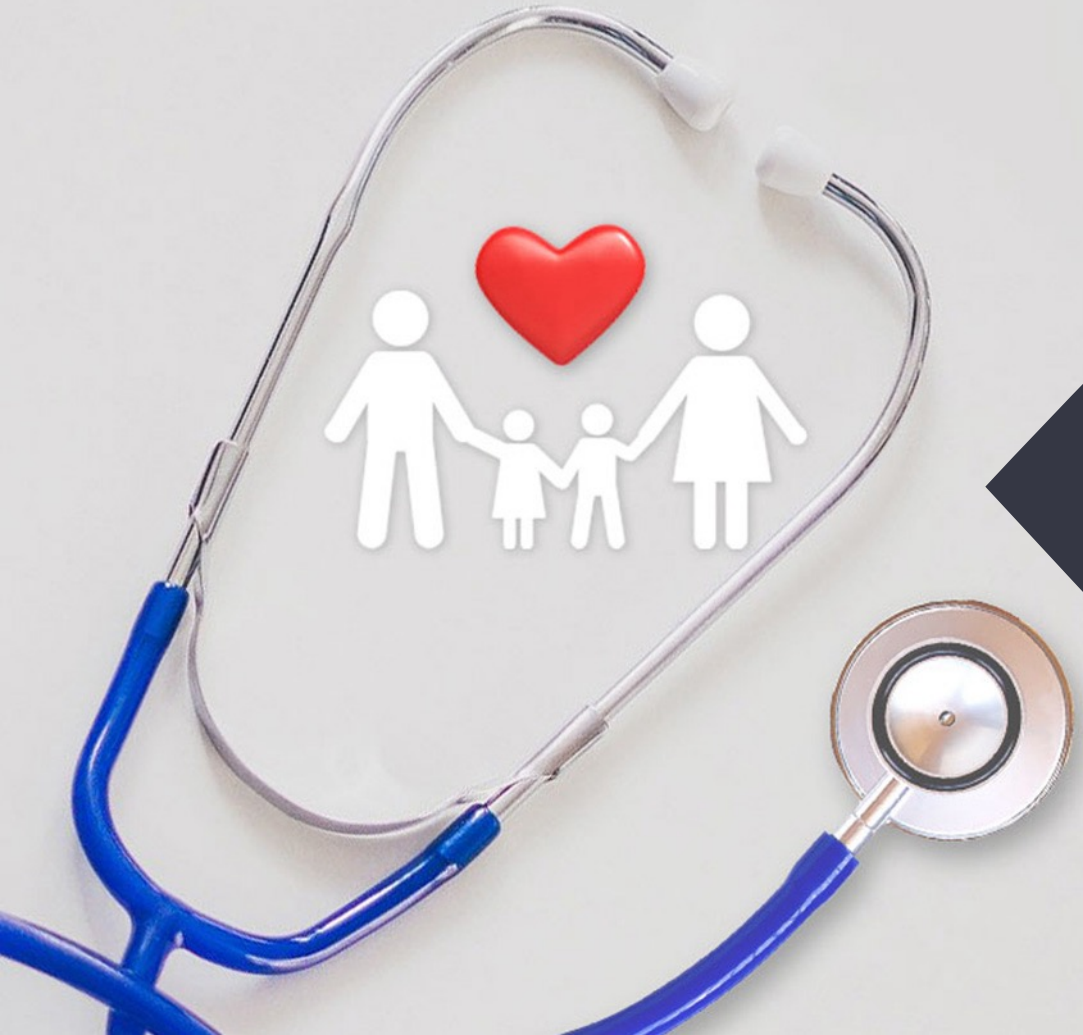


Capstone 2: Diabetes Prediction Project

Mengxiao Wang

Addressing Diabetes Prediction through Data Science

01 Purpose of project

- Diabetes is a chronic disease that affects millions worldwide. Early detection is important for effective management and prevention.

02 Problem Statement

- Many individuals remain undiagnosed due to lack of access to medical testing. There is an opportunity to leverage data science to predict diabetes risk using readily available health data.

03 Opportunity:

- Implementing a predictive model can provide early warnings and promote proactive healthcare measures.

Introduction of the Dataset

The diabetes dataset has 100,000 rows and 9 columns, with 3854 duplicate rows and no null values.

Resource: the Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset.

