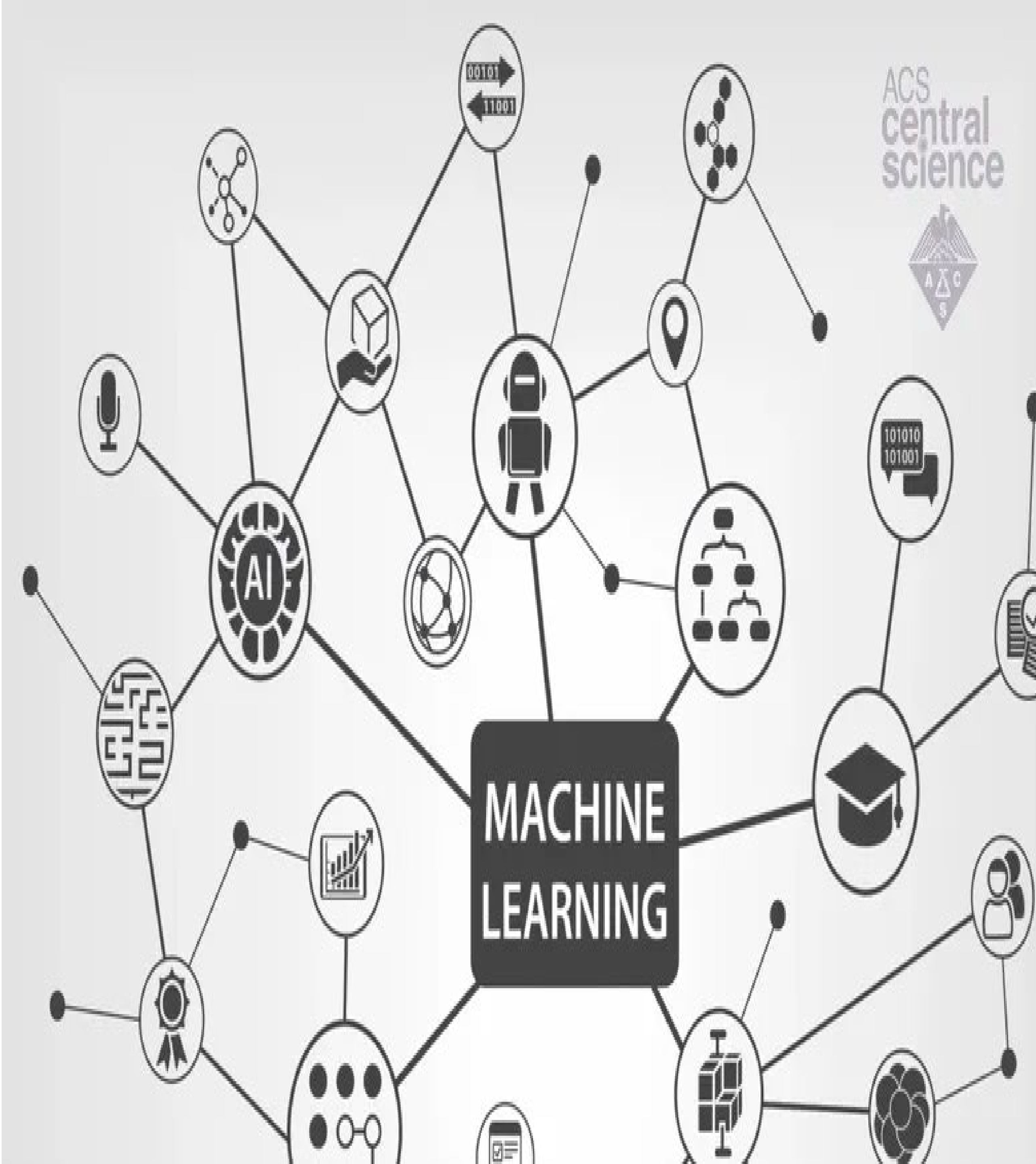
Supervised Learning Project

LIGHTHOUSE LABS

By: Rana Al-suffi



For this project:

