	721	0.0	NaN	38.1	NaN	NaN	NaN	NaN	NaN
996	134	0.0	NaN	21.1	NaN	NaN	NaN	NaN	NaN
997	405	0.0	NaN	42.1	NaN	NaN	NaN	NaN	NaN
998	91	0.0	NaN	32.0	NaN	NaN	NaN	NaN	NaN
999	187	1.0	NaN	32.0	NaN	NaN	NaN	NaN	NaN

Before Imputation

MentHlth	PhysHlth	DiffWalk	AgeLevel	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	DiabetesPedigreeFunction
0.0	8.0	0.0	7.0	NaN	NaN	NaN	NaN	NaN	NaN
0.0	10.0	1.0	9.0	NaN	NaN	NaN	NaN	NaN	NaN
1.0	2.0	0.0	10.0	NaN	NaN	NaN	NaN	NaN	NaN
3.0	0.0	0.0	7.0	NaN	NaN	NaN	NaN	NaN	NaN
30.0	7.0	0.0	10.0	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
NaN	NaN	NaN	1	1.0	114.0	66.0	36.0	200.0	0.289
NaN	NaN	NaN	2	2.0	96.0	68.0	13.0	49.0	0.647
NaN	NaN	NaN	2	2.0	123.0	48.0	32.0	165.0	0.520
NaN	NaN	NaN	3	4.0	123.0	80.0	15.0	176.0	0.443
NaN	NaN	NaN	3	1.0	128.0	98.0	41.0	58.0	1.321

After Imputation

	Diabetes_binary	HighChol	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	HvyAlcoholConsump	MentHlth
140577	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0	0.0
224647	0.0	1.0	27.0	1.0	0.0	0.0	0.0	0.0	0.0
37634	1.0	1.0	22.0	0.0	0.0	0.0	1.0	0.0	1.0
30040	0.0	0.0	23.0	1.0	0.0	0.0	1.0	0.0	3.0
69582	0.0	0.0	23.0	1.0	0.0	0.0	1.0	0.0	30.0
...	...	...	...	...	...	...	...	...	...
721	0.0	0.0	38.1	1.0	0.0	0.0	0.0	0.0	11.0
134	0.0	0.0	21.1	0.0	0.0	0.0	1.0	0.0	4.0
405	0.0	0.0	42.1	0.0	0.0	0.0	1.0	0.0	6.0
91	0.0	0.0	32.0	1.0	0.0	0.0	1.0	0.0	9.0
187	1.0	0.0	32.0	0.0	0.0	0.0	1.0	0.0	5.0

After Imputation

PhysHlth	DiffWalk	AgeLevel	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	DiabetesPedigreeFunction
8.0	0.0	7.0	5.0	133.9	75.0	30.0	189.0	0.4661
10.0	1.0	9.0	7.0	134.9	72.0	27.0	140.0	0.6080
2.0	0.0	10.0	6.0	139.9	73.0	26.0	158.0	0.5163
0.0	0.0	7.0	6.0	128.1	70.0	25.0	162.0	0.5814
7.0	0.0	10.0	6.0	132.7	73.0	27.0	151.0	0.5389
...	...	...	...	...	...	...	...	...
6.0	0.0	1.0	1.0	114.0	66.0	36.0	200.0	0.2890
6.0	0.0	2.0	2.0	96.0	68.0	13.0	49.0	0.6470
5.0	0.0	2.0	2.0	123.0	48.0	32.0	165.0	0.5200
8.0	0.0	3.0	4.0	123.0	80.0	15.0	176.0	0.4430
9.0	0.0	3.0	1.0	128.0	98.0	41.0	58.0	1.3210

