Predicting if Patient has Diabetes

Dipika Sharma

Bellevue University

Applied Data Science 680

Amirfarrokh Iranitalab

Project White Paper: Predict if Patient has Diabetes

Introduction:

Diabetes is a metabolic disease which can results in serious health issue if proper precaution not taken. It causes high blood sugar which can affect the hormone insulin of patient body. Hormone insulin helps body in moving the sugar from the blood to body cells and make sure it stored as energy. As part of the first project, I chose the dataset of medical science field. Using this analysis, we will predict if patient has diabetes or not.

Business Problem:

Diabetes is chronical disease, and it is important to identify if patient has the diabetes as soon as possible. Early identification of diabetes will give patient time to manage diet and use treatment to control it and reduce the risk of developing serious health problem.

I will be using the predictor variables like a patient's BMI, pregnancy details, insulin level, age, etc. to see if patient has diabetes or not.

Data Explanation:

I will be using the dataset from Kaggle website, this dataset is from National institutes of Diabetes, Digestive and kidney diseases.

The data can find at below location:

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

This dataset is good for our purpose as it consists of medical predictor variable like patient's BMI, pregnancy details, insulin level, age, etc. and one target variable, outcome.

This database has 768 rows and 9 columns.

	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

The detail of the columns is below.

- Pregnancies : It is a numeric field to show number of times patient get pregnant
- Glucose : Plasma glucose concentration 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) of the patient
- Outcome : Class variable (0 or 1) 268 of 768 are 1, the others are 0

Project Methodology:

As mentioned, the objective of the project is to predict if patient has diabetes based on previously observed values. A classic approach to Diabetes prediction includes medical analysis by expertise in medical field, but a more modern approach includes machine learning and sentiment

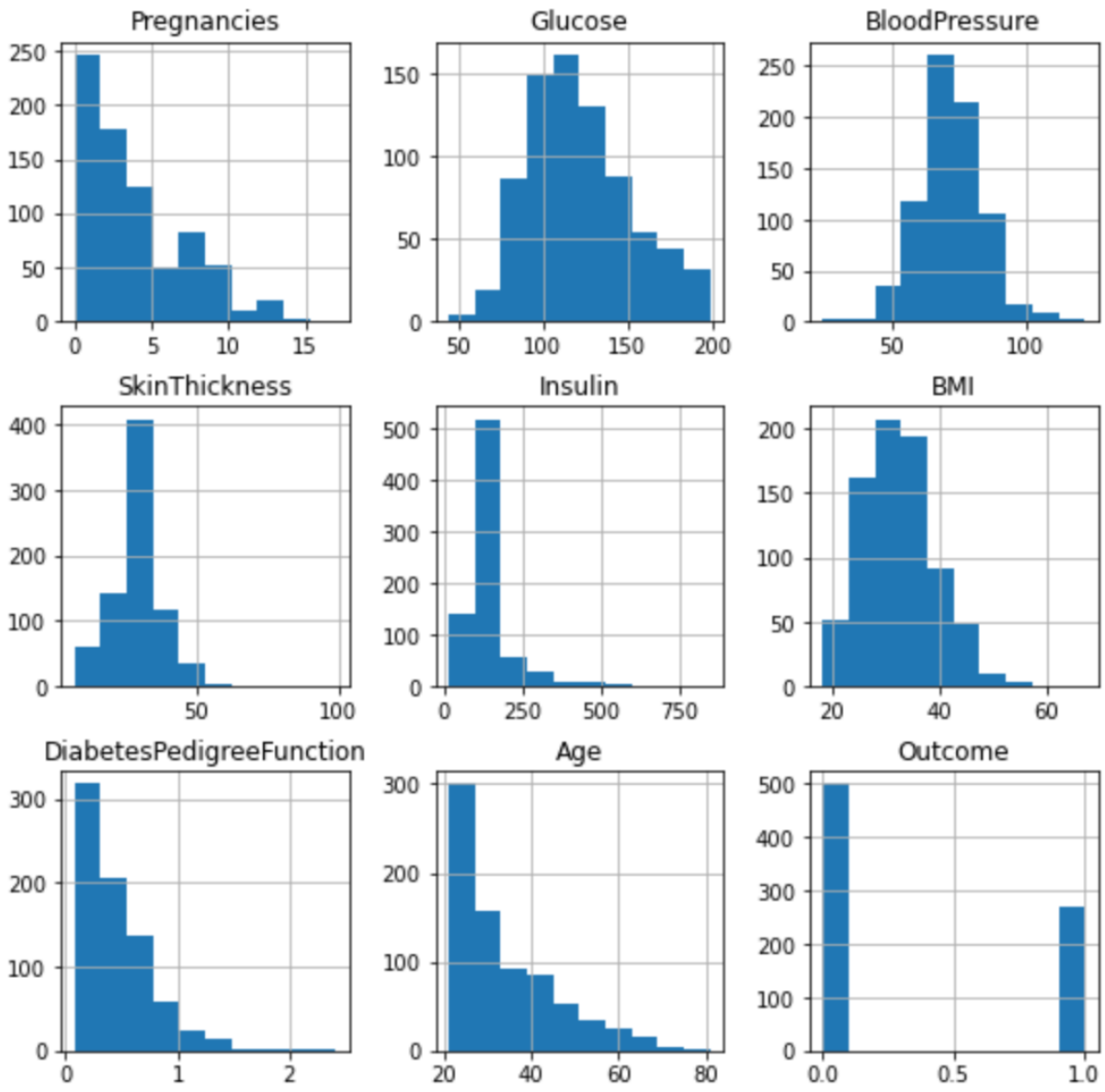
analysis. The model capable of this modern approach are supervised learning algorithm like decision tree and random forest algorithms. Both algorithms considered to be most widely used model when it comes to draw conclusions about a set of observation. This is appropriate models to deal with medical data and most widely known for efficiency, reliability, and capability when we need to perform predictive modeling.

