



DIABETES PREDICTION PROJECT

A Machine Learning Approach

WOGUNKA JOHN NNADI

Project Outline



Introduction



Objective



Exploratory
Data Analysis



Feature
Engineering



Model
Evaluation



Introduction

- This project focuses on developing machine learning models to predict diabetes risk based on various health indicators. The goal is to create accurate predictive models that can help identify individuals at risk of developing diabetes, potentially enabling earlier intervention and better health outcomes.

Dataset Overview

```
# General dataset info
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 100000 entries, 0 to 99999
```

```
Data columns (total 9 columns):
```

#	Column	Non-Null Count	Dtype
0	gender	100000 non-null	object
1	age	100000 non-null	float64
2	hypertension	100000 non-null	int64
3	heart_disease	100000 non-null	int64
4	smoking_history	100000 non-null	object
5	bmi	100000 non-null	float64
6	HbA1c_level	100000 non-null	float64
7	blood_glucose_level	100000 non-null	int64
8	diabetes	100000 non-null	int64

```
dtypes: float64(3), int64(4), object(2)
```

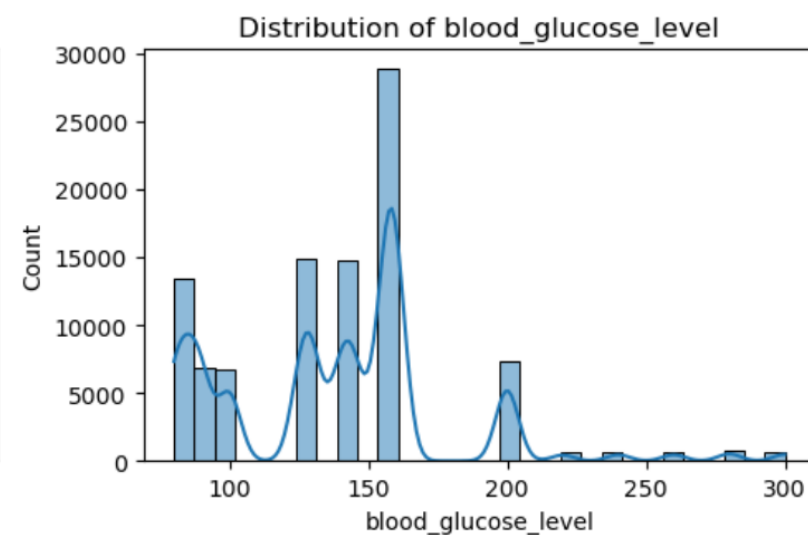
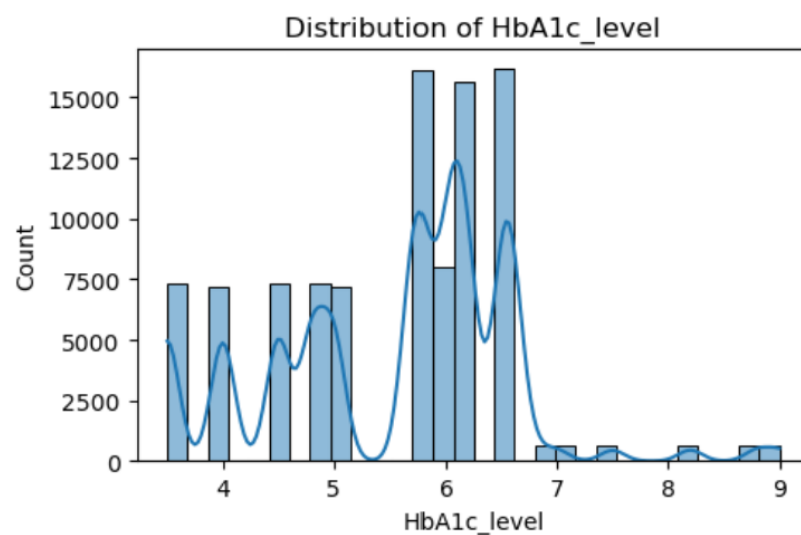
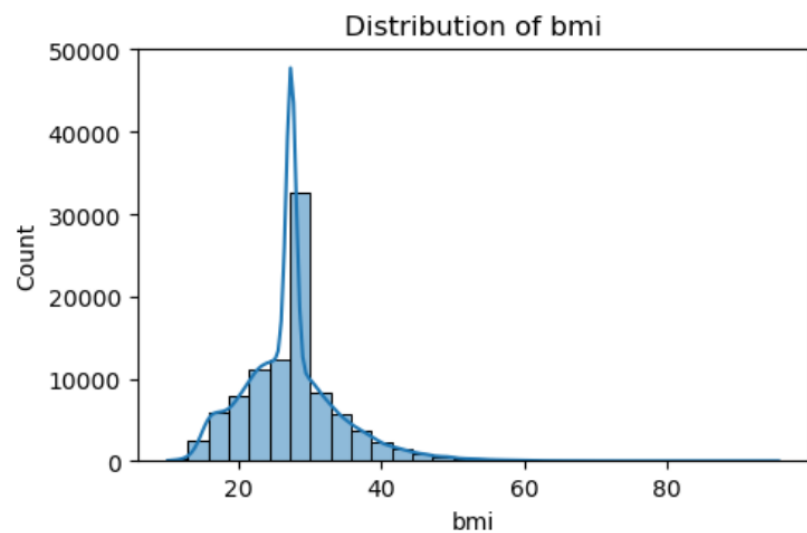
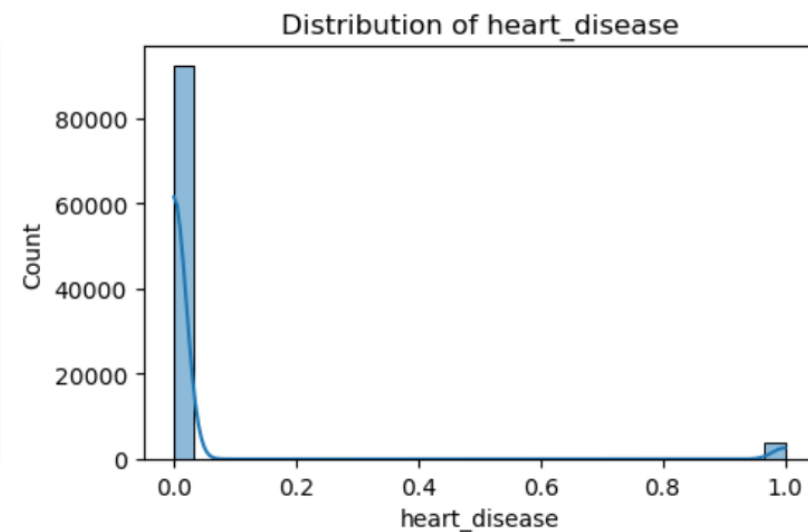
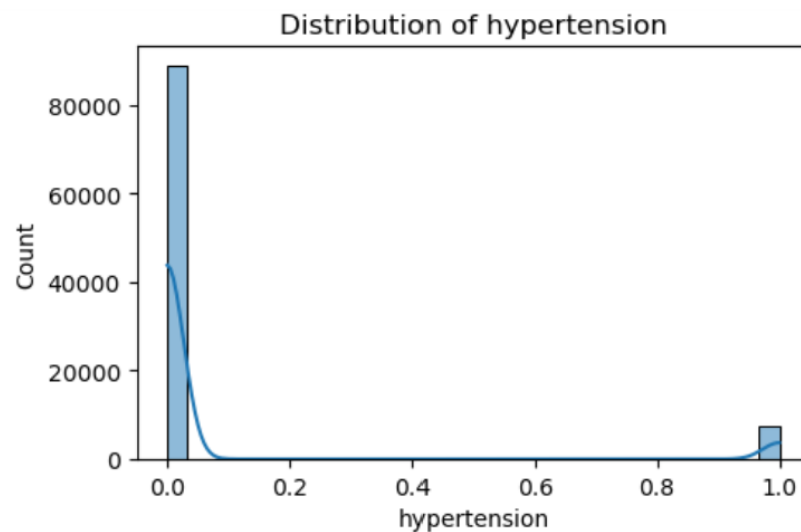
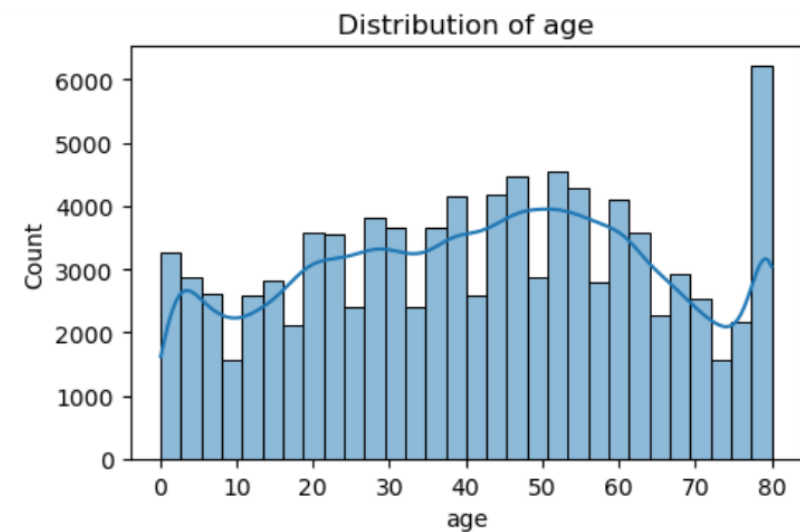
```
memory usage: 6.9+ MB
```