04

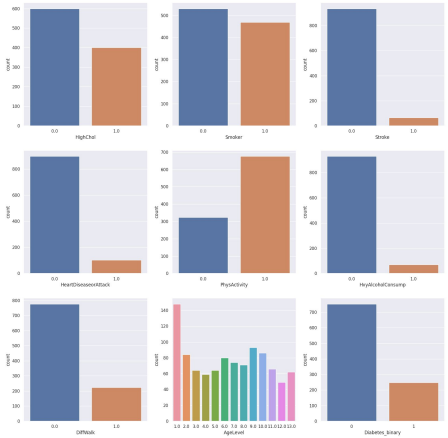
Exploratory Data Analysis

Uni-Variate Data Exploration

Summary statistics for numerical variables

	BMI	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	DiabetesPedigreeFunction	MentHlth	PhysHlth
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000
mean	31.300000	5.537000	129.913000	72.490000	29.231000	174.20100	0.588468	6.327000	7.467000
std	8.518792	2.965033	19.514257	7.973549	7.836648	71.80545	0.208337	7.821589	8.998215
min	0.000000	0.000000	56.000000	24.000000	7.000000	14.00000	0.085000	0.000000	0.000000
25%	25.975000	3.000000	122.000000	69.000000	25.000000	130.00000	0.489000	0.000000	1.000000
50%	30.000000	6.000000	132.000000	74.000000	29.000000	176.00000	0.631750	5.000000	5.000000
75%	35.325000	8.000000	142.000000	76.000000	33.000000	219.00000	0.677325	9.000000	9.000000
max	89.000000	17.000000	198.000000	110.000000	63.000000	744.00000	2.420000	30.000000	30.000000

