Supervised Learning:
Diabetes Prediction

June 28, 2023 Jennifer Zhao

Objective

The objective of this project is to develop a predictive model to diagnostically predict whether a patient has diabetes.

The dataset used for this analysis is the "Diabetes" dataset from the National Institute of Diabetes and Digestive and Kidney Diseases.

Through this project, we aim to gain insights from the dataset and build a reliable model for diabetes prediction.

Process


Exploring data and analysis
Data cleaning and preprocessing
Training a machine learning model

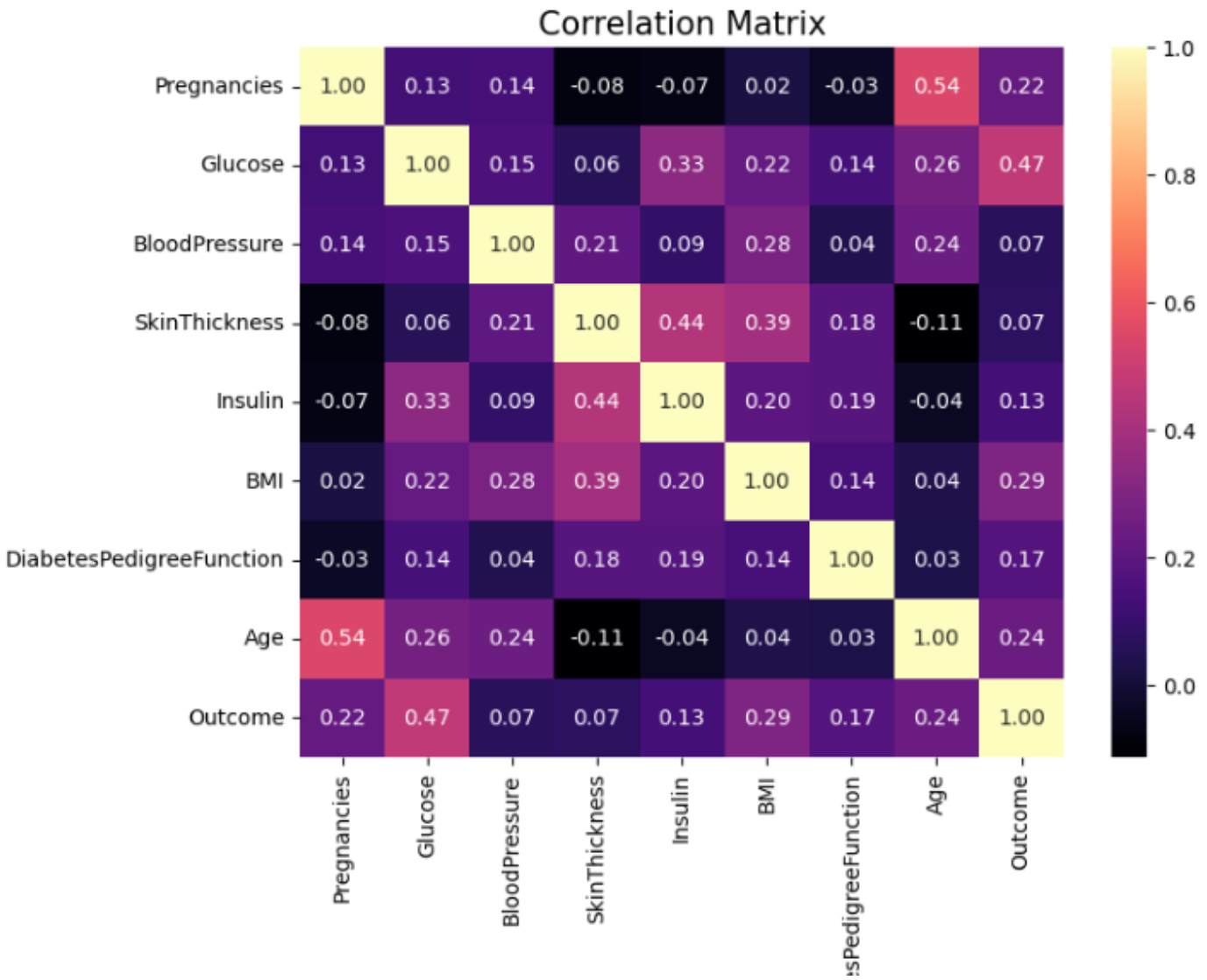
Exploring the Dataset

	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

Description of the dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Correlation Matrix



