Diabetes Prediction

Done By: Ronald Teo Boon Keat, Tay Wei Yang, Wong Zhen Wei Gabriel

Table of contents

Introduction

768 Datas of

Diabetic and Non-diabetic Individuals

Diabetes Dataset

Data Card   Code (212)   Discussion (3)

427   New Notebook

About Dataset

Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

Content

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Why Are We Interested In Diabetes?

The International Diabetes Federation (IDF) reports that about 537 million people worldwide have diabetes.

The number of cases has been **increasing over the past few decades,** and the IDF predicts 783 million people will have diabetes **by 2045 — an increase of 46%.**

Problem Statement

To devise a **method for early detection** of diabetes to prevent diabetic complications, using **diagnostic measurements to predict** whether an individual is likely to have diabetes.

Data Preparation
&
Exploratory Analysis

Preliminary Exploration

```
Number of NULL values for each attribute:

Pregnancies                   0
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

- Replaced zeros to NULL
- 30% of data for **Skin Thickness** and **Insulin** attribute each, is missing.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	$\boxed{0.074752}$	$\boxed{0.130548}$	0.292695	0.173844	0.238356	1.000000

Preliminary Exploration

Before

```
Number of NULL values for each attribute:

Pregnancies                   0
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

After

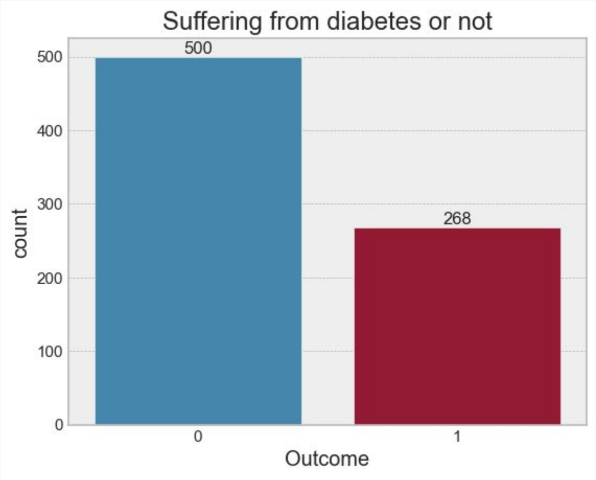
```
Number of NULL values for each attribute:

Pregnancies                 0
Glucose                     0
BloodPressure               0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

- Dropped **Skin Thickness** and **Insulin**
- Replaced NULL values with **median**

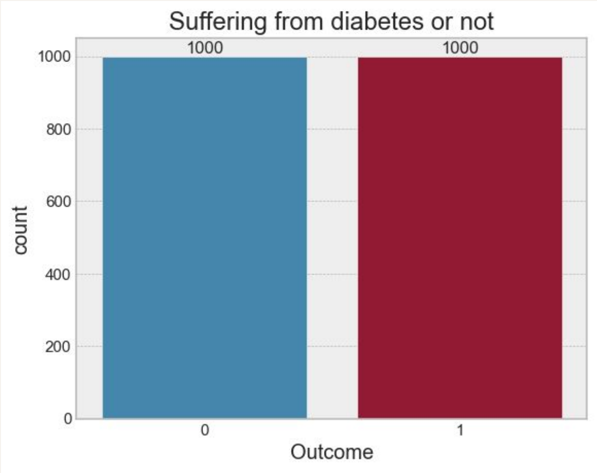
Upsampling

Before

768 Datas
65% Diabetics
35% Non-Diabetics



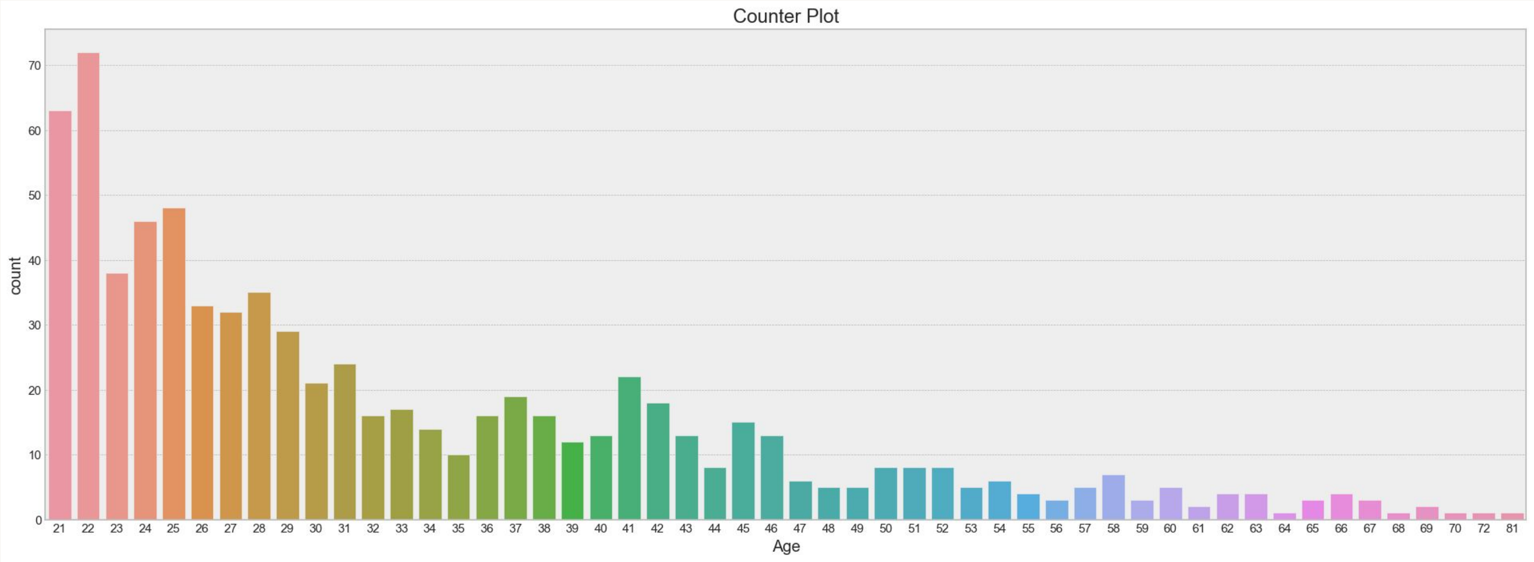
After



2000 Datas
50% Diabetics
50% Non-Diabetics

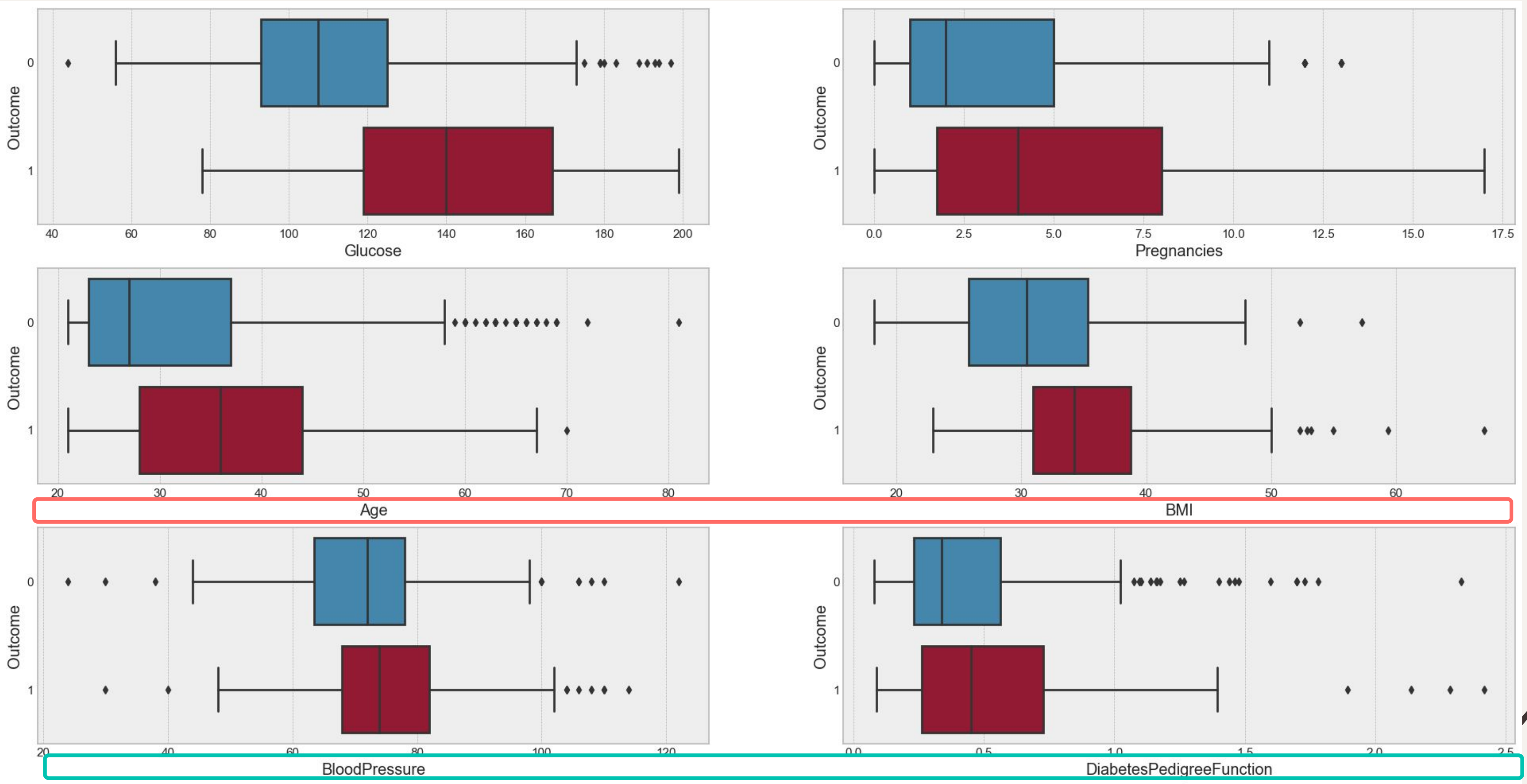
- Scikit resampling function
- Applied to some models elaborated subsequently