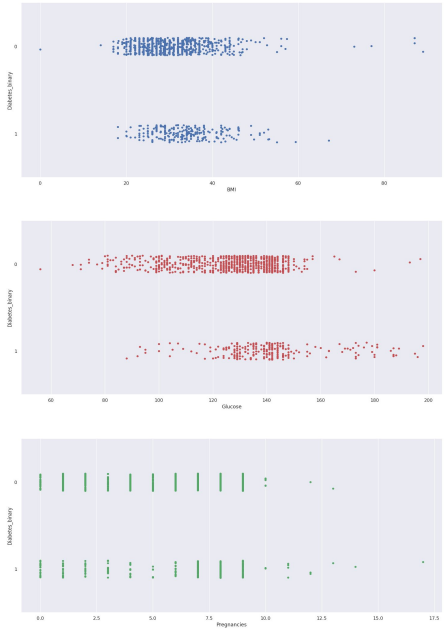
Countplots for categorical variables



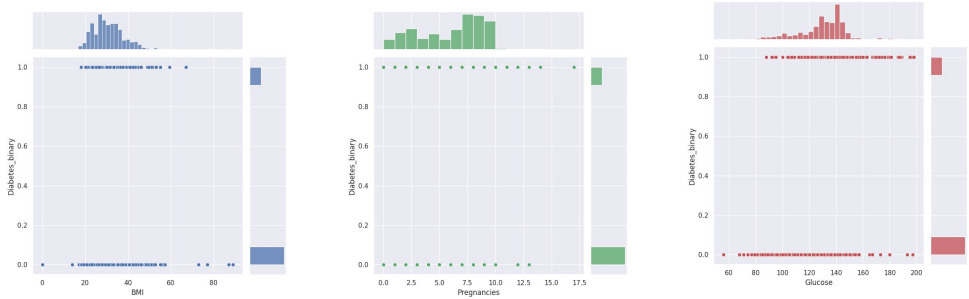
Visualisations for numerical variables

Bi-Variate Data Exploration

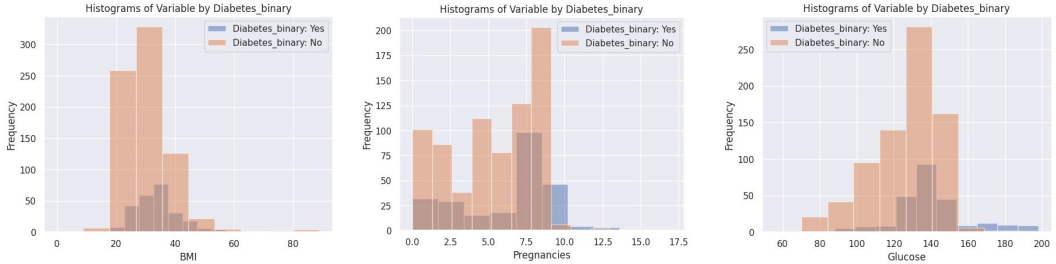
Swarm plot



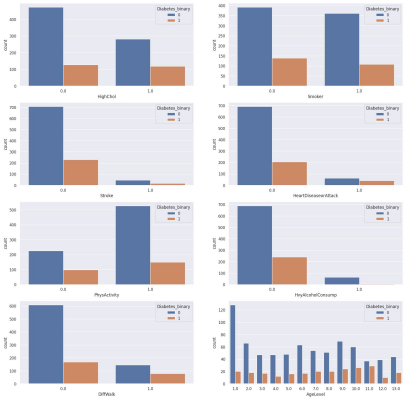
Joint plot



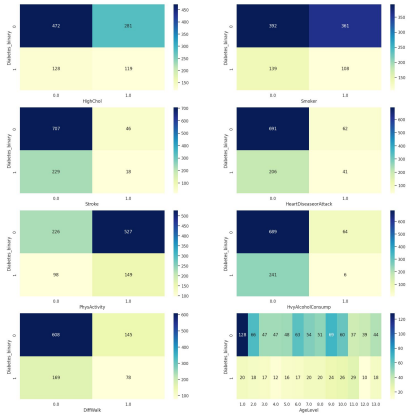
Histogram



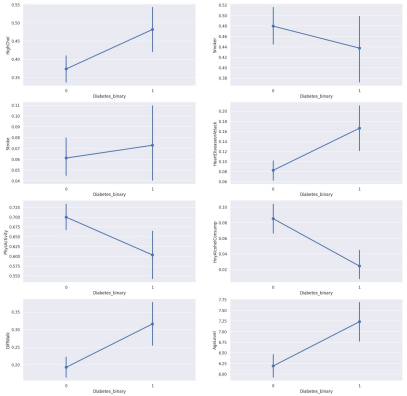
Countplots



Heatmap



Pointplot

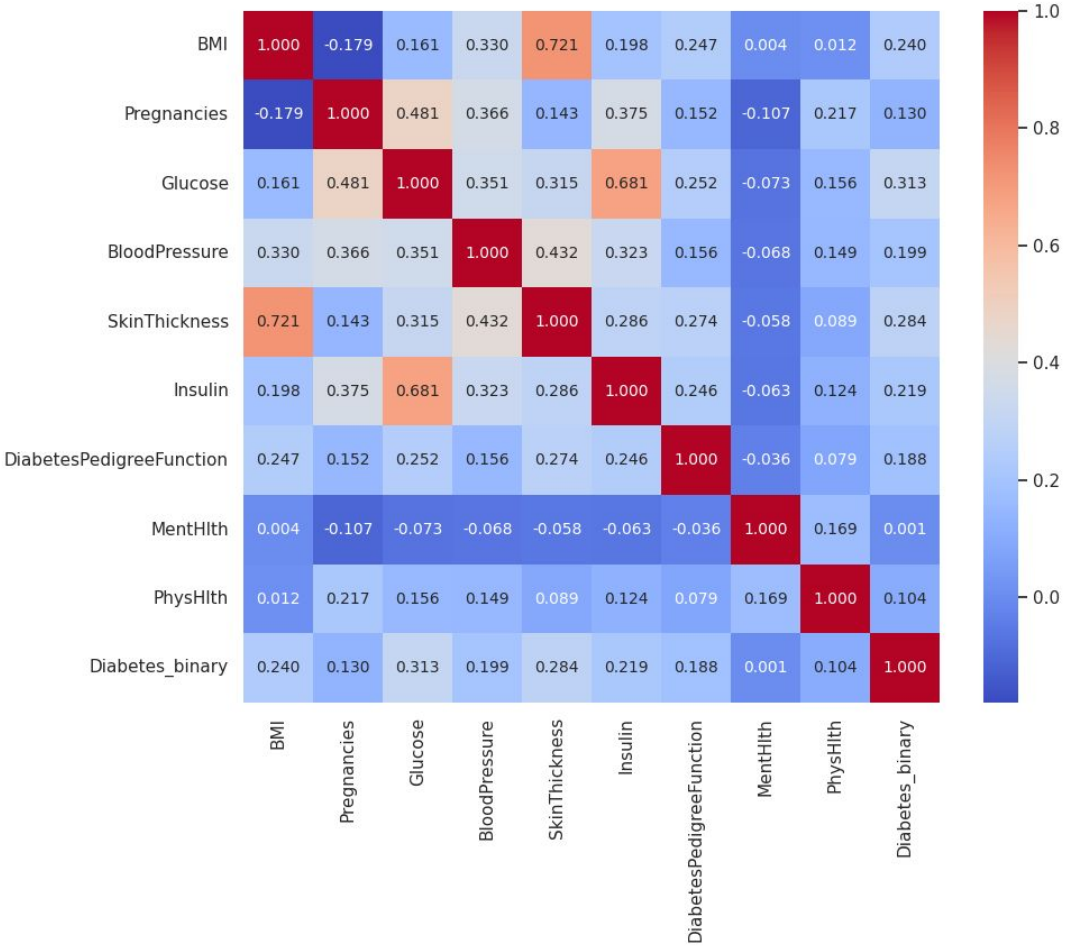


Chi-square test

```
Variable: Smoker
Chi-square statistic: 1.1640932290287953
P-value: 0.28061825547743013
Degrees of freedom: 1
Expected frequencies:
 [[399.843 353.157]
 [131.157 115.843]]
----------------------------------------
Variable: Stroke
Chi-square statistic: 0.25695256966528174
P-value: 0.6122215529313483
Degrees of freedom: 1
Expected frequencies:
 [[704.808  48.192]
 [231.192  15.808]]
----------------------------------------
Variable: HeartDiseaseorAttack
Chi-square statistic: 13.196861820278306
P-value: 0.0002804182810597116
Degrees of freedom: 1
Expected frequencies:
 [[675.441  77.559]
 [221.559  25.441]]
----------------------------------------
Variable: HighChol
Chi-square statistic: 8.69419308819602
P-value: 0.003192254051271038
Degrees of freedom: 1
Expected frequencies:
 [[451.8 301.2]
 [148.2  98.8]]
----------------------------------------
Variable: PhysActivity
Chi-square statistic: 7.493791515599021
P-value: 0.006191206449028067
Degrees of freedom: 1
Expected frequencies:
 [[243.972 509.028]
 [ 80.028 166.972]]
----------------------------------------
Variable: HvyAlcoholConsump
Chi-square statistic: 9.615457913944779