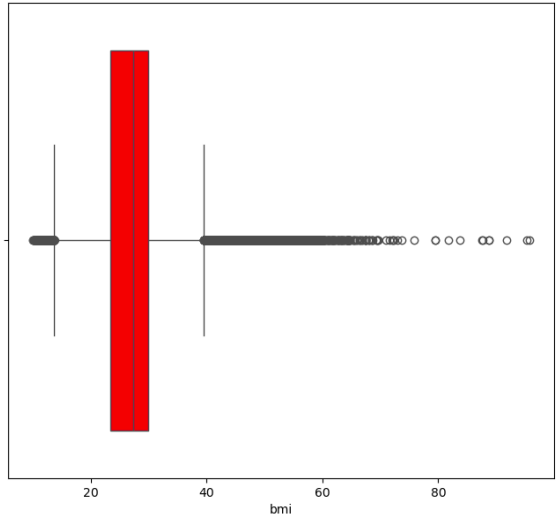
	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
59468	Female	2.0	0	0	No Info	27.32	5.0	158	0
67439	Female	2.0	0	0	No Info	27.32	6.0	85	0
67234	Female	2.0	0	0	No Info	27.32	6.0	145	0
97294	Female	2.0	0	0	No Info	27.32	6.2	145	0
89701	Female	2.0	0	0	No Info	27.32	6.5	155	0
...	...	...	...	...	...	...	...	...	...
54794	Male	80.0	0	0	No Info	27.32	6.6	159	0
75961	Male	80.0	0	0	No Info	27.32	6.6	159	0
46764	Male	80.0	0	0	No Info	27.32	6.6	160	0
73316	Male	80.0	0	0	No Info	27.32	6.6	160	0
77302	Male	80.0	0	0	No Info	27.32	6.6	160	0

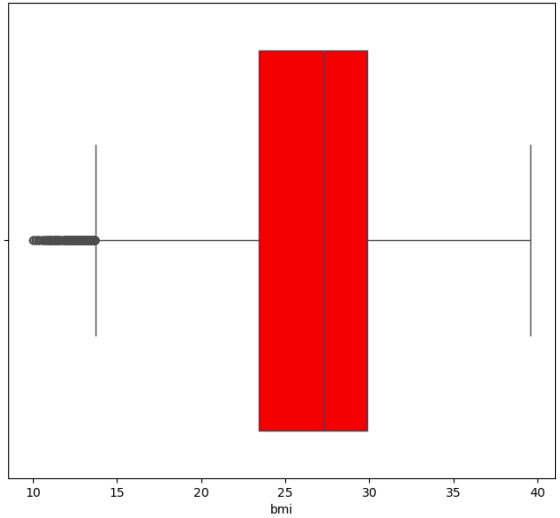
Preprocessing Procedures

Dropped **3854 duplicated rows**

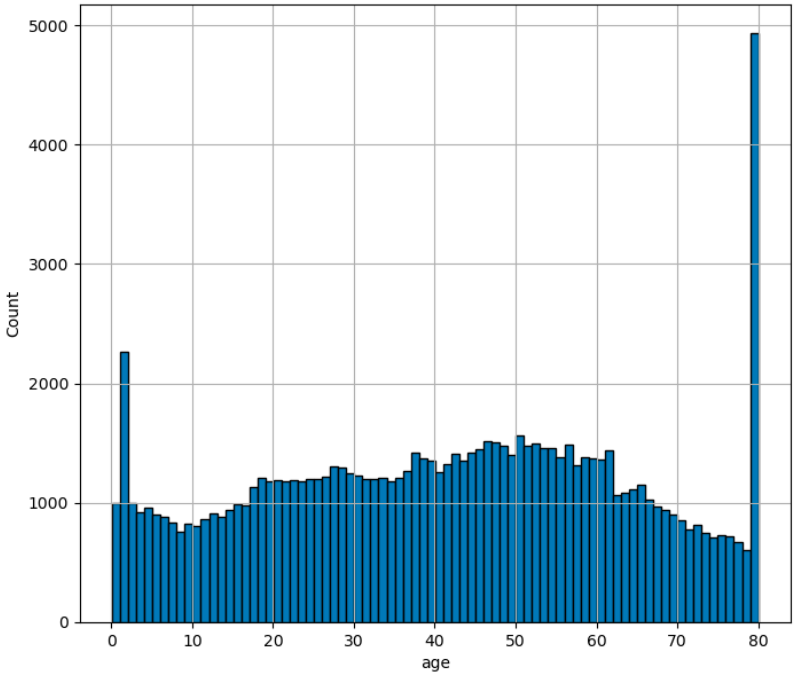
Handled **outliners for BMI, HbA1c_level and blood_glucose_level**