Analytic Visualisation

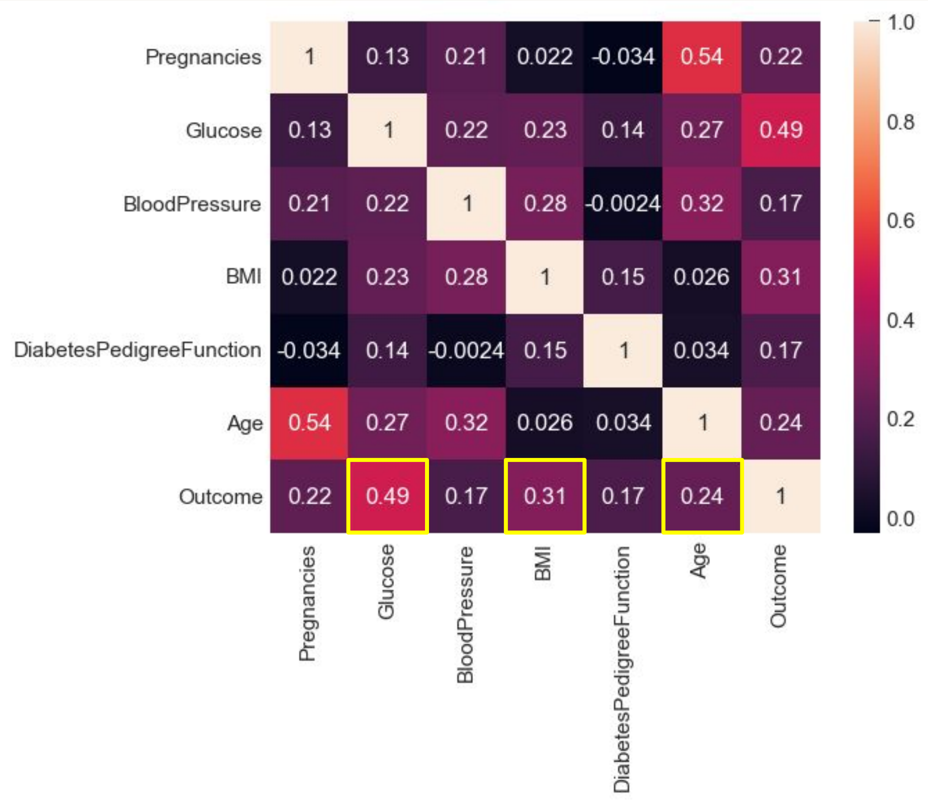


68% of individuals in our dataset is between the age of **21 to 44 years old**.

Boxplot For Each Attribute



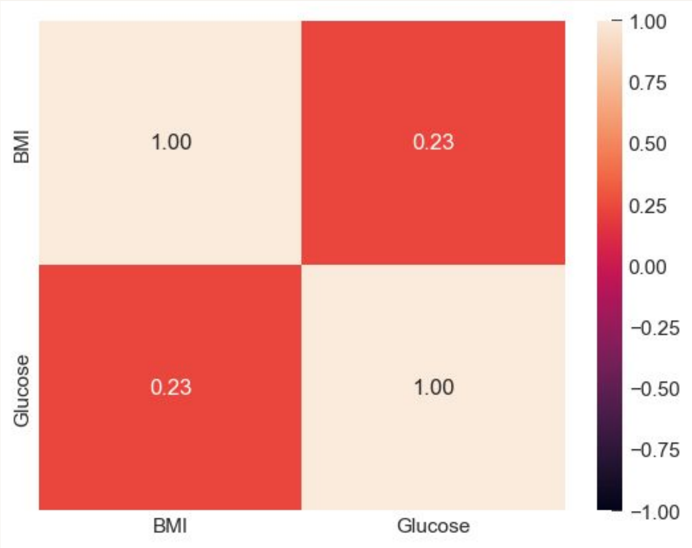
Correlation

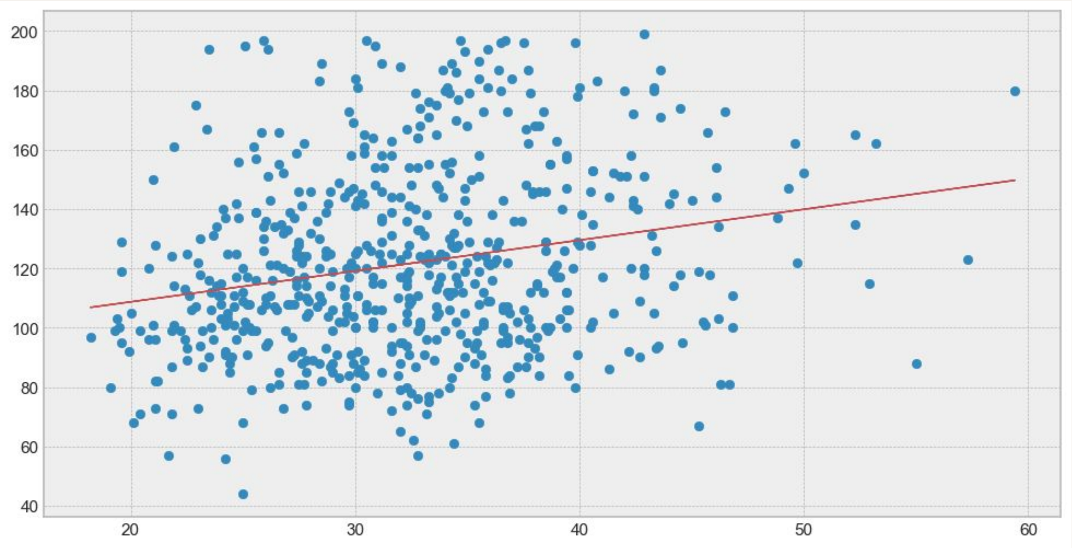


Data Analysis

Linear Regression

BMI & Glucose:

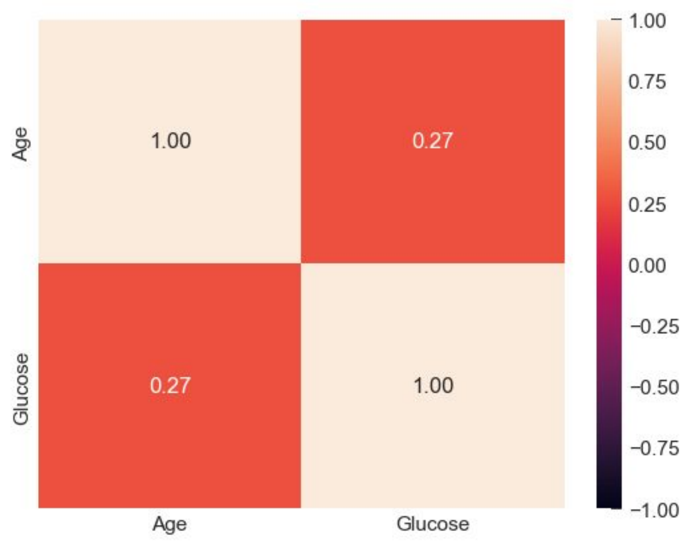


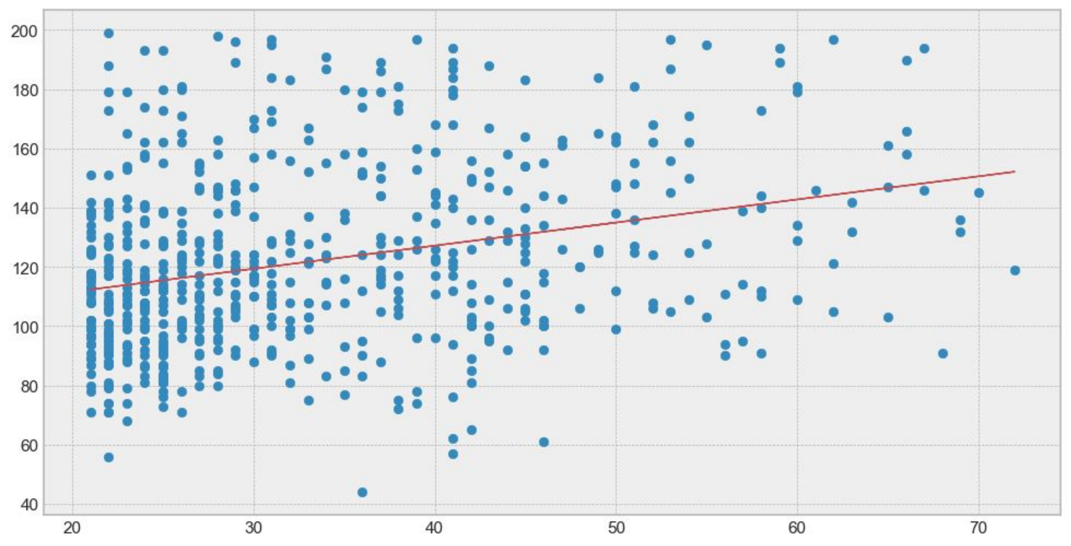


Score : 0.0847
Mean & Median: 31.99, 32

Linear Regression

Age & Glucose:

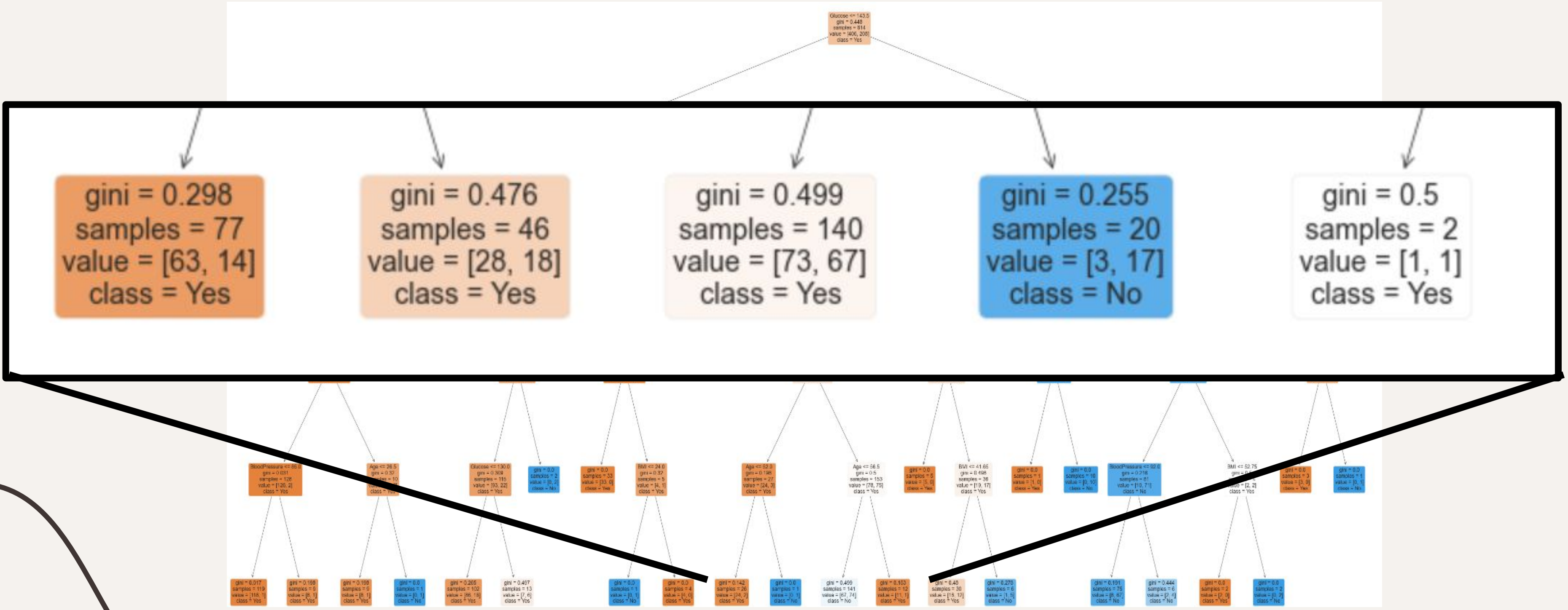




Score : 0.1068

Modeling

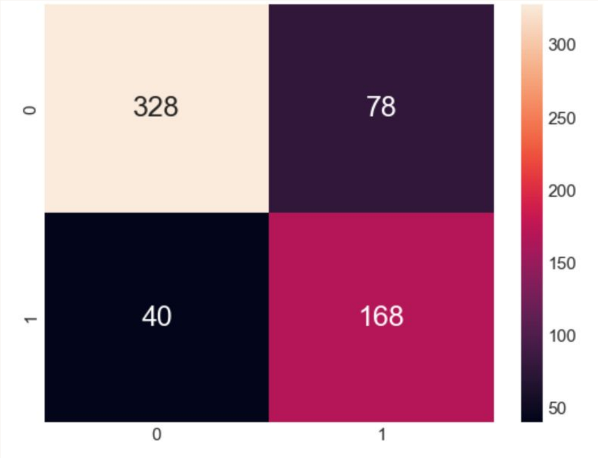
Decision Tree Classifier



Decision Tree Classifier

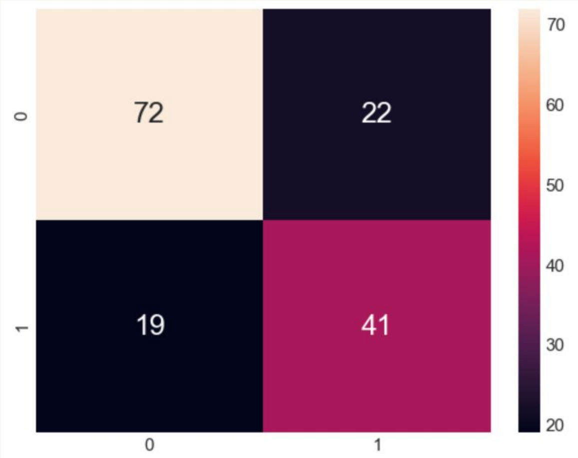
Train

