Prediction Of Diabetes Using Data Mining Techniques

CS6220 Data Mining Techniques - Final Project

Team Members: Paul Cruz, Chih-Ming Sun, Haocheng Yang

Introduction

Diabetes is a well-known disease affecting individuals all over the world. According to the WHO (World Health Association), an estimated **422** million people worldwide currently have diabetes with the majority living in low-and middle-income countries. In addition, **1.6** million deaths are directly attributed to diabetes each year.

Can diabetes be prevented?

Introduction

Though there are many factors that may contribute to a diabetes diagnosis, such as heredity, weight, and height, the major factor is blood sugar concentration. The early identification of abnormal glucose levels may be the only remedy to further complications and therefore a more serious diagnosis. Thus, analysis of the levels of sugar in the blood and how it may compare to other compounds measured is paramount in predictive modeling.

Methodology - Dataset

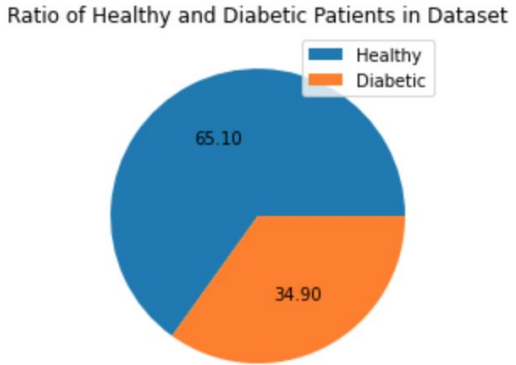
The dataset used for this study was provided by Kaggle, a data repository website. This dataset was provided to the repository courtesy of the University of California, Irvine (UCI), and included 70 sets of data recorded on diabetes patients ranging from several weeks to several months' worth of glucose, insulin, and lifestyle data per patient, as well as a description of the problem domain.

The dataset included the following tables: *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome.*

Methodology - Data Processing

Some basic analysis of the dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	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



Methodology - Data Processing

The data was first viewed to allow for proper analysis of its standard distribution. The data was then pre-processed to correct data inconsistencies, missing values, and to remove null values.

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

Original data

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

After process null value

Methodology - Data Analysis

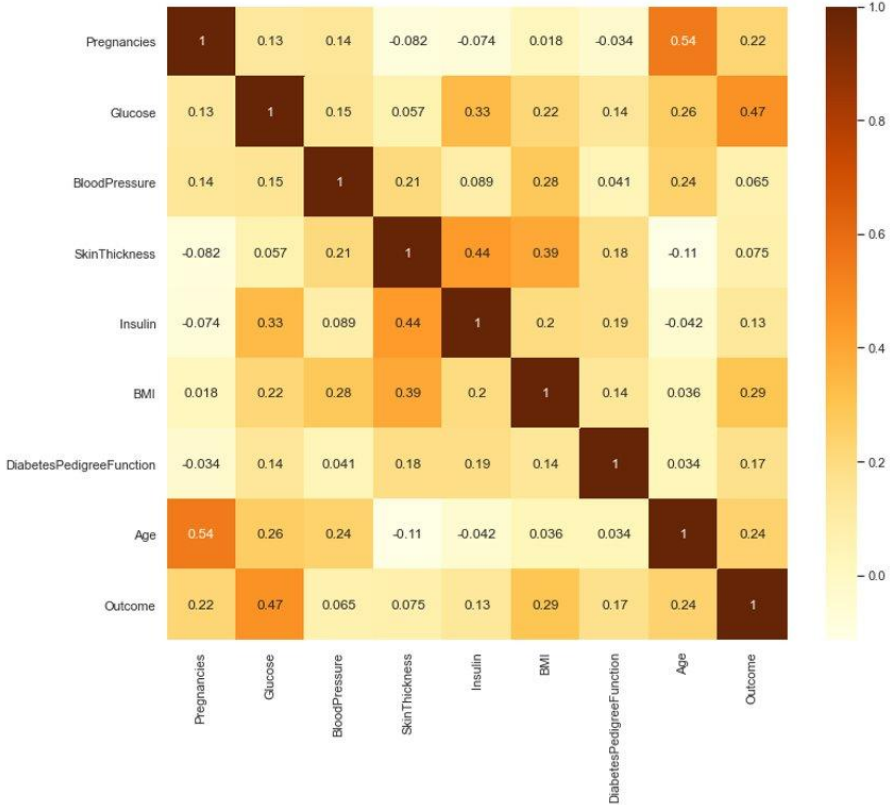
1. We first created a heatmap to understand the correlation coefficient of the features of the data.
2. Next, we chose to plot a histogram to analyze the association between outcome and diagnosis.
3. We then chose to evaluate the following models by splitting the data into **70%** training, and **30%** test: *Decision Tree, Support Vector Machine, K Neighbors Classifier, Gradient Boosting Classifier, Random Forest Classifier, Ada Boost, Gaussian NB,* and *Linear Regression.*
4. **Three models** were chosen to move forward with the analysis, in which a *confusion matrix* and *ROC curve* was plotted and analyzed.
5. After analyzing feature importance and accuracy scores, two features were found to have significance.