P-value: 0.0019294634637138828
Degrees of freedom: 1
Expected frequencies:
 [[700.29  52.71]
 [229.71  17.29]]
----------------------------------------
Variable: DiffWalk
Chi-square statistic: 15.596051572984061
P-value: 7.841816915333052e-05
Degrees of freedom: 1
Expected frequencies:
 [[585.081 167.919]
 [191.919  55.081]]
----------------------------------------
```

Multi-Variate Data Exploration

Correlation heatmap for numeric variables



Analysis of EDA

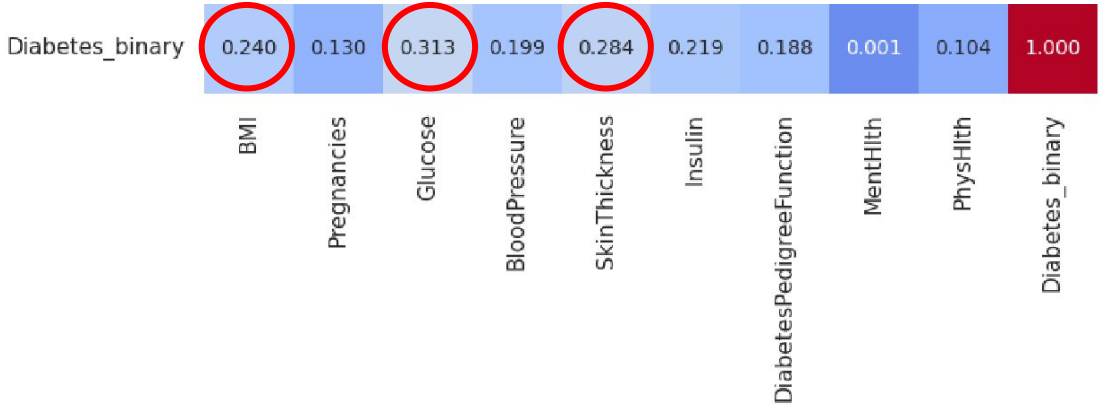
Numerical Variables

Feature variables chosen for machine learning

Glucose

BMI

SkinThickness



Highest correlation with diabetes_binary

Categorical Variables (Binary)

Feature variables chosen for machine learning

DiffWalk

(Difficulty walking or climbing stairs)

HeartDiseaseor Attack

HighChol

(High Cholesterol)

```
----------------------------------------
Variable: DiffWalk
Chi-square statistic: 16.88656311077369
P-value: 3.968149583168322e-05
Degrees of freedom: 1
Expected frequencies:
 [[565.308 172.692]
 [200.692  61.308]]
----------------------------------------
```

```
----------------------------------------
Variable: HeartDiseaseorAttack
Chi-square statistic: 17.775536796397557
P-value: 2.4855866128901678e-05
Degrees of freedom: 1
Expected frequencies:
 [[653.868  84.132]
 [232.132  29.868]]
----------------------------------------
```

```
----------------------------------------
Variable: HighChol
Chi-square statistic: 17.519189721553722
P-value: 2.8442244166505004e-05
Degrees of freedom: 1
Expected frequencies:
 [[436.896 301.104]
 [155.104 106.896]]
----------------------------------------
```