```
Train data:
Classification Accuracy          : 0.8078175895765473
TN: 328
FN: 40
TP: 168
FP: 78
True Positive Rate:  0.8076923076923077
False Positive Rate:  0.1921182266009852
```



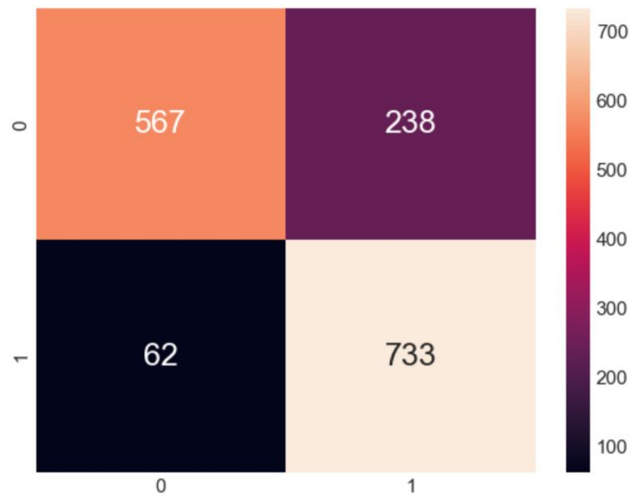
Test

```
Test data:
Classification Accuracy          : 0.7337662337662337
TN: 72
FN: 19
TP: 41
FP: 22
True Positive Rate:  0.6833333333333333
False Positive Rate:  0.23404255319148937
```



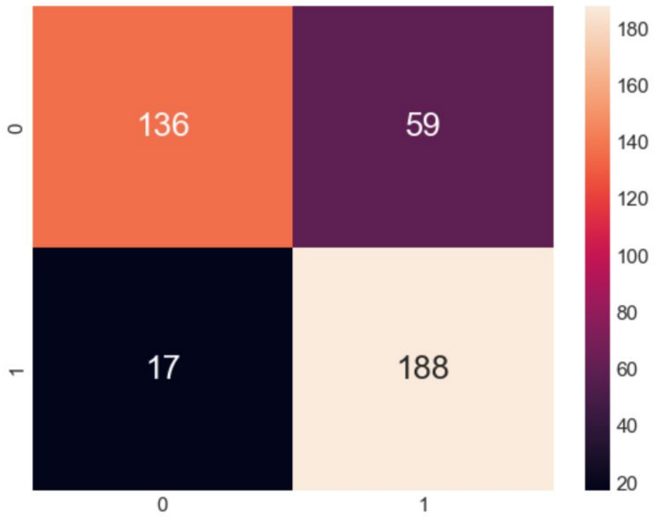
Decision Tree Classifier (With balanced class)

Train