Result - Heatmap

The heatmap helps us to understand the correlation coefficient of the features of the dataset.

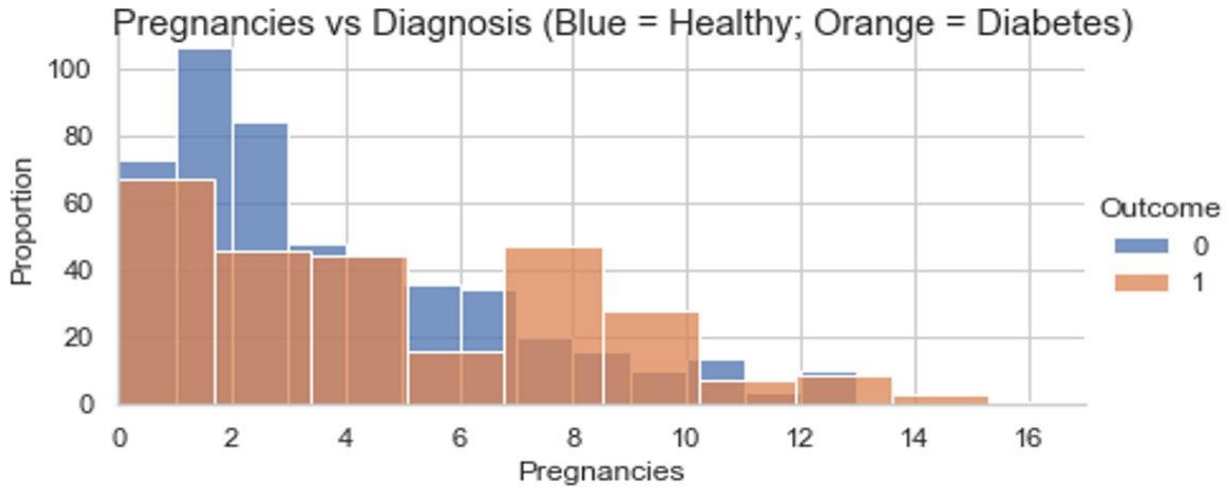
Intuitively form the heatmap, the darker the color is, the more correlated the two features are.

From the heatmap, we can clearly see that **Glucose**, **BMI**, **Age** and **Pregnancies** are four of the most correlated feature to lead to Diabetes.