As part of this project both algorithms were performed, and Random Forest model was selected over decision tree algorithm because of its ease of application and interpretability.

Analysis:

I started the analysis with cleaning process, where I learned that some of the columns have missing values which I planned to replace with the mean or median of the respective column. The missing values can impact the result accuracy and reliability hence it is very important to handle them carefully in preprocessing step.

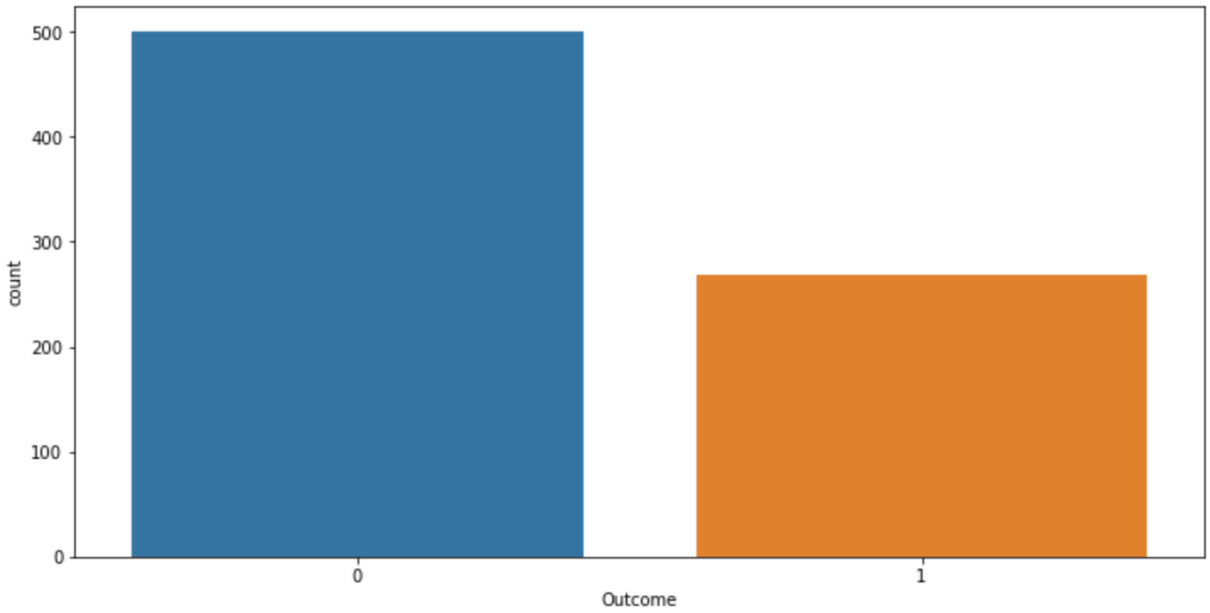
Let’s visualize the data in the histogram to see the changes after imputing missing values.



