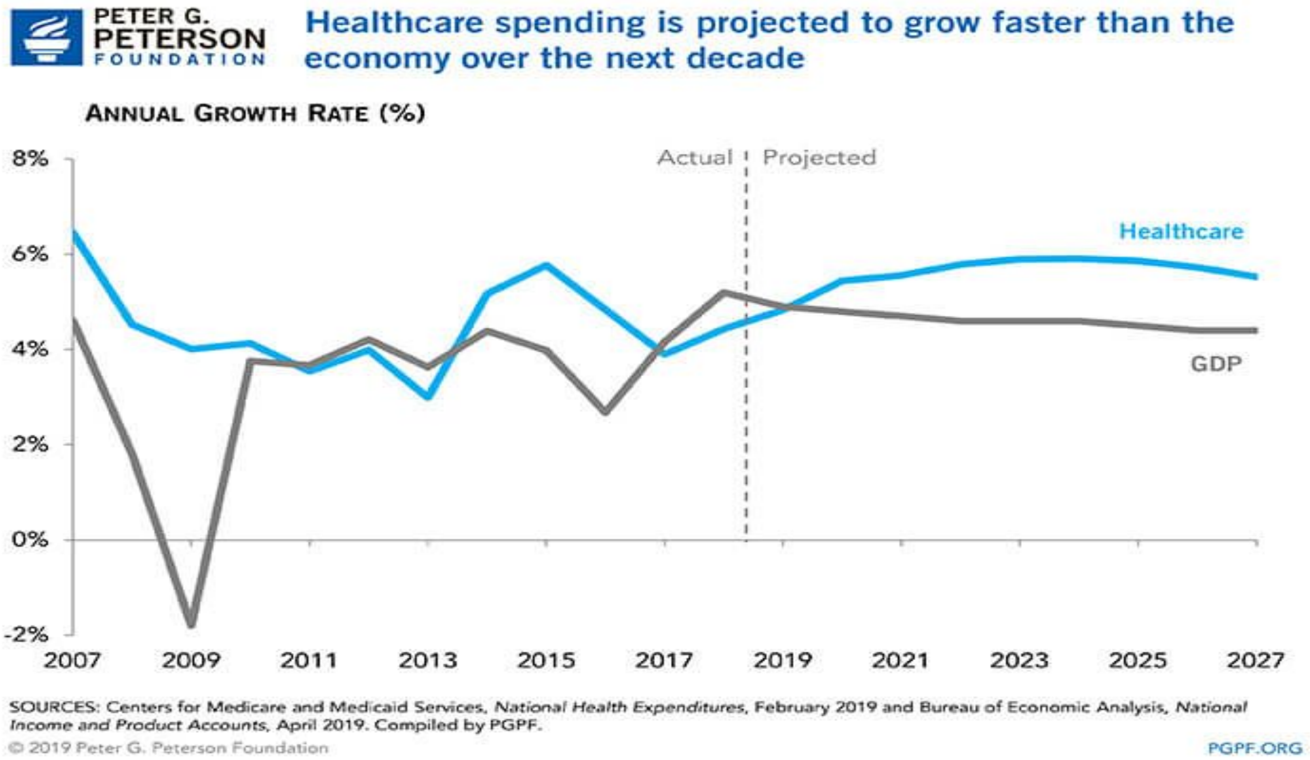
# Evaluation of ML Models to Predict Diabetes