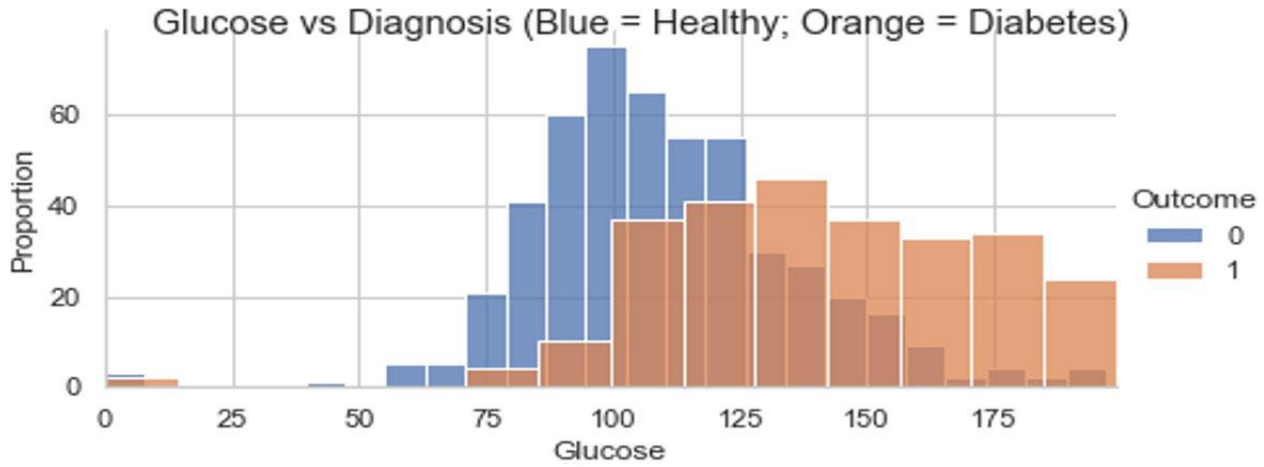
Result - Histogram (Pregnancies)

Pregnancies vs diagnosis showed that an around half of Pregnancies resulted in an outcome of a women getting diabetes.



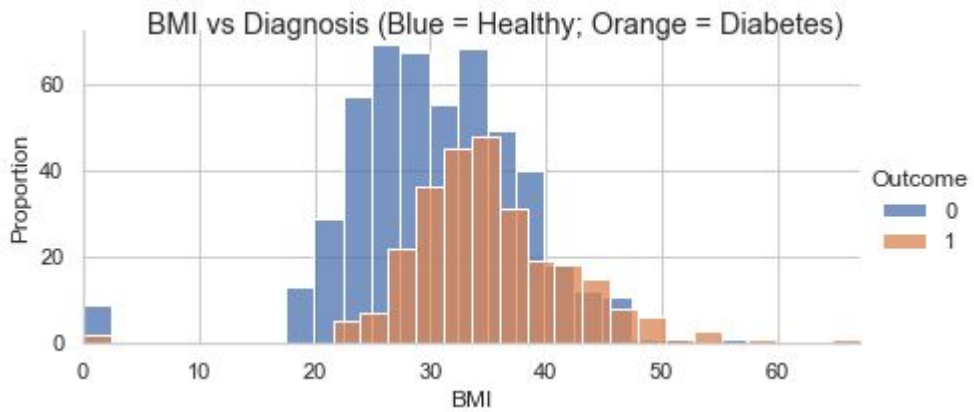
Result - Histogram (Glucose)

Glucose vs diagnosis showed that glucose levels over 76 and mainly above 125 was correlated with diabetes diagnosis.



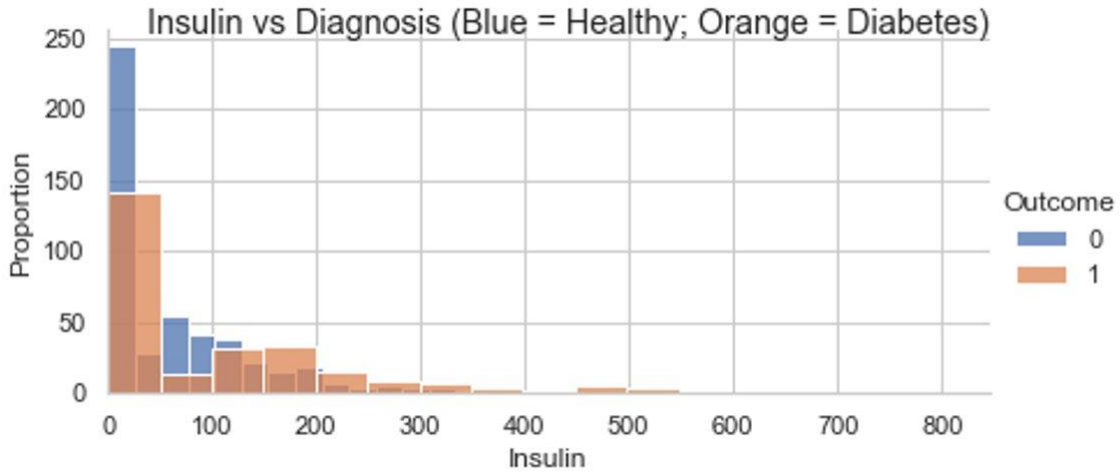
Result - Histogram (BMI)

BMI vs diagnosis showed that both healthy and diabetic individuals have two similar bell shape histograms with an obvious shift of mean.