Mapping **the smoking_history column** for simplification reason

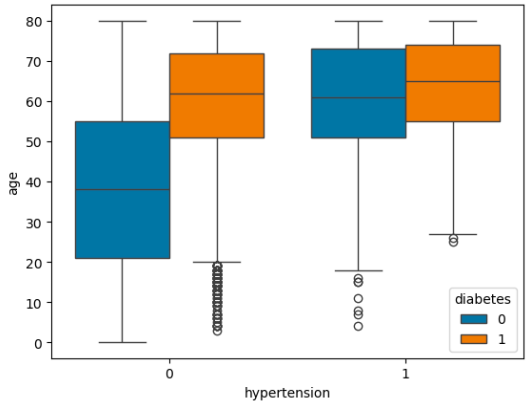




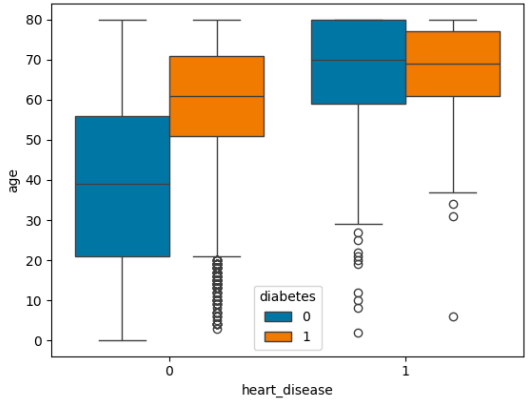
Findings from EDA



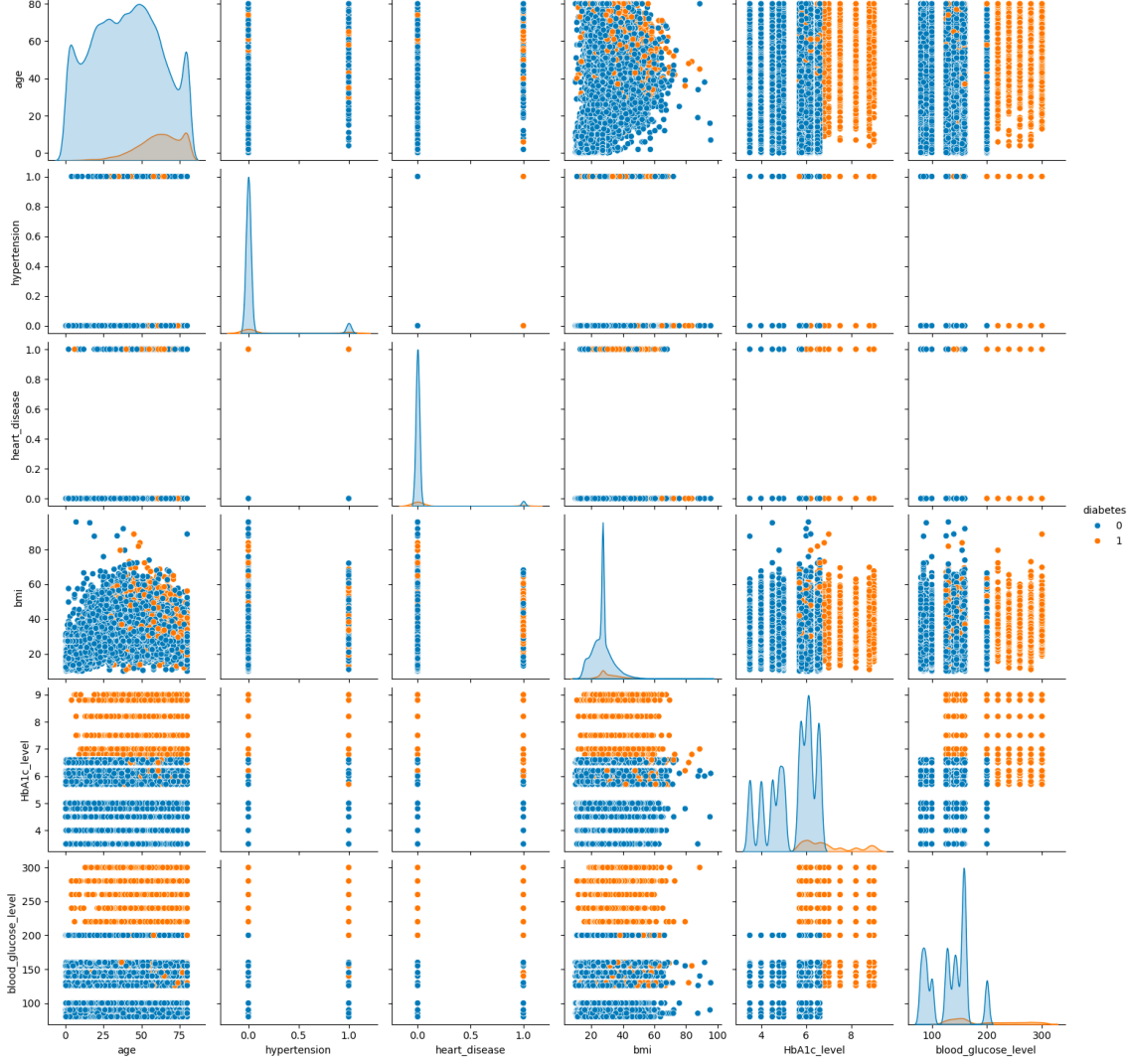
there are noticeable peaks at age 0 and age 80