Fathia Mohamed

# Diabetes Crisis

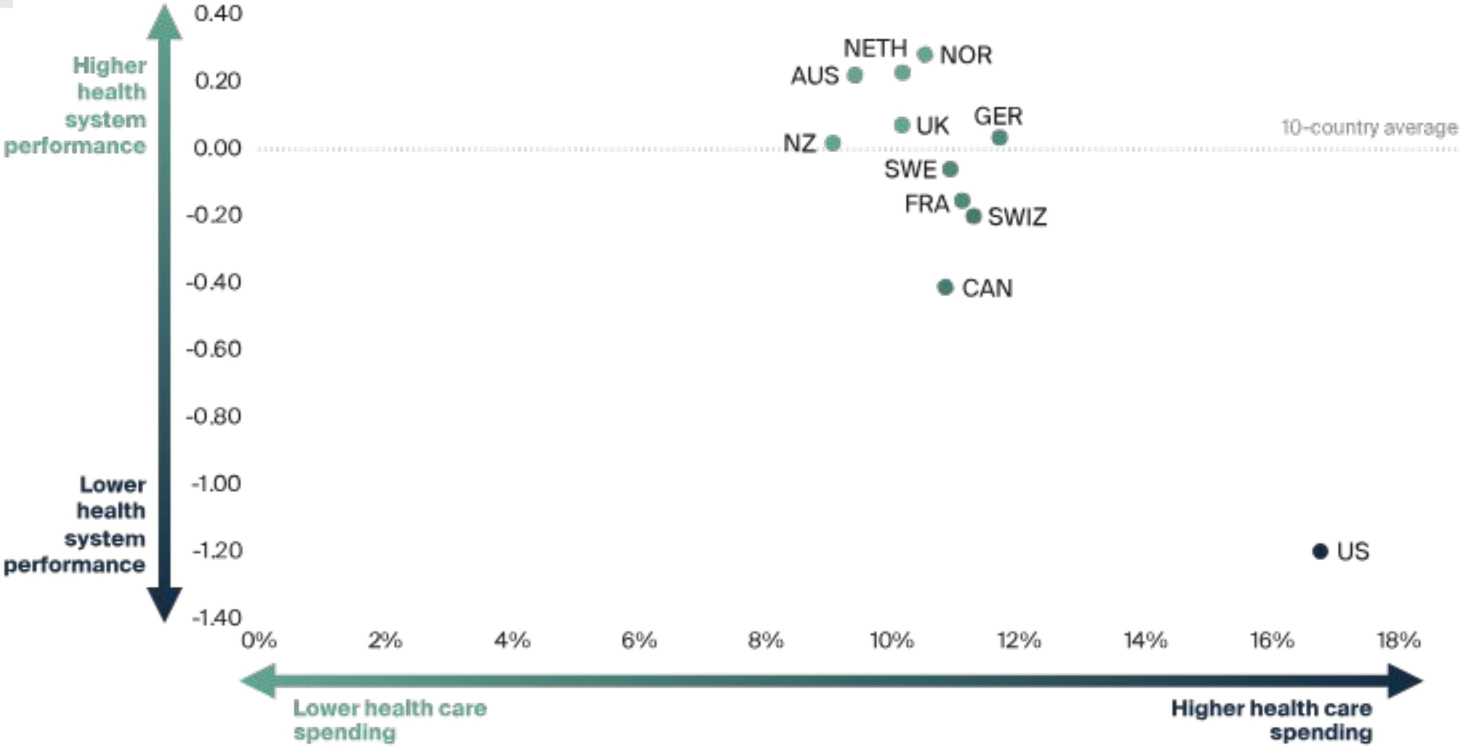
- 37.3 million people in the US have diabetes
- 96 million Americans have prediabetes
- 1 in 4 adults with diabetes don’t know they have diabetes
- Can lead to other health issues such as heart disease, kidney disease, nerve damage etc.

# Healthcare Spending in America




Source

# Spending vs. Performance




Source

# Benefits of ML in Healthcare

- Predictive tools allows for early intervention
- Decrease in healthcare costs

# National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age
- Outcome (target variable)

| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 3 | 78 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | 1 |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |

| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |

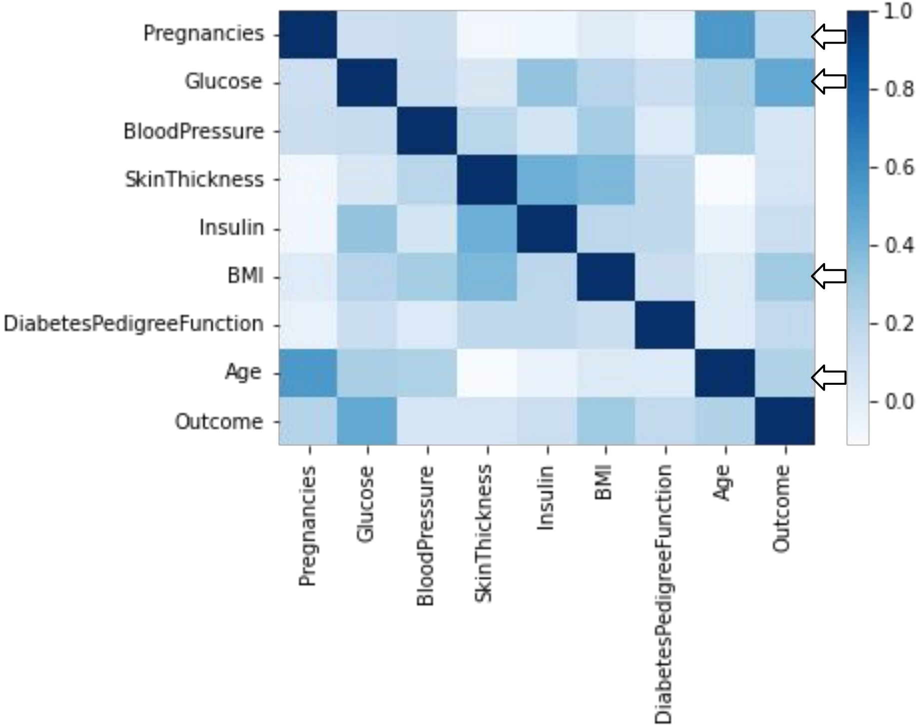
# Statistics - Diabetes vs. No Diabetes

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 268.000000 | 268.000000 | 268.000000 | 268.000000 | 268.000000 | 268.000000 | 268.000000 | 268.000000 | 268.0 |
| **mean** | 4.865672 | 141.257463 | 70.824627 | 22.164179 | 100.335821 | 35.142537 | 0.550500 | 37.067164 | 1.0 |
| **std** | 3.741239 | 31.939622 | 21.491812 | 17.679711 | 138.689125 | 7.262967 | 0.372354 | 10.968254 | 0.0 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.088000 | 21.000000 | 1.0 |
| **25%** | 1.750000 | 119.000000 | 66.000000 | 0.000000 | 0.000000 | 30.800000 | 0.262500 | 28.000000 | 1.0 |
| **50%** | 4.000000 | 140.000000 | 74.000000 | 27.000000 | 0.000000 | 34.250000 | 0.449000 | 36.000000 | 1.0 |
| **75%** | 8.000000 | 167.000000 | 82.000000 | 36.000000 | 167.250000 | 38.775000 | 0.728000 | 44.000000 | 1.0 |
| **max** | 17.000000 | 199.000000 | 114.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 70.000000 | 1.0 |