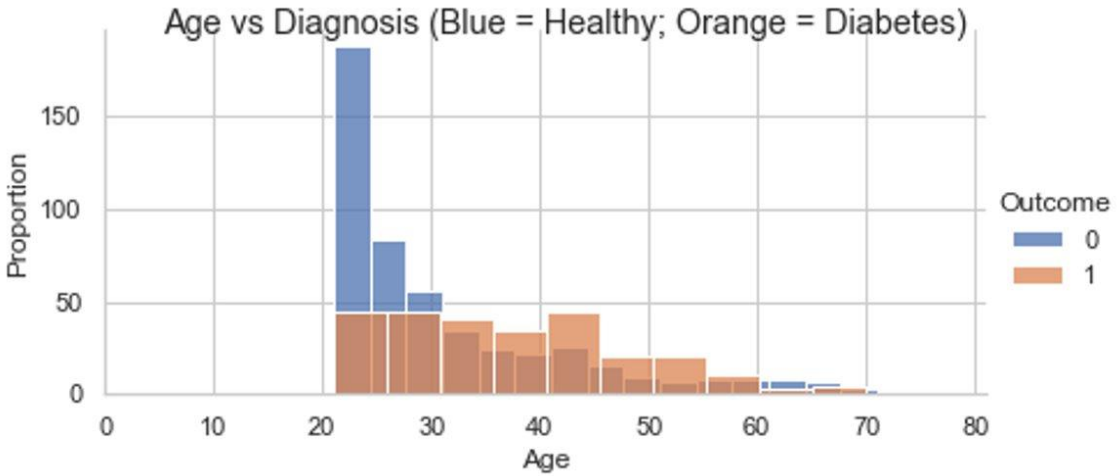
Result - Histogram (Insulin)

Insulin vs diagnosis found the same, that health and diabetic patients exhibited similar insulin levels.



Result - Histogram (Age)

Age vs diagnosis showed that diabetes in this dataset is found between the ages of 22-70 years old, and people around 40-50 has a high potential to get diabetes.



Result - Modelling

Initially, we build **7** classifying models, and analyzed their prediction accuracy.

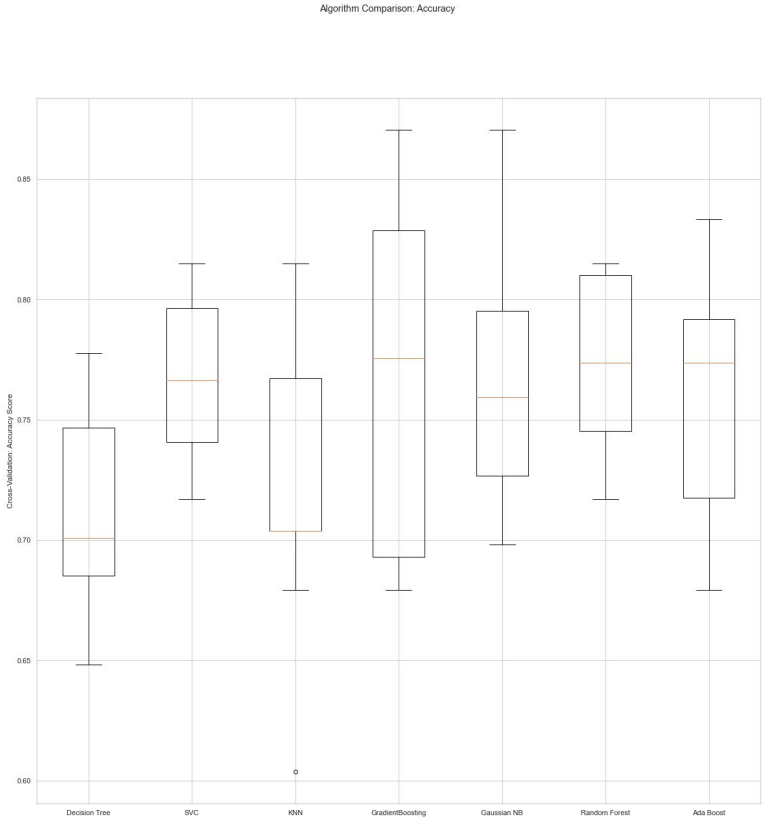
They are:

1. Decision Tree
2. Support Vector Machine
3. K Neighbors Classifier
4. Gradient Boosting Classifier
5. Random Forest Classifier
6. Ada Boost
7. Gaussian NB

```
Decision Tree: 0.709469 (0.041069)
SVC: 0.767051 (0.031590)
KNN: 0.722257 (0.058806)
GradientBoosting: 0.766981 (0.067470)
Gaussian NB: 0.768903 (0.052252)
Random Forest: 0.772711 (0.035685)
Ada Boost: 0.759679 (0.046430)
```

Result - Modelling

Then we compared the accuracy of each model by drawing the boxplot.



Result - Modelling

Again, We changed the iteration time and used cross validation score to see the accuracy of each classifier.

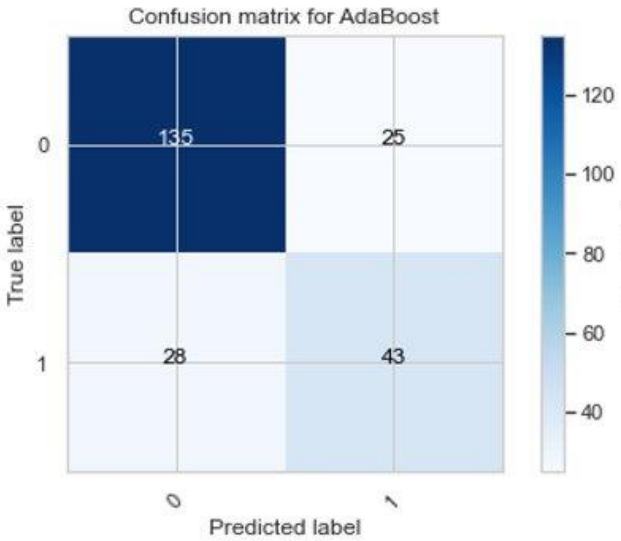
After comparing and analyzing, we chose 3 models to explore their capability of predicting diabetes.

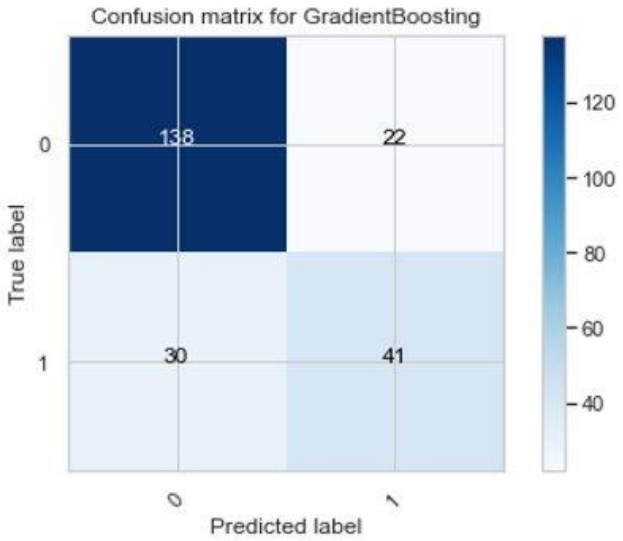
They are: **RandomForest**, **GradientBoosting**, and **Ada Boost**