- The dataset that I used was diabetes.csv.
- I used Jupyter lab
- Libraries include pandas, NumPy,

  sklearn, and seaborn

LIGHTHOUSE LABS

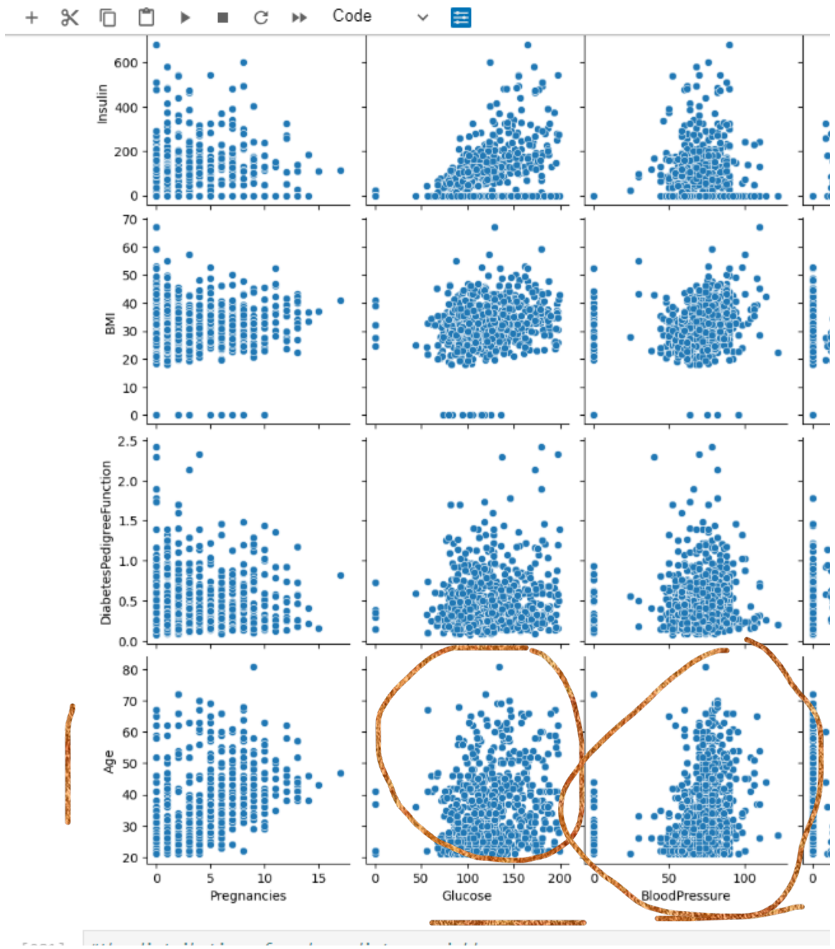
Processing the data:

Upload the data using pandas

Explore the data using different methods to see what data type there is in each columns.

All data type were numerical

Separate the dataset to predictor variables and target and try to find the relationships between each predictor variable with each other using pairplot.

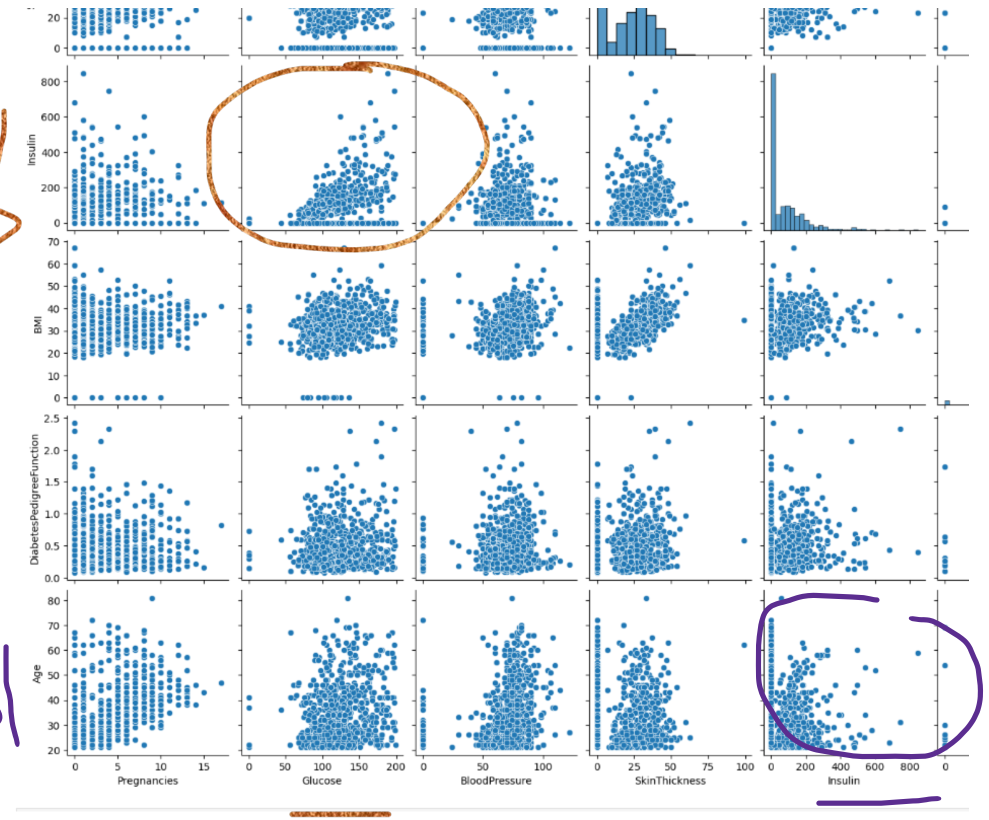


From this plot, if we exclude the outliers, we can see that there is a relationship between age, glucose level and blood pressure. The older the individual is the higher the glucose level and blood pressure.