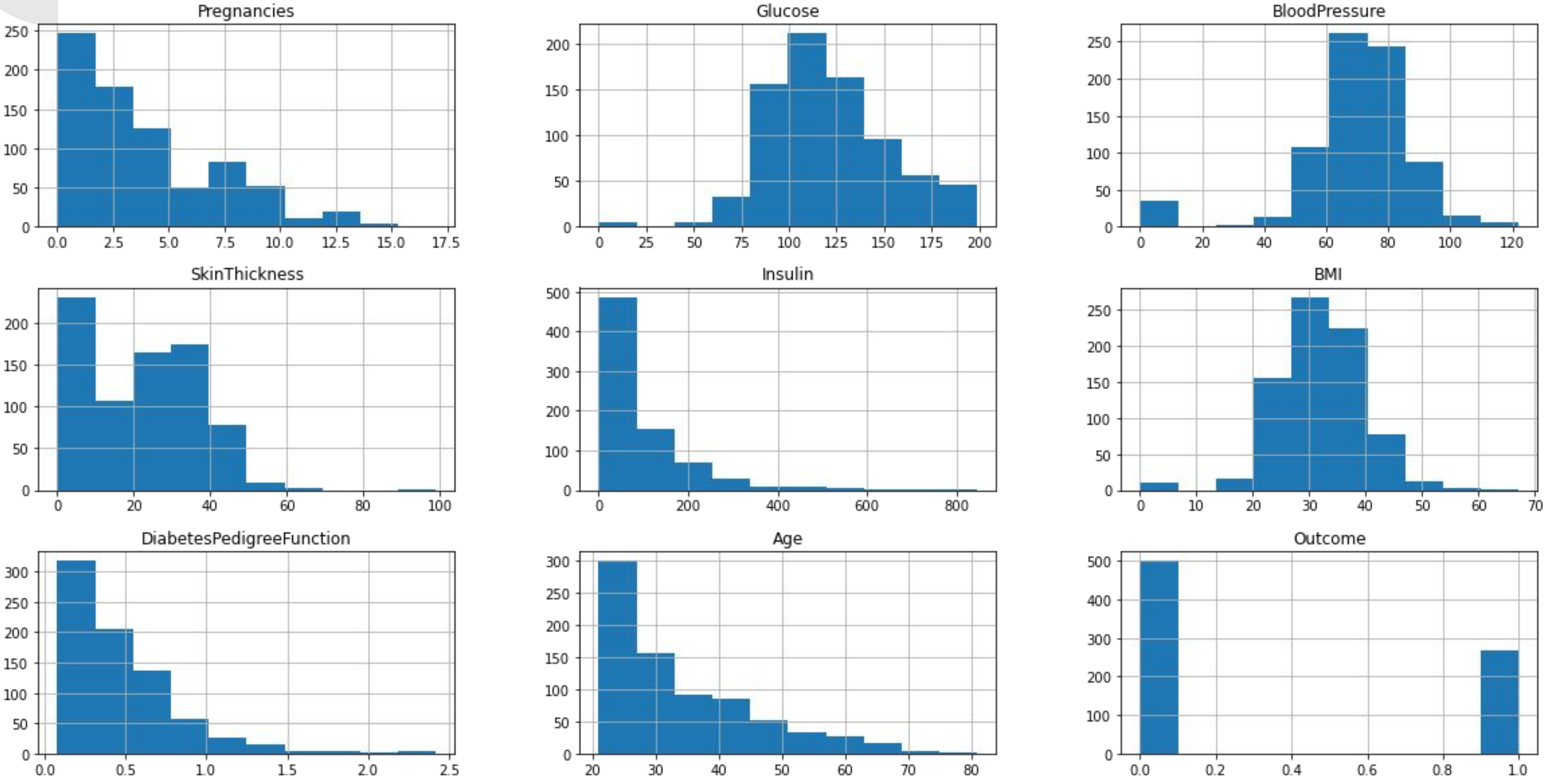
| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 500.000000 | 500.0000 | 500.000000 | 500.000000 | 500.000000 | 500.000000 | 500.000000 | 500.000000 | 500.0 |
| **mean** | 3.298000 | 109.9800 | 68.184000 | 19.664000 | 68.792000 | 30.304200 | 0.429734 | 31.190000 | 0.0 |
| **std** | 3.017185 | 26.1412 | 18.063075 | 14.889947 | 98.865289 | 7.689855 | 0.299085 | 11.667655 | 0.0 |
| **min** | 0.000000 | 0.0000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.0 |
| **25%** | 1.000000 | 93.0000 | 62.000000 | 0.000000 | 0.000000 | 25.400000 | 0.229750 | 23.000000 | 0.0 |
| **50%** | 2.000000 | 107.0000 | 70.000000 | 21.000000 | 39.000000 | 30.050000 | 0.336000 | 27.000000 | 0.0 |
| **75%** | 5.000000 | 125.0000 | 78.000000 | 31.000000 | 105.000000 | 35.300000 | 0.561750 | 37.000000 | 0.0 |
| **max** | 13.000000 | 197.0000 | 122.000000 | 60.000000 | 744.000000 | 57.300000 | 2.329000 | 81.000000 | 0.0 |

# Correlation to Occurence of Diabetes

| | 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 | $\boxed{0.221898}$ |
| **Glucose** | 0.129459 | 1.000000 | 0.152590 | 0.057328 | 0.331357 | 0.221071 | 0.137337 | 0.263514 | $\boxed{0.466581}$ |
| **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 | $\boxed{0.292695}$ |
| **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 | $\boxed{0.238356}$ |
| **Outcome** | 0.221898 | 0.466581 | 0.065068 | 0.074752 | 0.130548 | 0.292695 | 0.173844 | 0.238356 | 1.000000 |


Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0

# Distribution of all Variables



# Missing Values in the Data

- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI

# Handling Missing Data

- Three types of missing data
  - Missing completely at random (MCAR) - mean, median, mode
  - Missing at random (MAR) - multiple imputation, regression imputation
  - Missing not at random (MNAR) - pattern substitution, maximum likelihood estimation

# Multiple Imputation with Mice Forest Algorithm

- Data missing at random (MAR)
- Training a model over multiple iterations to predict the missing values

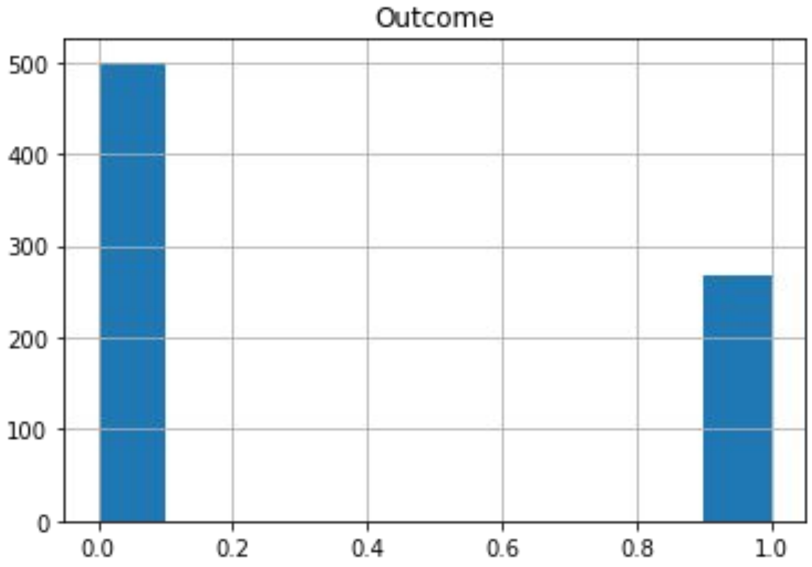
| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | None | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85 | 66 | 29 | None | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183 | 64 | None | None | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148.0 | 72.0 | 35.0 | 130.0 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85.0 | 66.0 | 29.0 | 37.0 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183.0 | 64.0 | 21.0 | 120.0 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89.0 | 66.0 | 23.0 | 94.0 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137.0 | 40.0 | 35.0 | 168.0 | 43.1 | 2.288 | 33 | 1 |

| | |
|---|---|
| Pregnancies | 0 |
| Glucose | 5 |
| BloodPressure | 35 |
| SkinThickness | 227 |
| Insulin | 374 |
| BMI | 11 |
| DiabetesPedigreeFunction | 0 |
| Age | 0 |
| Outcome | 0 |

| | |
|---|---|
| Pregnancies | 0 |
| Glucose | 0 |
| BloodPressure | 0 |
| SkinThickness | 0 |
| Insulin | 0 |
| BMI | 0 |
| DiabetesPedigreeFunction | 0 |
| Age | 0 |
| Outcome | 0 |