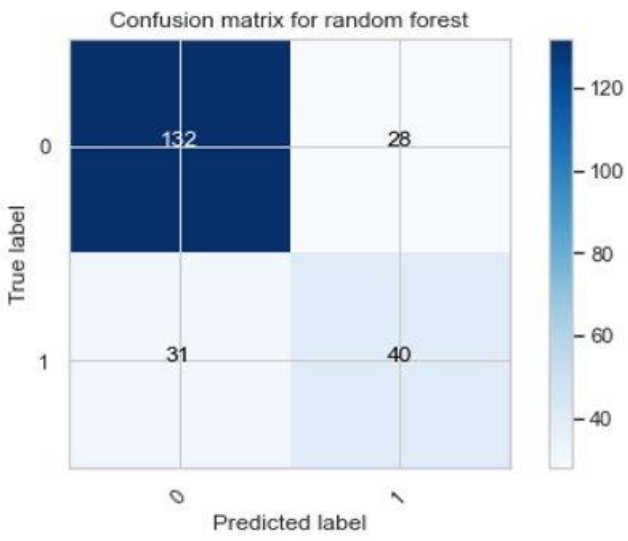
```
Decision Tree = DecisionTreeClassifier
SVC = Support Vector Machine SVC
KNN = KNeighborsClassifier
GradientBoosting = GradientBoostingClassifier
Gaussian NB = GaussianNB
Random Forest = RandomForestClassifier
Ada Boost = AdaBoostClassifier
GradientBoosting = GradientBoostingClassifier
```

```
       Decision Tree | Accuracy: 71.51% (+/- 1.94%)
                 SVC | Accuracy: 75.23% (+/- 3.42%)
                 KNN | Accuracy: 72.80% (+/- 7.01%)
    GradientBoosting | Accuracy: 76.90% (+/- 6.24%)
         Gaussian NB | Accuracy: 76.91% (+/- 2.53%)
       Random Forest | Accuracy: 78.57% (+/- 6.04%)
           Ada Boost | Accuracy: 73.91% (+/- 8.92%)
```

Result - Confusion Matrix

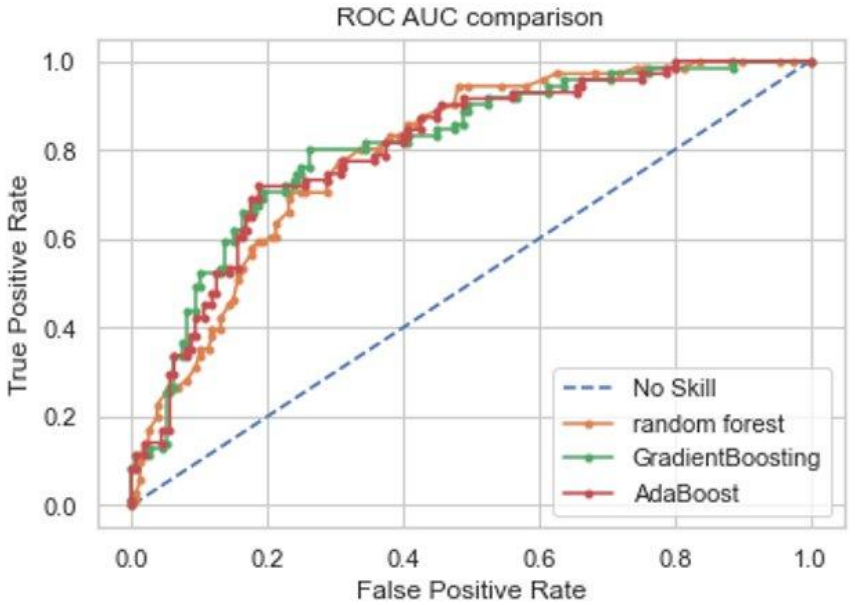






Result - ROC

ROC AUC comparison was plotted with **Baseline**, **Random Forest, Gradient Boosting**, and **Ada Boost.** The result of the plot shows that all three models are very good prediction model, while the Gradient Boosting is the best predictive model.



Result - Accuracy & Feature Importance

Again, we calculated the accuracy of the three models as well as the importance of each features, and the result is below.