As we can see there is no gaps between the range of data in histogram and data looks pretty clean now.

The next visualization I done is for outcome to understand if data is balanced or not.

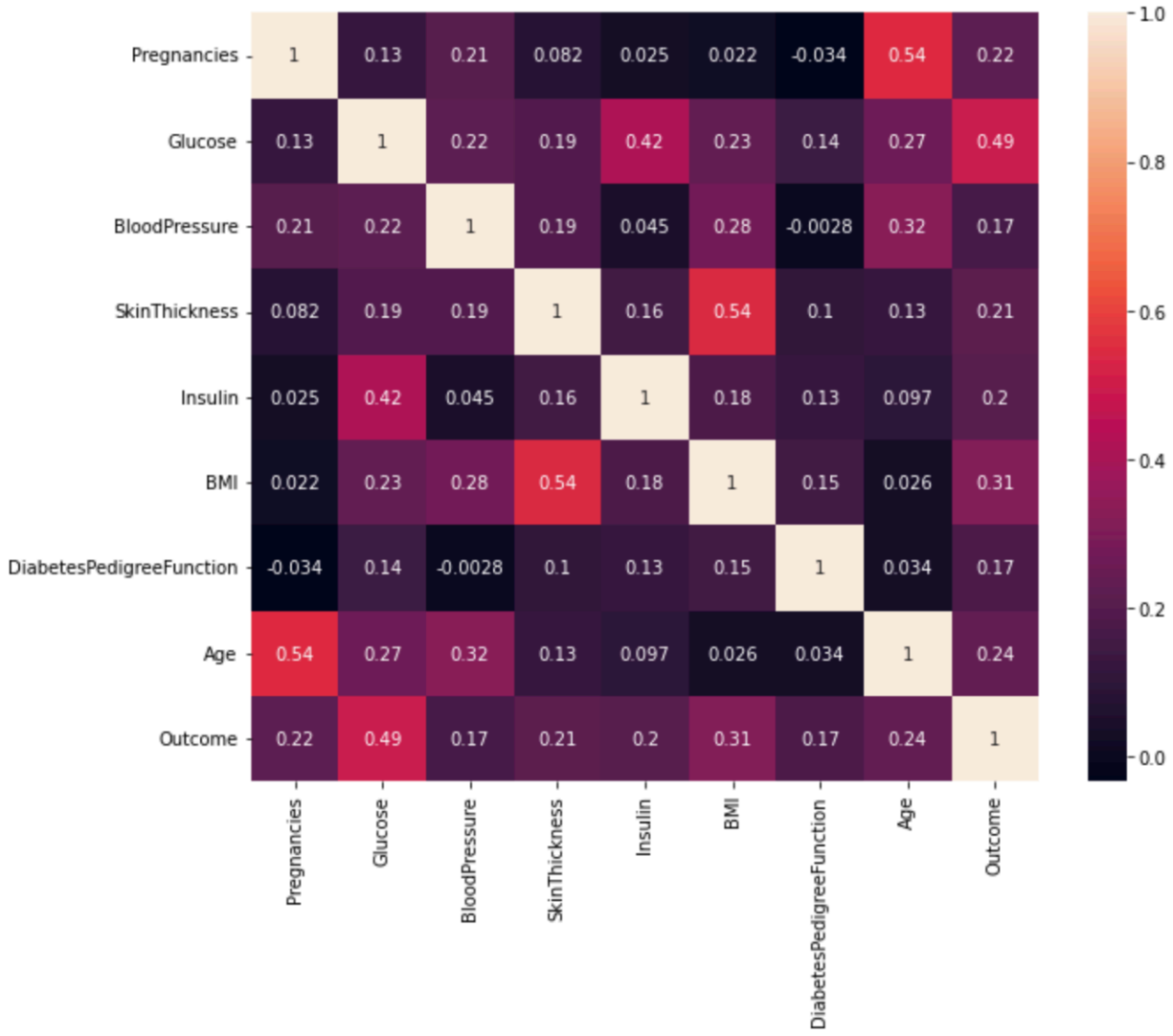
```
<AxesSubplot:xlabel='Outcome', ylabel='count'>
```