Highest chi-square statistic and lowest p-value with diabetes_binary

05

Machine Learning

Model Selection

Logistic Regression

Random Forest

Support Vector Machine

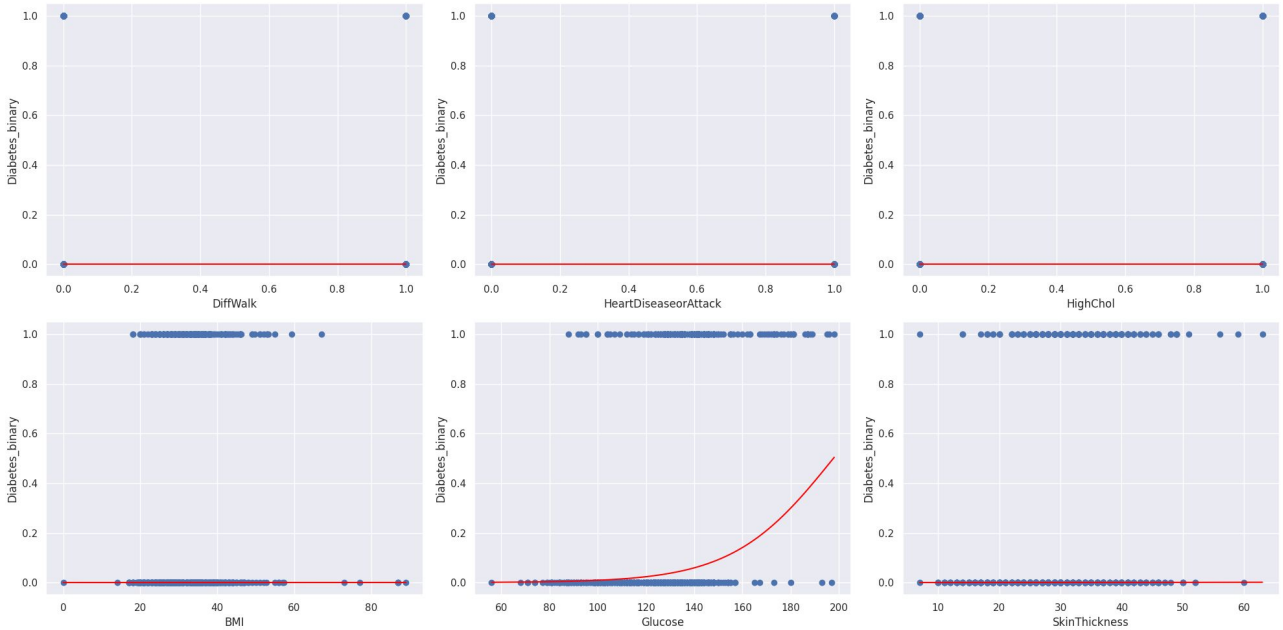
All three models can handle both categorical binary predictor variables (DiffWalk, HeartDiseaseorAttack, HighChol) and numerical predictor variables (BMI, Glucose, SkinThickness). They can also handle the binary response variable (Diabetes_binary).

Model Training

Logistic Regression

- Logistic regression is a statistical method used for binary classification
- It estimates the probability of an event using a logistic function
- An optimization algorithm is used to fit the model parameters
- It can predict the probability of having diabetes based on predictor variables