As well…..

If we ignore the outliers, we can see a negative relationship between Glucose level and Insulin, the higher the insulin the lower the glucose level.

We can see another relationship between age and insulin. It shows that insulin level decreases with age.



- To observe each variable, I used displot. From that I noticed that some of the variables have a value of zero.
- These variables are:
    - BMI
    - Glucose Level
    - Blood pressure
    - Skin thickness
    - Insulin
- For the first four, I replace zero values with the mean value for each column.
- I didn't replace the zero values for insulin with the mean as doing that might skew the data, instead I will remove all the rows with zero insulin. I will process the data without removing the insulin zero values and compare it to the result when removing the rows with zero insulin values.

me

mean absolute error

Training ML model

I used three models:

- □ Linear Regression
- □ Decision Tree
- □ GaussianNB

For evaluating, Decision Tree & GaussianNB, used:

- Accuracy testing
- Precision
- F1-score
- roc_auc_score

For evaluating, Linear Regression & Ridge Regression, used:

- MAE (Mean Absolute Error)
- MSE(Mean Squared Error)

The Result

Before removing zero insulin values from the data:

Decision Tree:

- Training set score: 0.79
- Test set score: 0.72
- Model F1 score with criterion gini index: 0.55
- Model accuracy score with criterion gini index: 0.72
- Model precision score with criterion gini index: 0.70
- Model roc_auc_score with criterion gini index: 0.67

GaussianNB:

- Training set score: 0.76
- Test set score: 0.74
- GaussianNB F1 score: 0.63
- Model accuracy score with GaussianNB index: 0.74
- Model precision score with GaussianNB index: 0.63
- Model roc_auc_score with GaussianNB index: 0.71

LIGHTHOUSE LABS

For Linear Regression and Ridge Model:

Linear Regression:

- The train score for lr model is 0.3312619381748778
- The test score for lr model is 0.29588640885791295
- Model mean absolute error with Linear Regression index: 0.32
- Model precision score with Linear Regression index: 0.40

Ridge Model:

- The train score for ridge model is 0.3312619248620772
- The test score for ridge model is 0.2959046742755318
- Model mean absolute error with Ridge Regression index: 0.32
- Model precision score with Ridge Regression index: 0.40

- Both Decision tree and GaussianNB are better models compared to linear regression as linear regression has 33% prediction, however Decision tree has 72% prediction and GaussianNB has 74%

❑ Decision Tree has 72% accuracy whereas GaussianNB has 74% which makes GaussianNB a better model for prediction in general. However, precision, which refers to the number of true positives, is 64% in GaussianNB and 70% in decision tree.

❑ roc_auc_score, which tells us how efficient the model is, is higher in GaussianNB (71%) compared to 67% in Decision Tree .

❑ As this is a health dataset, giving more false positives is better then less false positives as it usually needs to be confirmed by a blood test anyway. Therefore, I would say GaussianNB is a better model.

After removing zero insulin values from the data:

Decision Tree:

- Training set score: 0.85
- Test set score: 0.76
- Model F1 score with criterion gini index: 0.65
- Model accuracy score with criterion gini index: 0.76
- Model precision score with criterion gini index: 0.73
- Model roc_auc_score with criterion gini index: 0.74

GaussianNB:

- Training set score: 0.77
- Test set score: 0.77
- GaussianNB F1 score: 0.65
- Model accuracy score with GaussianNB index: 0.77
- Model precision score with GaussianNB index: 0.66
- Model roc_auc_score with GaussianNB index: 0.74

- Although I ran the linear regression and ridge regression again, the result was like the result before remove the zeros from the insulin column. I won't include it in my conclusion.
- After removing the zero values from the insulin column, the model performance increased to a higher level. The Decision Tree model prediction jumped to 76% from 72%. For GaussianNB, it jumped to 77% from 74%.
- All other testing increased in both models; however, I still think GaussianNB is better because of the reason above.
- Removing the zero values had a good impact on the model as it makes the data more reliable.
- For further investigation, I will remove rows with zero insulin only from patients who don’t have diabetes.

Thank you