As we can clearly see that the observation outcome of the dataset is unevenly distributed as the patients with no diabetes is almost double the patient who has diabetes. It is important to handle the imbalanced data to avoid building biased model and inaccurate predictions.

Next step is to learn about the relationship between the dataset by using the correlation matrix

```
<AxesSubplot:>
```



Looking at the heating map above we can clearly see that outcome is strongly related to glucose. Also, it is related to features like BMI, Age, pregnancies, and skin thickness.

The pair plot also created to understand which feature has strong relationship and unique pattern. Please see **Appendix, Table 1: Pair Plot.**

The next important step before we start working on the model is scaling. Scaling is important as it avoid the biasness in model and improve the model accuracy. For our project we are using the

standard scaling this will help us to normalize the train dataset to make sure all the features contribute at same level to the model and the model result will not get dominate by the larger values of single feature.

The next step is to split the data into training and testing datasets. The training dataset will be used to prepare the model and generate a prediction. The testing dataset, which is much smaller than the training dataset, will be the part of the data that we will try to predict.

Conclusion:

The two-model decision tree and random forest were trained to performed prediction, but random forest show the accuracy of 77% and came out as best model in this project scenario.

The confusion matrix is created to see how the Random Forest model performing:

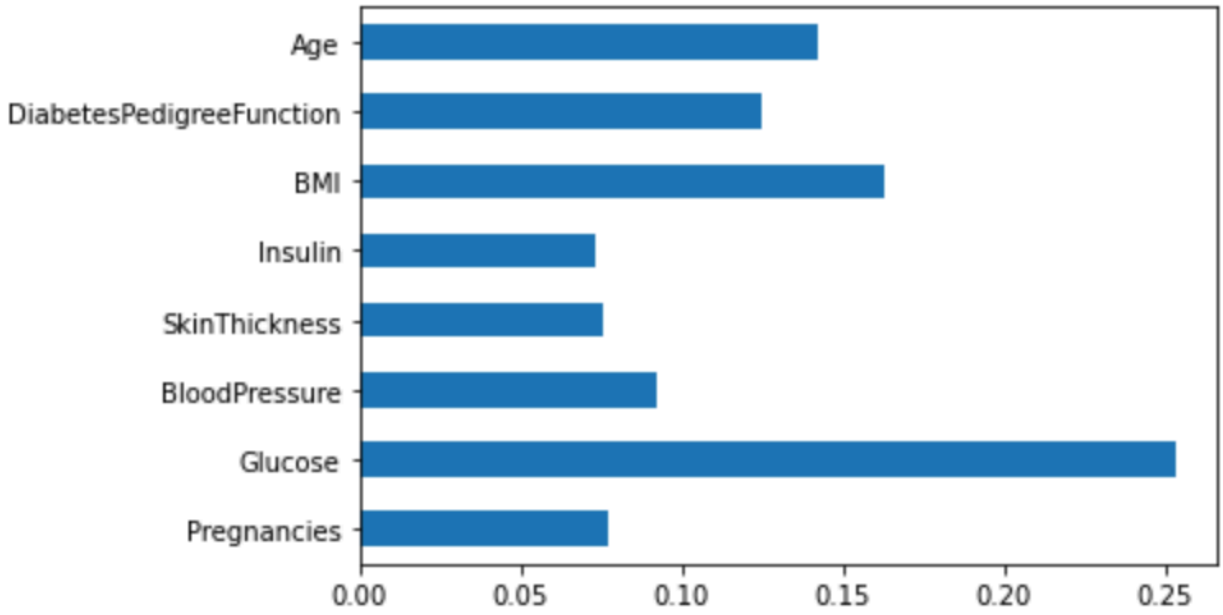
```
[[136  26]
 [ 32  60]]
              precision    recall  f1-score   support

           0       0.81      0.84      0.82       162
           1       0.70      0.65      0.67        92

    accuracy                           0.77       254
   macro avg       0.75      0.75      0.75       254
weighted avg       0.77      0.77      0.77       254
```