Logistic Regression



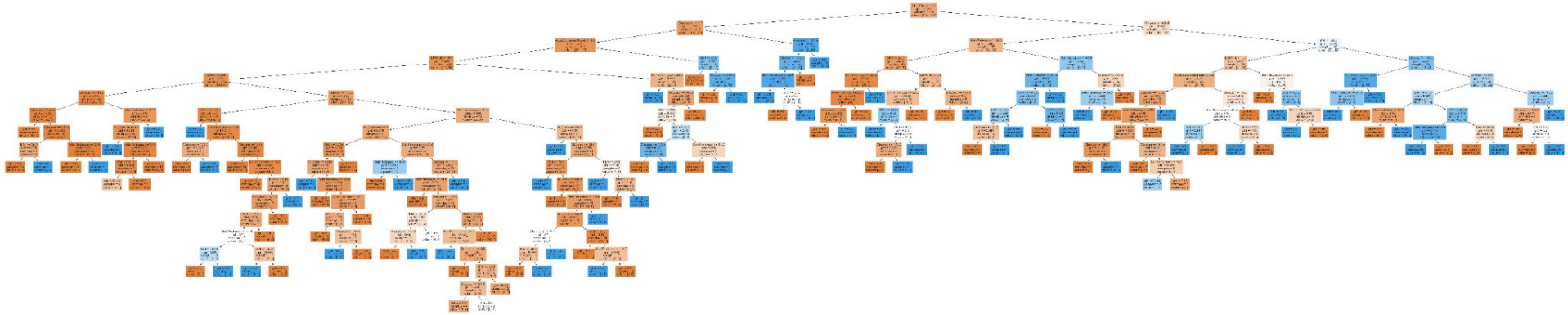
Response variable:
'Diabetes_binary' with 1 indicating presence and 0 indicating absence

Predictor variables:
'Diff Walk', 'HeartDiseaseorAttack', 'HighChol', 'BMI', 'Glucose', and 'SkinThickness'

Random Forest

- 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
- The algorithm aggregates the predictions for the final outcome
- Suitable for classification and regression tasks

Random forest Classification



Variables

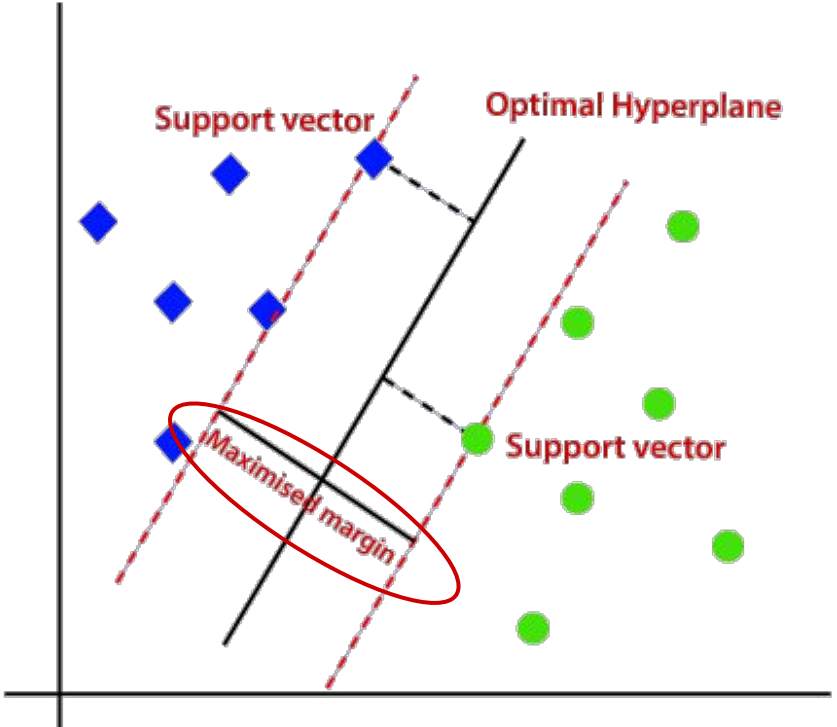
- **Numerical**
- **Categorical (Binary)**

- Splits the data into intervals
- Creates decision trees based on the intervals

- 0.5 is used as the threshold for deciding whether a sample belongs to one class or another.

Support Vector Machine (SVM)

SVM is an algorithm that finds a hyperplane that best separates the data points of different classes and has the maximum margin



Prediction using our models

Logistic Regression Model

```
Patient details:
DiffWalk: 1
HeartDiseaseorAttack: 0
HighChol: 0
BMI: 15
Glucose: 150
SkinThickness: 25
```

```
The patient has no diabetes
Probability of having no diabetes: 0.71
```

SVM Model

```
Patient details:
DiffWalk: 1
HeartDiseaseorAttack: 0
HighChol: 0
BMI: 15
Glucose: 150
SkinThickness: 25
```

```
The patient has no diabetes
Probability: 0.7
```