Accuracy of random forest: 0.76
DecisionTreeClassifier - Feature Importance:

	Variable	absCoefficient
1	Glucose	0.287808
5	BMI	0.142343
7	Age	0.121921
6	DiabetesPedigreeFunction	0.114688
0	Pregnancies	0.099902
2	BloodPressure	0.084220
4	Insulin	0.082514
3	SkinThickness	0.066602
8	Outcome	NaN

Accuracy of GradientBoosting: 0.77
DecisionTreeClassifier - Feature Importance:

	Variable	absCoefficient
1	Glucose	0.440522
5	BMI	0.152091
7	Age	0.109702
6	DiabetesPedigreeFunction	0.096314
4	Insulin	0.073429
0	Pregnancies	0.067637
2	BloodPressure	0.045053
3	SkinThickness	0.015252
8	Outcome	NaN

Accuracy of AdaBoost: 0.77
DecisionTreeClassifier - Feature Importance:

	Variable	absCoefficient
1	Glucose	0.22
5	BMI	0.18
2	BloodPressure	0.16
4	Insulin	0.14
6	DiabetesPedigreeFunction	0.14
7	Age	0.08
0	Pregnancies	0.06
3	SkinThickness	0.02
8	Outcome	NaN

Result - Reduce features

Two features were found to have the most importance; **Glucose** and **BMI**.

We dropped columns and only kept the two columns Glucose and BMI. And we repeated the steps to use Gradient Boosting Classifier, and the final accuracy is **75%**, with *Glucose absolute coefficient* being **0.62**, and *BMI* **0.37**.

```
Accuracy of GradientBoostingClassifier in Reduced Feature Space: 0.75

 GradientBoostingClassifier - Feature Importance:

   Variable  absCoefficient
0  Glucose        0.622342
1      BMI        0.377658
2  Outcome             NaN
```

Discussion

1. The model evaluated in this study shows how data mining techniques via machine learning predictive modelling can be used to predict human disease.
2. Our results show how machine learning and classification algorithms can be used to predict diabetes.
3. The two features that were found to be most important were BMI and glucose.

Discussion

1. For BMI, there is a certain relationship between the satisfaction rate of blood glucose control and overweight or obesity, which explains the importance of BMI in the classification of control satisfaction.

2. This tells us that lifestyle choices can make a difference between whether you develop diabetes or not.

Discussion

- The results show that there are many predictive models that can be applied, but only certain models will work best which is very dependent on the selected dataset.
- Therefore, when focusing on a certain disease, several appropriate classification algorithms should be selected based on the characteristics of the dataset.
- By comparing the classification accuracy of many classification algorithms on the dataset, the most effective classification algorithm can be selected and used as the diagnostic model.

Discussion

- In general, the performance of machine learning algorithms is evaluated using predictive accuracy.
- In addition, despite the claims that machine learning classification algorithms can generate sufficient and effective decision-making, very few have really permeated the clinical practice.
- Therefore, the practice of using machine learning algorithms to predict disease should be under further study in the biomedical research and development field, and we hope that this study provided a good example of how this can be done.

Discussion

- One of the most important real-world medical problems is the detection of diabetes at its early stage.
- In this study, systematic efforts are made in finding a system which results in the prediction of diabetes through data science via predictive modelling algorithms.
- Though this work may not be the final solution to the prediction of diabetes, it serves as an example of the power data science may have on the prediction of other diseases as well.

Discussion

- During this work, three machine learning classification algorithms are chosen and evaluated on various measures.
- The results determine the adequacy of the system with an achieved accuracy of 79% using the Random Forest classification algorithm.
- In the future, this system with the use of machine learning classification algorithms may be able to be used to predict or diagnose other diseases as well. The work therefor be extended and improved for the automation of diabetes analysis using the methods described here as well as other machine learning algorithms.

References

- Huang, R. (2021). Prediction of Pima Indians Diabetes with Machine Learning Algorithms. UCLA. ProQuest ID: Huang_ucla_0031N_19508. Merritt ID: ark:/13030/m5md4q60. Retrieved from https://escholarship.org/uc/item/6rh07945
- Altıntaş, Ergin. “UCI Diabetes Data Set.” Kaggle, 1 May 2020, www.kaggle.com/ealtintas/uci-machine-learning-repository-diabetes-data-set.
- World Health Organization (WHO). “Diabetes.” Kaggle, 2 April 2021, https://www.who.int/health-topics/diabetes#tab=tab_1.

Contribution

Paul Cruz:

Chih-Ming Sun:

Haocheng Yang:

Thank you!