```
Train data:
Classification Accuracy          : 0.8125
TN: 567
FN: 62
TP: 733
FP: 238
True Positive Rate:  0.9220125786163522
False Positive Rate:  0.2956521739130435
```

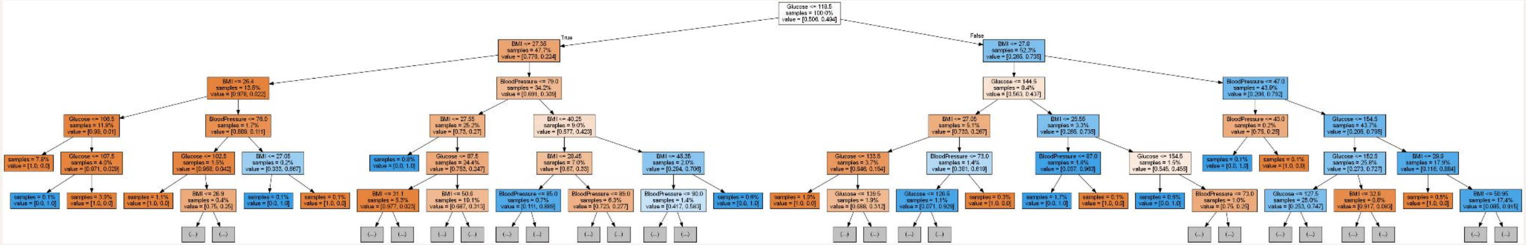
Test

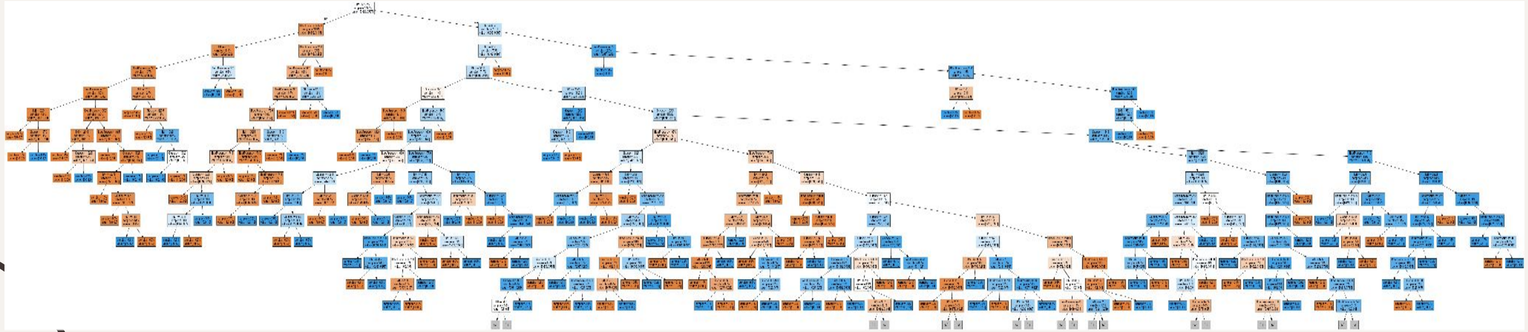