See Appendix, Table 2: Confusion matrix Decision Tree

It is important to detect the diabetes in patient early, delay in detection can cause complication to organs like heart and kidneys. With this project we are aiming to offer a machine learning tool which can predict diabetes with accuracy and precision. Using the Random Forest model will add value and help patient to take early precaution to be healthy.

We also analyzed the all the features to understand which features has more importance and can change the prediction. Used the feature importances from the Random Forest and visualize the data to have clear understanding. Importance of the features and as a result found out that the

```
<AxesSubplot:>
```



As we can see above the graph shows that the glucose is an important feature which has most of the influence in predicting if patient has diabetes or not.

Assumptions:

All the analysis has been performed using the dataset available online assuming that the data is representative of the larger environment from National institutes of Diabetes, Digestive and kidney diseases. The biggest assumption is that this dataset is collected without any bias since we are not sure how this data is collected and the original purpose of collecting the dataset. Any bias in dataset can lead us to bias model and to inaccurate prediction of outcome.

Limitations:

One of the limitations is we using the historical dataset, we are not sure how old is this dataset. Also, this dataset has only 8 features as predictors variable, if we are missing any important features that can changes the prediction of outcome then we build the model with certain limitation. In order to deal with this limitation, we need to try this model on different dataset and a proper review should happen on this model by medical expertise to understand if we are missing anything in this model.

Challenges and Risks:

With all the assumption and limitation of the dataset the biggest challenge is to make sure the model we build is accurate. Some of the potential risk we can see depending on how the missing data, misbalancing and scaling of the data is handled in the preprocessing steps as all these factors can lead us to inaccurate prediction and bias in model.

Future Uses/Additional Applications:

The model we build here using machine learning can help patients in detecting diabetes risk early. Which in turns help them to make plan for treatments, start monitoring the diet to manage diabetes, identification of health risk and work on preventative strategies.

In order to advance our machine learning model, we can analyze more patient data that is available to us ethically like medical history and other lab tests to identify patient who can develop diabetes even before symptoms start showing up. This early detection can help patient to take early precaution to reduce the risks.

Recommendations:

Ensemble method is considered to be the essential when it comes to prediction in machine learning. We can use ensemble method to improve accuracy in predicting the diabetes in patient. This method allows us to combine the outcome of multiple models effectively instead of relying of any single model status which results in producing the robust and reliable outcome in making prediction.

Implementation Plan:

The next step is to partner with heath organizations, share the importance of the machine learning model with them, get access to the patient data so the model can be tested more on real time patient data.

Once the testing is done, we can implement the model in hospital, urgent care and any other medical services where medical support is provided to patient and the predictor variable like glucose and other important information of patient can be obtained. Once the model has all the predictor variables, it can identify health risk and alert the providers to make preventive strategies to safe patient life or reduce risk. Regular checks can be done to make sure of the consistency of the model and flow of the data.

Ethical Consequences:

Make sure all the ethical analysis guidelines should be followed while preforming data analysis steps.

1. To ensure that legal and ethical ways are used to collect the data and make sure no personal information of patients get misused or disclosed without the consent.
2. Make sure that all the algorithm and models that are used in the project do not show any biases and discrimination towards any group.

3. There should be transparency in the data analysis methods so that everyone understands the analysis easily.
4. Will make sure the security of data so there will no unauthorized access to the data.
5. Finally, will ensure there will no negative impacts on population because of the project result.