Random Forest

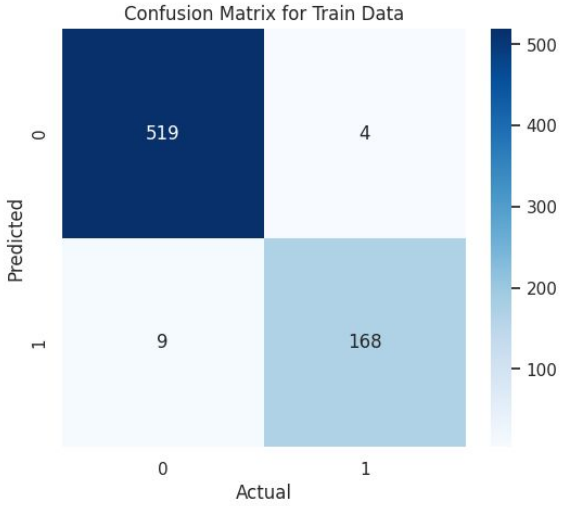
```
Patient details:
HeartDiseaseorAttack: 0
HighChol: 0
DiffWalk: 1
BMI: 15
Glucose: 150
SkinThickness: 25
```

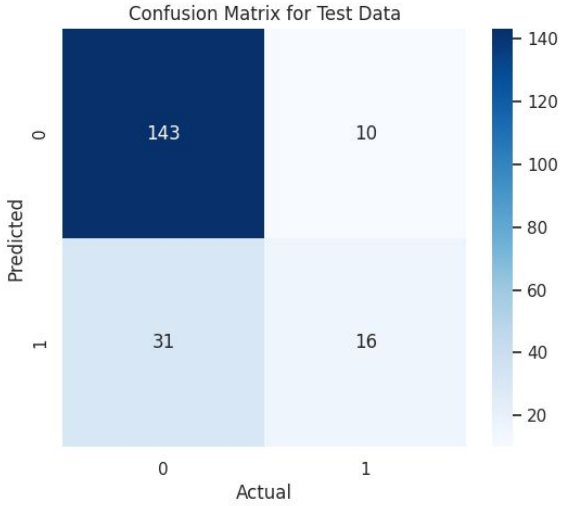
```
The patient does not have diabetes
Probability of not having diabetes: 0.81
```

Consistent and similar prediction results

Model Evaluation

Logistic Regression





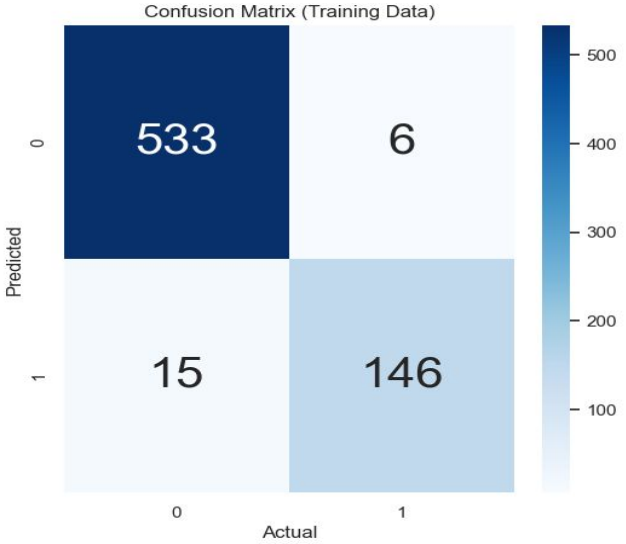
TPR = 0.949
FPR = 0.008
TNR = 0.992
FNR = 0.051