Elder people more easy to get hypertension. Also, hypertension may independently contribute to the risk of diabetes regardless of age.



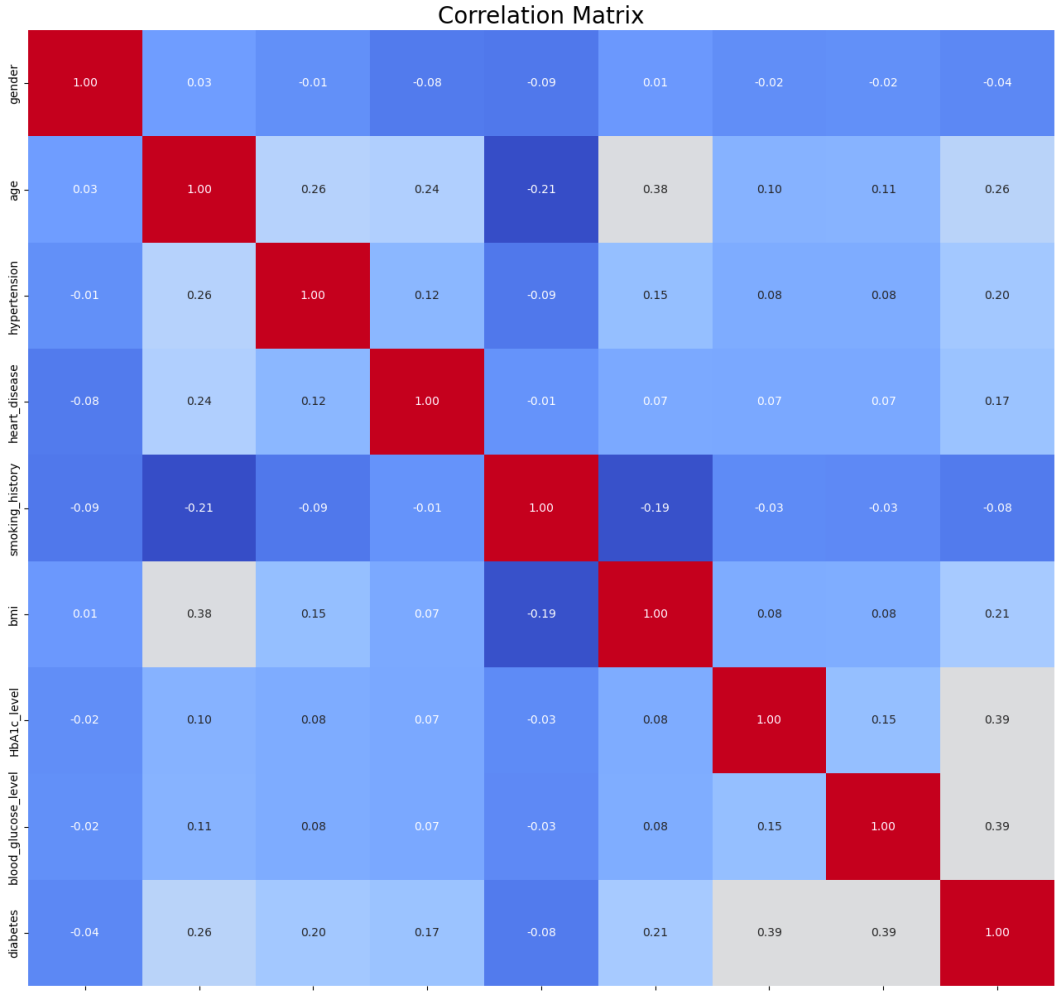
Same with hyperextension, age and heart_disease are strongly related, heart_disease may independently contribute to the risk of diabetes with only a small influence from age.



According to the pairplot, we can see that

- Age is a strong predictor of diabetes, with older individuals at higher risk.
- Hypertension and heart disease are associated with a higher prevalence of diabetes.
- BMI, HbA1c level, and blood glucose level are significant indicators of diabetes. Higher values in these features are strongly associated with the presence of diabetes.