References:

1. World health Organization: - What is Diabetes? https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Diabetes%20is%20a%20chronic%20disease,hormone%20that%20regulates%20blood%20glucose.
2. Data Source : - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

Appendix:

Table 1: Pair Plot

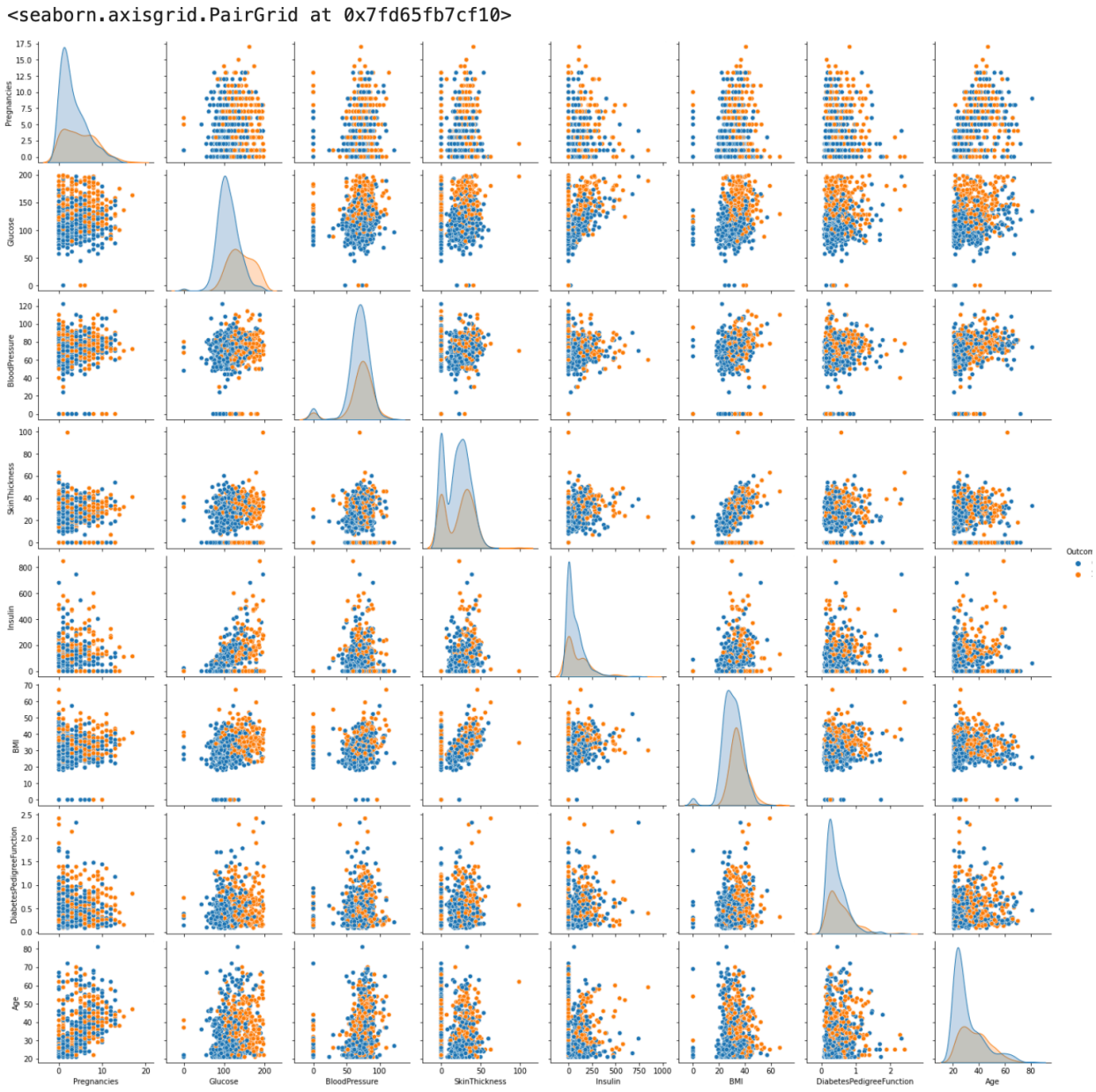


Table 2: Confusion Matrix Decision Tree

```
[[128  34]
 [ 40  52]]
              precision    recall  f1-score   support

           0       0.76      0.79      0.78       162
           1       0.60      0.57      0.58        92

    accuracy                           0.71       254
   macro avg       0.68      0.68      0.68       254
weighted avg       0.70      0.71      0.71       254
```