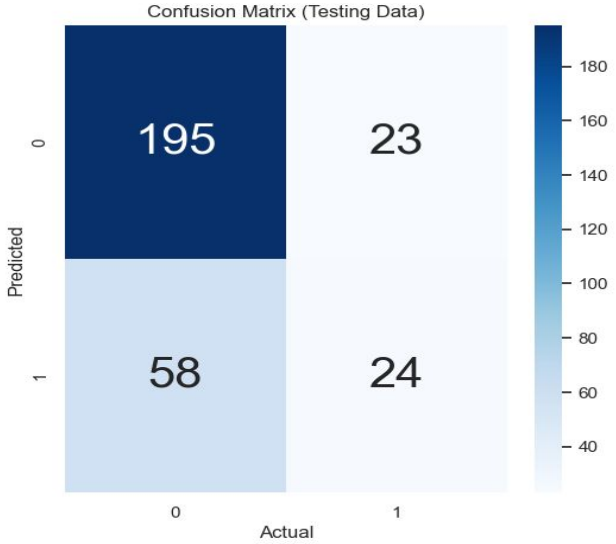
-> Accuracy of 98.8%

TPR = 0.340
FPR = 0.065
TNR = 0.935
FNR = 0.660

-> Accuracy of 79.5%

Random forest classification





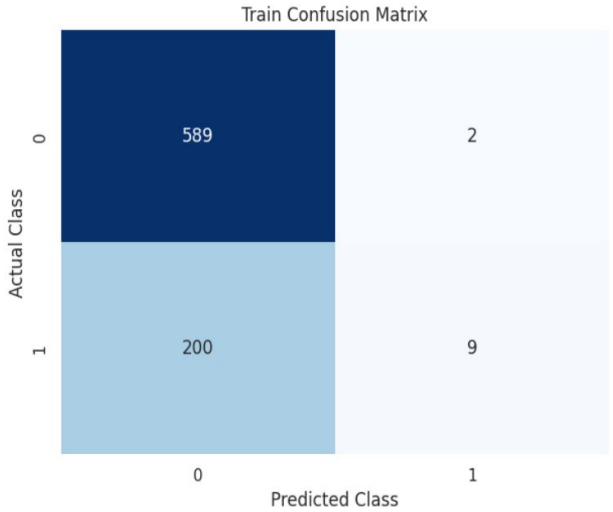
TPR: 0.907
FPR: 0.011
TNR: 0.011
FNR: 0.093

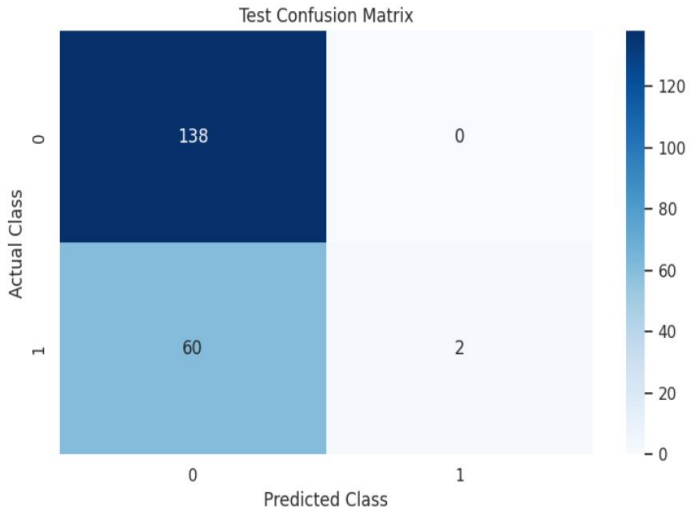
-> Accuracy: 97.0%

TPR: 0.400
FPR: 0.114
TNR: 0.886
FNR: 0.600

->Accuracy: 74.0%

Support Vector Machine





TPR = 0.043
FPR = 0.003
TNR = 0.997
FNR = 0.957

-> Accuracy of 74.8%

TPR = 0.032
FPR = 0
TNR = 1
FNR = 0.968

-> Accuracy of 70.0%

Comparing across all three machine learning models

1) Accuracy Scores

   - ❖ Logistic Regression model has the highest accuracy score for train and test data

2) False Negative Rate (FNR)

   - ❖ Logistic Regression model generally has the lower FNR for train and test data

Logistic Regression Model shows the best prediction for diabetes

Improving the performance of chosen model

Tuning process

- Feature selection
- Ensembling
- Regularisation
- Data augmentation

06

Conclusion

Conclusion

- Develop a machine learning model that accurately discerns the possibility of a person developing diabetes
- Explored new ways of visualising and predicting our data
    - Point plot
    - Histograms of the numeric variable separately for each level of the binary variable
    - Chi square statistic for categorical variable
    - Logistic Regression
    - Random Forest Classification
    - Support Vector Machine Algorithm
- Necessity for Model Evaluation

References

- https://www.einsteinmed.edu/centers/global-health/global-diabetes-institute/about-us/global-diabetes/index.html (7th leading cause of death in 2030)
- https://www.moh.gov.sg/news-highlights/details/speech-by-mr-ong-ye-kung-minister-for-health-at-world-diabetes-day-2021#:~:text=Locally%2C%20one%20in%20three%20individuals,will%20be%20living%20with%20diabetes. (Locally in sg, one in three people get diabetes)
- https://vizhub.healthdata.org/gbd-compare/#0 (Graph of global diabetes)
- https://www.narayanahealth.org/blog/diabetes-mellitus-guide/ (Picture of person taking glucose level)