I intend to use hypothesis testing to determine whether my assumption is valid.



HbA1c Level and Blood Glucose Level have strong correlation with diabetes of 0.41 and 0.42.

Age, BMI and hypertention, heart_disease have moderate correlation with diabetes of 0.26, 0.21, 0.2, and 0.17.

Smoking_histroy and gender have low correlation with diabetes

Modeling

	precision	recall	f1-score	support
0	0.97	0.99	0.98	17562
1	0.84	0.63	0.72	1668
accuracy			0.96	19230
macro avg	0.90	0.81	0.85	19230
weighted avg	0.95	0.96	0.95	19230

Logistic Regression(standardization is used)

My precision, recall, and F1-score are significantly higher for individuals without diabetes compared to those with diabetes It is because my dataset is not balanced. Therefore, I need to implement upsampling to address this issue.

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.87	0.86	17532
1	0.87	0.84	0.85	17534
accuracy			0.86	35066
macro avg	0.86	0.86	0.86	35066
weighted avg	0.86	0.86	0.86	35066

After upsampling

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.87	0.86	1686
1	0.87	0.84	0.85	1707
accuracy			0.85	3393
macro avg	0.86	0.86	0.85	3393
weighted avg	0.86	0.85	0.85	3393

After downsampling

	precision	recall	f1-score	support
0	0.90	0.90	0.90	17439
1	0.90	0.91	0.90	17627
accuracy			0.90	35066
macro avg	0.90	0.90	0.90	35066
weighted avg	0.90	0.90	0.90	35066

Accuracy: 0.9007585695545542

After smote

Modeling

Decision Tree

	precision	recall	f1-score	support
0	0.97	0.97	0.97	17534
1	0.69	0.74	0.72	1696
accuracy			0.95	19230
macro avg	0.83	0.86	0.84	19230
weighted avg	0.95	0.95	0.95	19230

Original classification_report

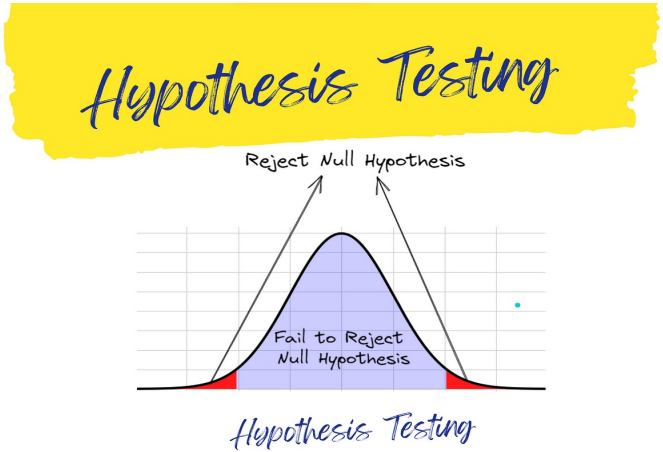
	precision	recall	f1-score	support
0	0.88	0.87	0.87	1686
1	0.87	0.88	0.88	1707
accuracy			0.87	3393
macro avg	0.87	0.87	0.87	3393
weighted avg	0.87	0.87	0.87	3393

After downsampling

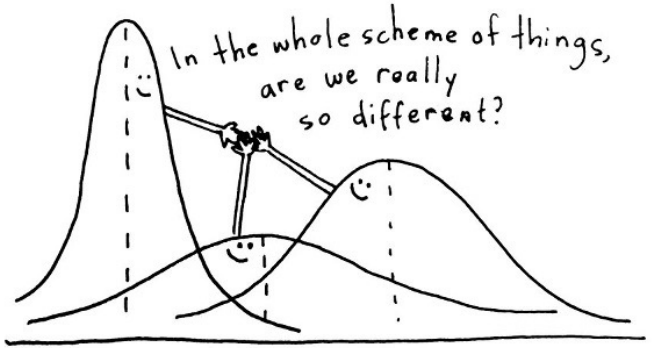
	precision	recall	f1-score	support
0	0.97	0.97	0.97	17439
1	0.97	0.97	0.97	17627
accuracy			0.97	35066
macro avg	0.97	0.97	0.97	35066
weighted avg	0.97	0.97	0.97	35066

After Smote

Next Step



ANOVA Testing



Model Tuning:

•Hyperparameter Optimization: Use techniques like Grid Search or Random Search to find the best hyperparameters for your models.

•Ensemble Methods: Consider using ensemble methods to combine the strengths of different models and improve overall performance.


Questions?