```
Test data:
Classification Accuracy          : 0.81
TN: 136
FN: 17
TP: 188
FP: 59
True Positive Rate:  0.9170731707317074
False Positive Rate:  0.08292682926829269
```

Random Forest

Average Accuracy: 0.96. Depth 5 & 14 respectively:

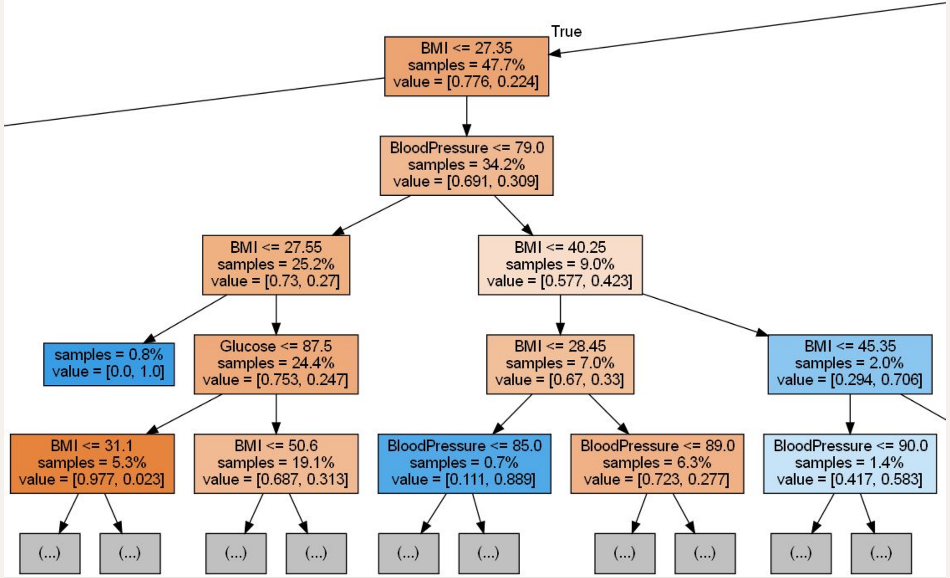




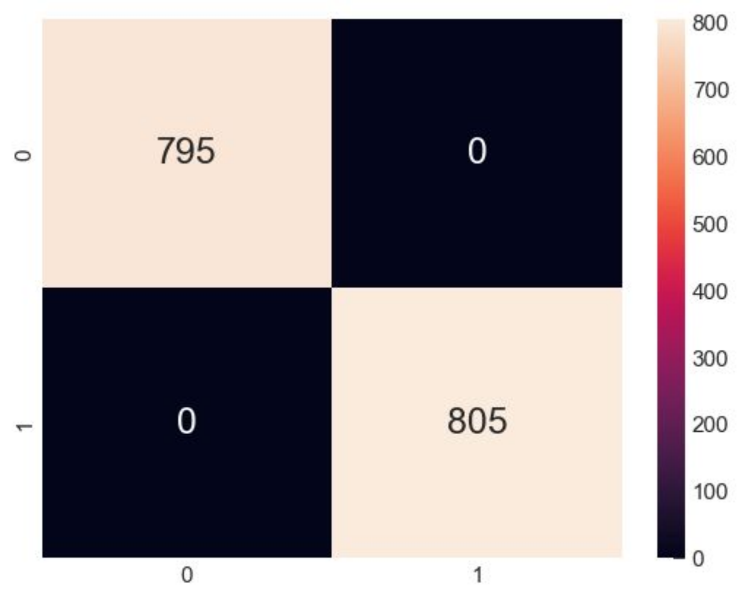
```
Accuracy: 0.9625
Accuracy: 0.9575
Accuracy: 0.9675
Accuracy: 0.9525
Accuracy: 0.9475
Accuracy: 0.96
Accuracy: 0.9425
Accuracy: 0.9525
Accuracy: 0.9475
Accuracy: 0.965
Accuracy: 0.96
Accuracy: 0.975
Accuracy: 0.965
Accuracy: 0.945
Accuracy: 0.9675
Accuracy: 0.965
Accuracy: 0.965
Accuracy: 0.9875
Accuracy: 0.955
Accuracy: 0.96
Average Accuracy: 0.96
```

Random Forest

Depth 5 Zoomed in:

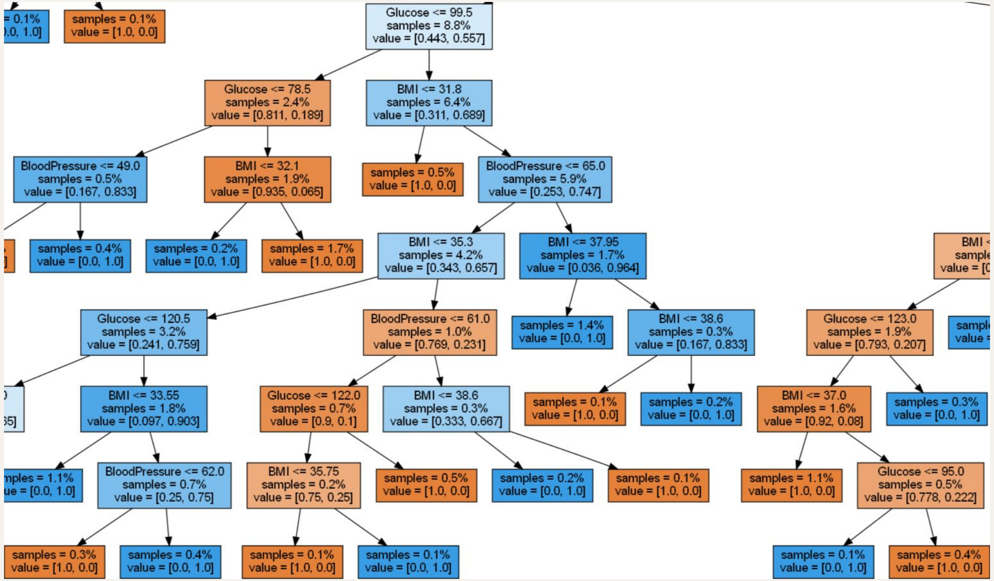


Heatmap:

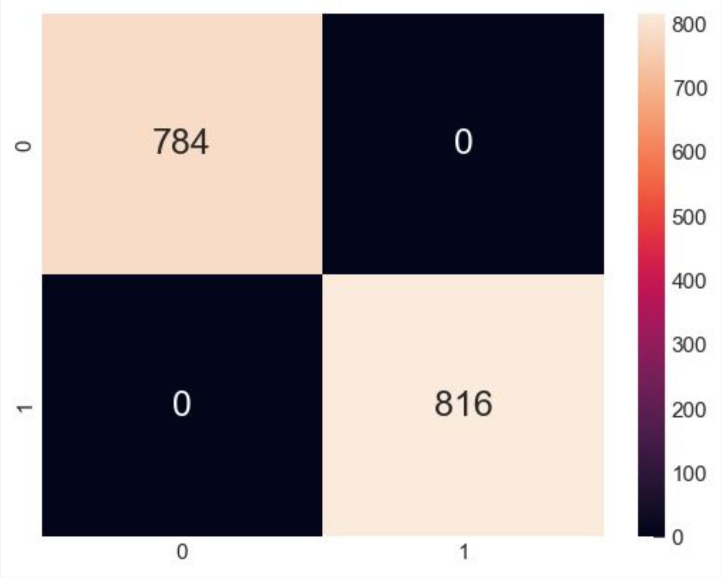


Random Forest

Depth 14 Zoomed in:



Heatmap:



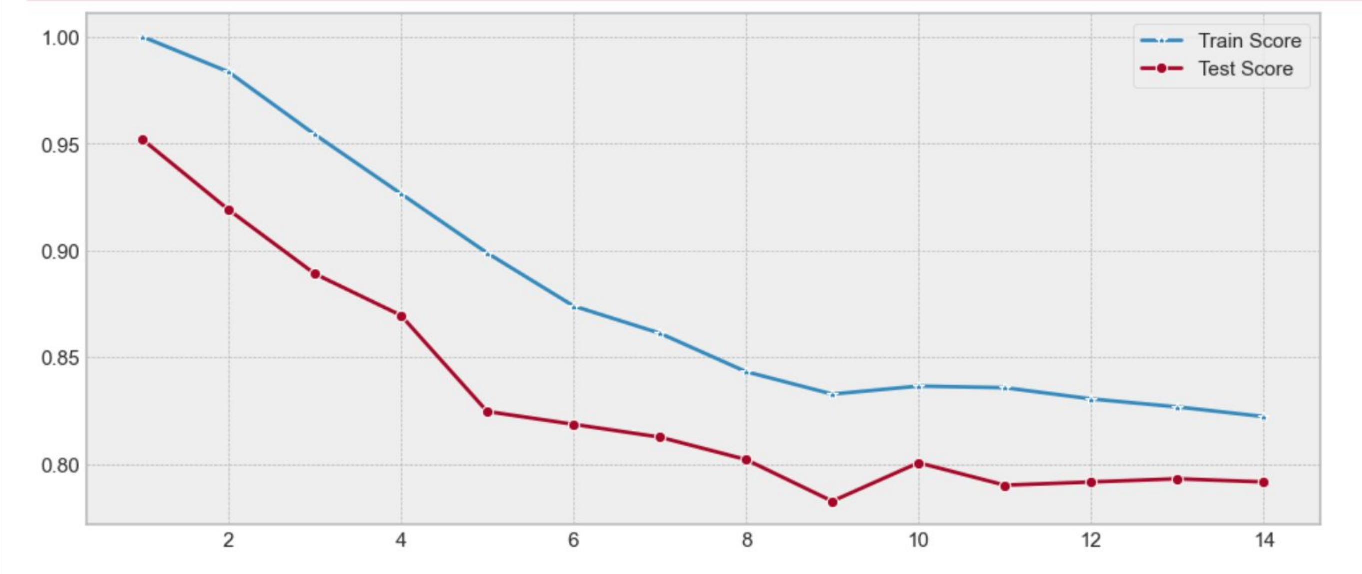
K-Nearest Neighbors (KNN)

1. The algorithm is simple and easy to implement.
2. There’s no need to build a model, tune several parameters, or make additional assumptions.

However,

1. The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

K-Nearest Neighbors (KNN)



Choose a value K with the highest accuracy. (In this case 1)

K-Nearest Neighbors (KNN)