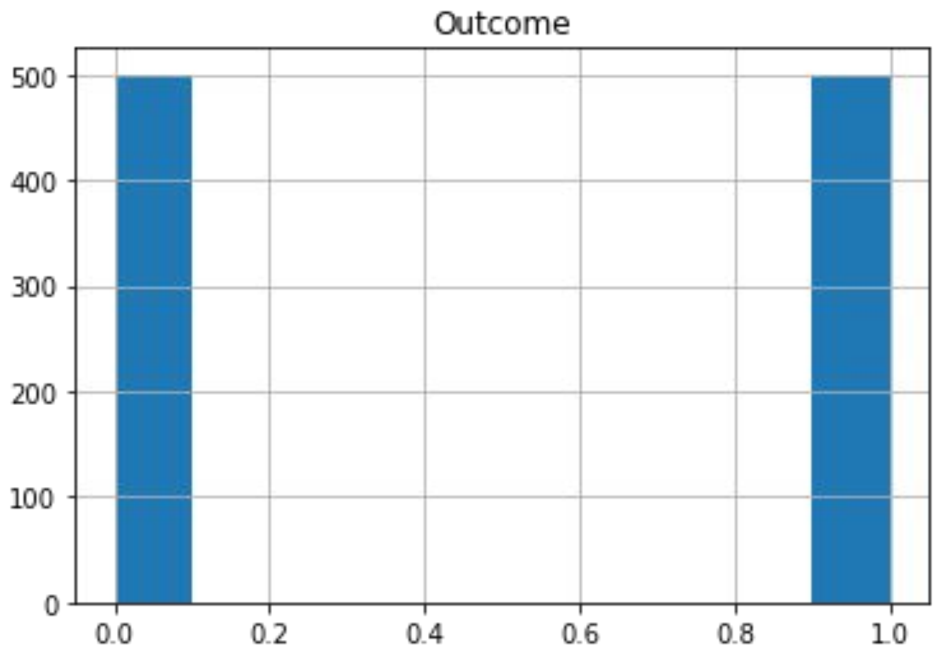
# Distribution of the Outcome of Diabetes



# Synthetic Minority Oversampling Technique (SMOTE)

- Undersampling vs. Oversampling

- Generates new instances for the minority class


Outcome
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0

# Decision Tree Model

- Supervised machine learning model
- Used for classification
  - Occurrence of diabetes vs. no occurrence of diabetes

# Random Forest Model

- Supervised machine learning model
- Used for classification
- Combines multiple decision trees to make predictions
- More accurate than decision trees

# Accuracy of Models

- Decision Tree: 79%
- Random Forest: 82%

# Evaluation of Machine Learning Models

- Confusion matrix

| | |
|---|---|
| Count of True Positives | Count of False Positives |
| Count of False Negatives | Count of True Negatives |

# Evaluation of Decision Tree Model

From the test set of the machine learning model, 98 of those points were correctly predicted to have diabetes

```
array([[ 98,  27],
       [ 24, 101]])
```

From the test set of the machine learning model, 101 of those points were correctly predicted to not have diabetes

# Evaluation of Random Forest Model

From the test set of the machine learning model, 101 of those points were correctly predicted to have diabetes

```
array([[101,   24],
       [ 19,  106]])
```

False positives is less than the decision tree

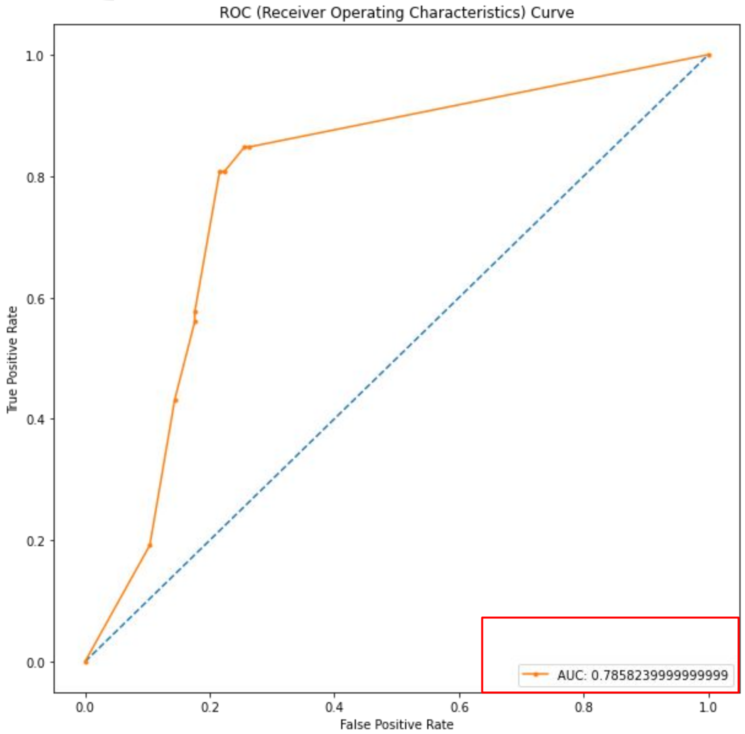
From the test set of the machine learning model, 106 of those points were correctly predicted to not have diabetes