Objective

- The goal of this project is to leverage on Machine Learning to predict the likelihood of diabetes onset, allowing for timely and targeted preventive measures. This initiative will empower Stark Health to enhance patient outcomes, reduce the burden on healthcare resources, and play a proactive role in combating diabetes.

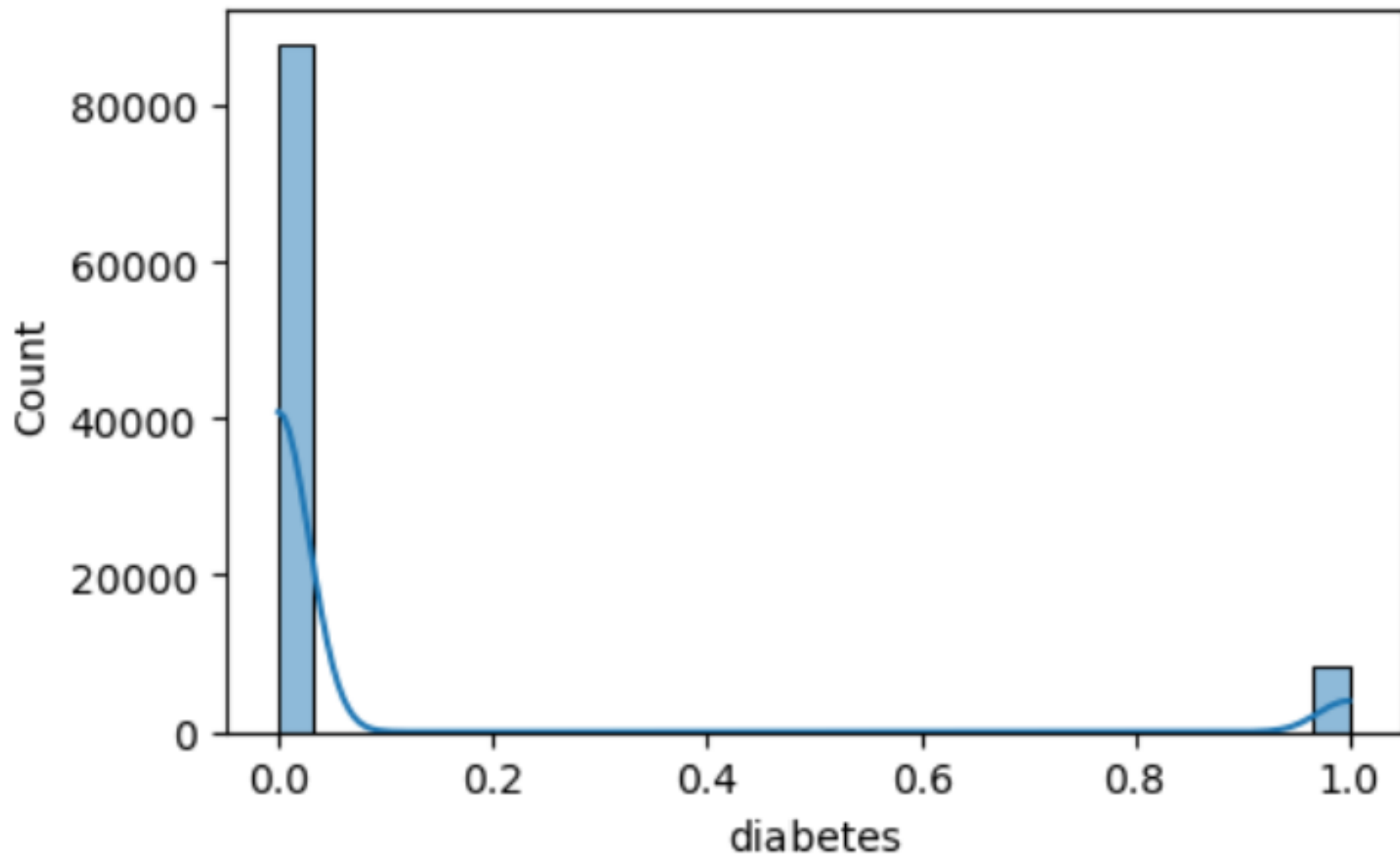


EDA



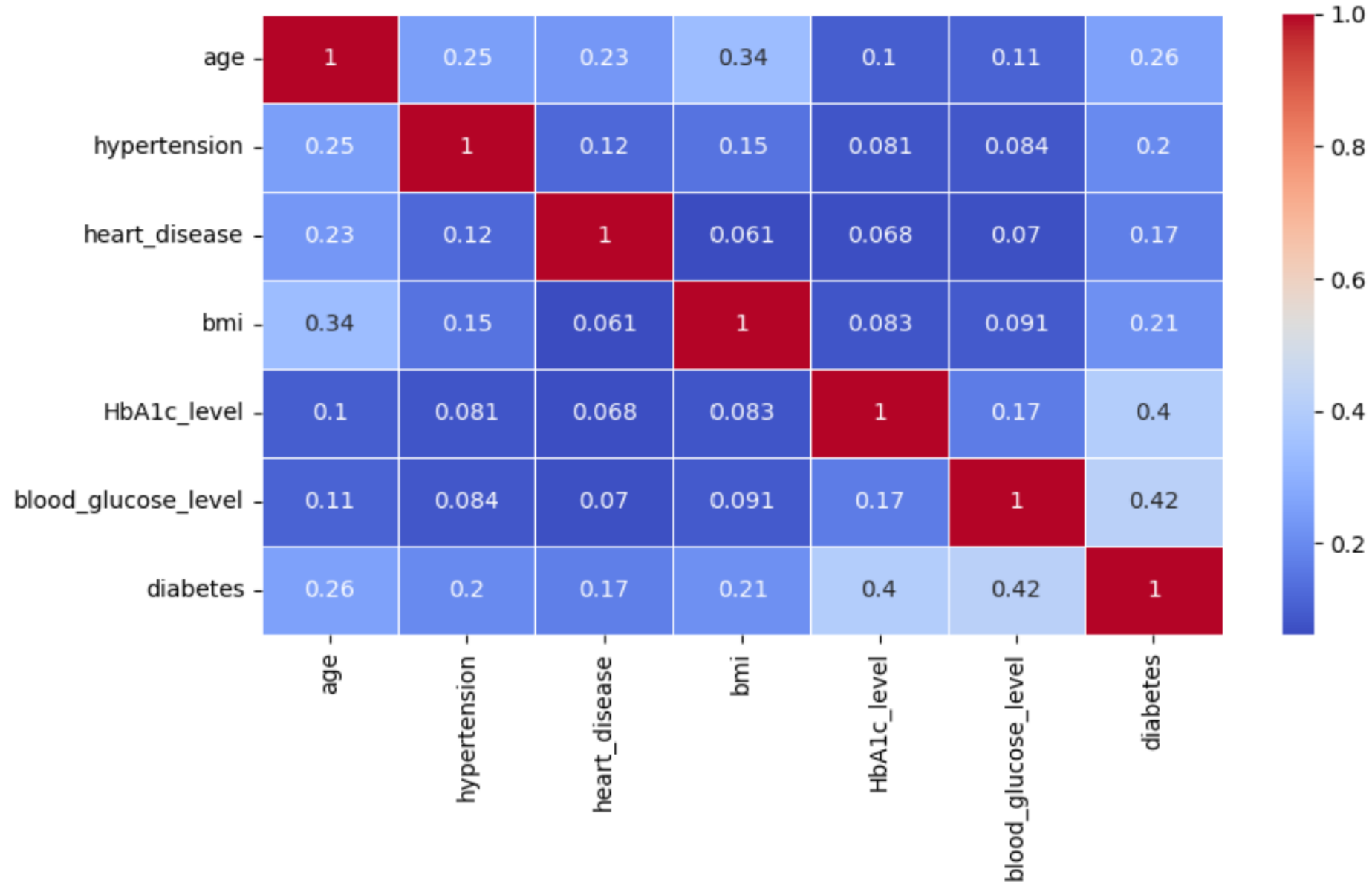
EDA

Distribution of diabetes



Feature Engineering

Correlation Matrix of Numerical Features



Model Evaluation

Accuracy	ROC AUC	F1 Score
0.9591263651	0.9593249826	0.7322888283
0.9676547062	0.9543493228	0.7926666667
0.9705668227	0.975550635	0.8076138681
0.963650546	0.9281299634	0.7444241316



Conclusion

- The evaluation results reveal that overall model performance is good. XGBoost achieved the highest ROC AUC (approximately 0.976) and has the best F1 score (approximately 0.808), while all models show high accuracy.