```
Test data:
Classification Accuracy          : 0.952023988005997
TN: 313
FN: 11
TP: 322
FP: 21
True Positive Rate:  0.9669669669669669
False Positive Rate:  0.03303303303303303
```

Comparing accuracy of all Models

Decision tree without class balancing:

```
Test data:
Classification Accuracy         : 0.7337662337662337
TN: 72
FN: 19
TP: 41
FP: 22
True Positive Rate:  0.6833333333333333
False Positive Rate:  0.23404255319148937
```

Decision tree with class balancing:

```
Test data:
Classification Accuracy         : 0.81
TN: 136
FN: 17
TP: 188
FP: 59
True Positive Rate:  0.9170731707317074
False Positive Rate:  0.08292682926829269
```

Random Forest:

```
test data:
Classification Accuracy         : 0.99
TN: 193
FN: 0
TP: 203
FP: 4
True Positive Rate:  1.0
False Positive Rate:  0.02030456852791878
```

K-Nearest Neighbors:

```
Test data:
Classification Accuracy         : 0.952023988005997
TN: 313
FN: 11
TP: 322
FP: 21
True Positive Rate:  0.966966966966967
False Positive Rate:  0.03303303303303303
```

Insights & Recommendations

Insights & Recommendations

1. **Glucose** has the **highest correlation** with diabetic outcome. The root of decision tree utilizes glucose to split the node.

2. **Upsampling** imbalanced data **improved the model performance** especially on the non-diabetic class.

3. **Random forest tree** would be the **best classifier** to determine whether a person is likely to have diabetes with an accuracy of 99%.

Conclusion

Conclusion

- There are **other factors** such as lifestyle and environmental factors that could also influence risk of diabetes.

- The insights gained from this project can help **promote changes in diet and lifestyle** to reduce risk of diabetes.

- With data science and artificial intelligence, it is possible to detect diabetes early with **reasonable accuracy** using models.

New Things We Learnt

- Class imbalance can affect the accuracy of the model
- Going back to clean dataset after modelling can help improve accuracy
- Modelling sometimes requires us to go back to the start to clean our dataset to improve our models
- Using other models like random forest and KNN can help to produce more accurate models

Thank You

Reference

Afolabi, S. (2022, September 17). *Here's what I've learnt about sklearn.resample*. Medium. Retrieved April 14, 2023, from https://towardsdatascience.com/heres-what-i-ve-learnt-about-sklearn-resample-ab735ae1abc4

Brownlee, J. (2021, January 21). *Failure of classification accuracy for imbalanced class distributions*. MachineLearningMastery.com. Retrieved April 14, 2023, from https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

Harrison, O. (2019, July 14). *Machine learning basics with the K-nearest neighbors algorithm*. Medium. Retrieved April 14, 2023, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Khageshor, G. (2023, April 7). *Diabetes Prediction - Eda & Prediction*. Kaggle. Retrieved April 14, 2023, from https://www.kaggle.com/code/khageshorgiri/diabetes-prediction

Yiu, T. (2021, September 29). *Understanding random forest*. Medium. Retrieved April 14, 2023, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2