# Evaluation of Machine Learning Models

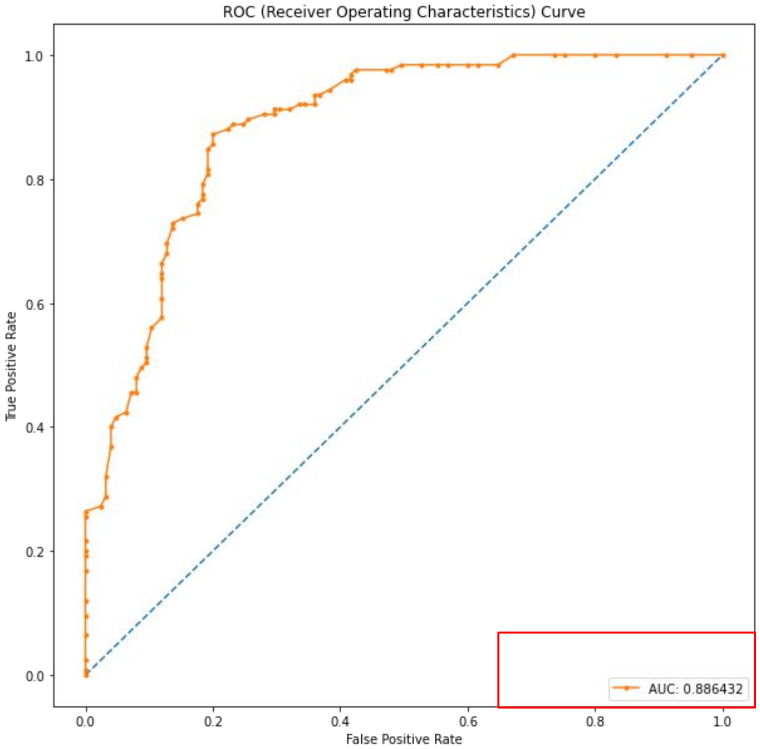
- AUC Curve - area under the curve
    - True positive rate vs. false positive rate

- Ranges 0-1

# Decision Tree




# Random Forest



# Conclusions

- Random forest model performed better
- Using a confusion matrix and AUC curve to evaluate the performance of both models
- Great potential for use in healthcare

# References


Chugh, V. (2022, October 4). Which Metric Should I Use? Accuracy vs. AUC. *KD Nuggets*. https://www.kdnuggets.com/2022/10/metric-accuracy-auc.html

Healthcare Costs For Americans Projected to Grow at an Alarmingly High Rate. (2019, May 1). *Peter G. Peterson Foundation*. https://www.pgpf.org/blog/2019/05/healthcare-costs-for-americans-projected-to-grow-at-an-alarmingly-high-rate

Prabhakaran, S. (n.d.). MICE imputation – How to predict missing values using machine learning in Python. *Machine Learning +*. https://www.machinelearningplus.com/machine-learning/mice-imputation/

Risk Factors for Type 2 Diabetes (2022, July). *National Institute of Diabetes and Digestive and Kidney Diseases*. https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

R, S. E. (2023, July 5). Understand Random Forest Algorithms With Examples. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Saini, A. (2023, September 13). Decision Tree Algorithm – A Complete Guide. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

Satpathy, S. (2023, July 24). SMOTE for Imbalanced Classification with Python. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/#h-smote-synthetic-minority-oversampling-technique

Schneider, E. C., Shah, A., Doty, M. M., Tikkanen, R., Fields, K., Williams II, R. D. (2021, August 4). Health Care in the U.S. Compared to Other High-Income Countries. *The Commonwealth Fund*. https://www.commonwealthfund.org/publications/fund-reports/2021/aug/mirror-mirror-2021-reflecting-poorly?utm_source=google&utm_medium=cpc&utm_campaign=Mirror_Mirror_-_Universal_Coverage&utm_adgroup=Mi

Top Techniques to Handle Missing Values Every Data Scientist Should Know. (2023, January). *Datacamp*. https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values