Data analysis and preprocessing

1.Column Analysis:
- •Check column meaning and type.
- •Verify dataset shape, min, and max values.

2.Duplicates Check:
- •Ensure no duplicates exist in the dataset.

3.Handling Missing and Zero Values:
- •Treat zeros as null values, except for specific columns.
- •Drop rows with zero values in selected columns.
- •Replace missing values with the mean value for a specific column.

4.Outlier Handling:
- •Use the Interquartile Range (IQR) method to identify outliers.
- •Preserve extremely high values for "Insulin" without modification.

Data analysis and preprocessing

1.Column Analysis:
- •Check column meaning and type.
- •Verify dataset shape, min, and max values.

2.Duplicates Check:
- •Ensure no duplicates exist in the dataset.

3.Handling Missing and Zero Values:
- •Treat zeros as null values, except for specific columns.
- •Drop rows with zero values in selected columns.
- •Replace missing values with the mean value for a specific column.

4.Outlier Handling:
- •Use the Interquartile Range (IQR) method to identify outliers.
- •Preserve extremely high values for "Insulin" without modification.

Model Building and Evaluation

1.Define X and y:
- Define the input features (X) and target variable (y) for the model.

2.Split Dataset:
- Split the dataset into training and testing sets.
- Use the training set to train the model and the testing set to evaluate its performance.

3.Model Building:
- Build the Random Forest model.
- Build the Decision Tree model.
- Build the XGBoost model.

4.Compare and Analyze Metrics:
- Evaluate and compare the performance metrics of the Random Forest, Decision Tree, and XGBoost models.

Comparing Machine Learning Models

Random Forest classifier, which exhibits the highest level of accuracy in predicting.

Random Forest Model:

```
Accuracy on training set: 0.800
Accuracy on test set: 0.786
```

Decision tree:

```
[[75 24]
 [15 40]]
              precision    recall  f1-score   support

           0       0.83      0.76      0.79        99
           1       0.62      0.73      0.67        55

    accuracy                           0.75       154
   macro avg       0.73      0.74      0.73       154
weighted avg       0.76      0.75      0.75       154
```

XGBoost:

```
[[70 29]
 [19 36]]
              precision    recall  f1-score   support

           0       0.79      0.71      0.74        99
           1       0.55      0.65      0.60        55

    accuracy                           0.69       154
   macro avg       0.67      0.68      0.67       154
weighted avg       0.70      0.69      0.69       154
```

Key Findings

EDA: When analyzing the data, it is crucial to pay attention to zero values. It is important to understand the units of each column and determine the normal range to gain insights into potential outliers. Before decided the outlier, I searched up if the max value of Insulin and Glucose are reasonable.

Data Patterns: Exploratory data analysis revealed patterns and correlations within the diabetes dataset. For instance, there was a positive correlation between glucose level and diabetes diagnosis, indicating that higher glucose levels are associated with an increased risk of diabetes. Factors like BMI and age also showed some degree of correlation with diabetes occurrence.

Feature Importance: The Random Forest model identified certain features as significant predictors of diabetes. These features, such as glucose level, body mass index (BMI), and age, exhibited higher importance in influencing the model's predictions.

Model Performance: Among the three models compared, the Random Forest model outperformed the others in terms of accuracy. This suggests that the ensemble approach employed by Random Forest, which combines multiple decision trees, can effectively handle the complexity of the diabetes dataset and yield more accurate predictions.

Thank you!