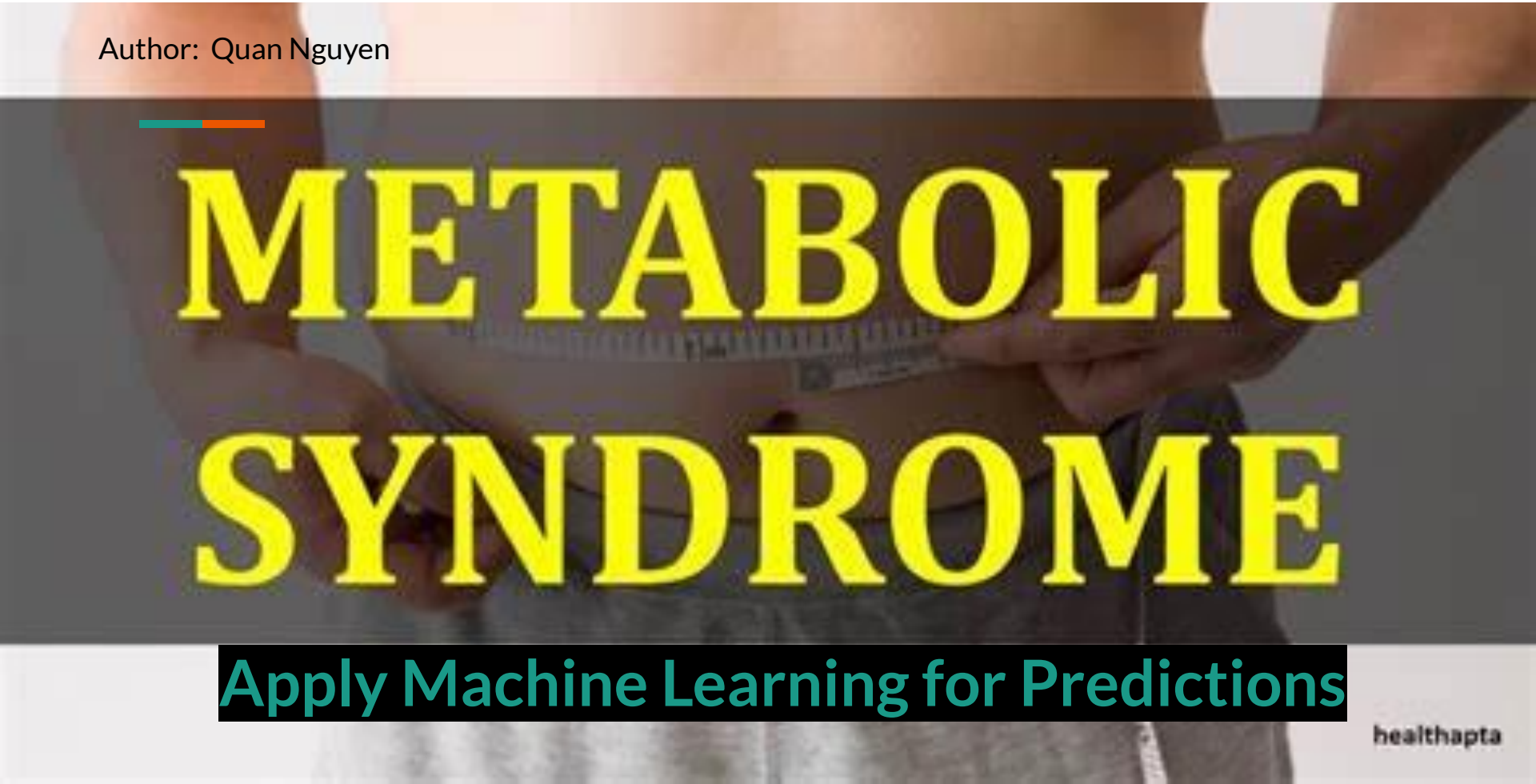
Author: Quan Nguyen



Apply Machine Learning for Predictions

Insulin Resistance

High Blood Pressure

METABOLIC SYNDROME

High Triglyceride Levels

Low HDL Cholesterol

healthydwell.com

Data Overview

nhanes National Health and Nutrition Examination Survey

Data Dictionary

Feature	Description
seqn	Respondent sequence number
Age	Age in Years
Sex	Male or Female
Marital	Marital status
Income	Income in Dollars
Race	Racial Group
WaistCirc	Waist Circumference
BMI	Body Mass Index
Albuminuria	Albumin in Urine
UrAlbCr	Ratio of Albumin to Creatinine in Urine
UricAcid	Uric Acid in Blood
BloodGlucose	Glucose in Blood
HDL	High Density Lipoprotein
Triglycerides	Triglycerides in Blood
MetabolicSyndrome	Presence or not of at least three of the five metabolic risk factors

**The dataset includes of 2401 rows, and 15 columns.*

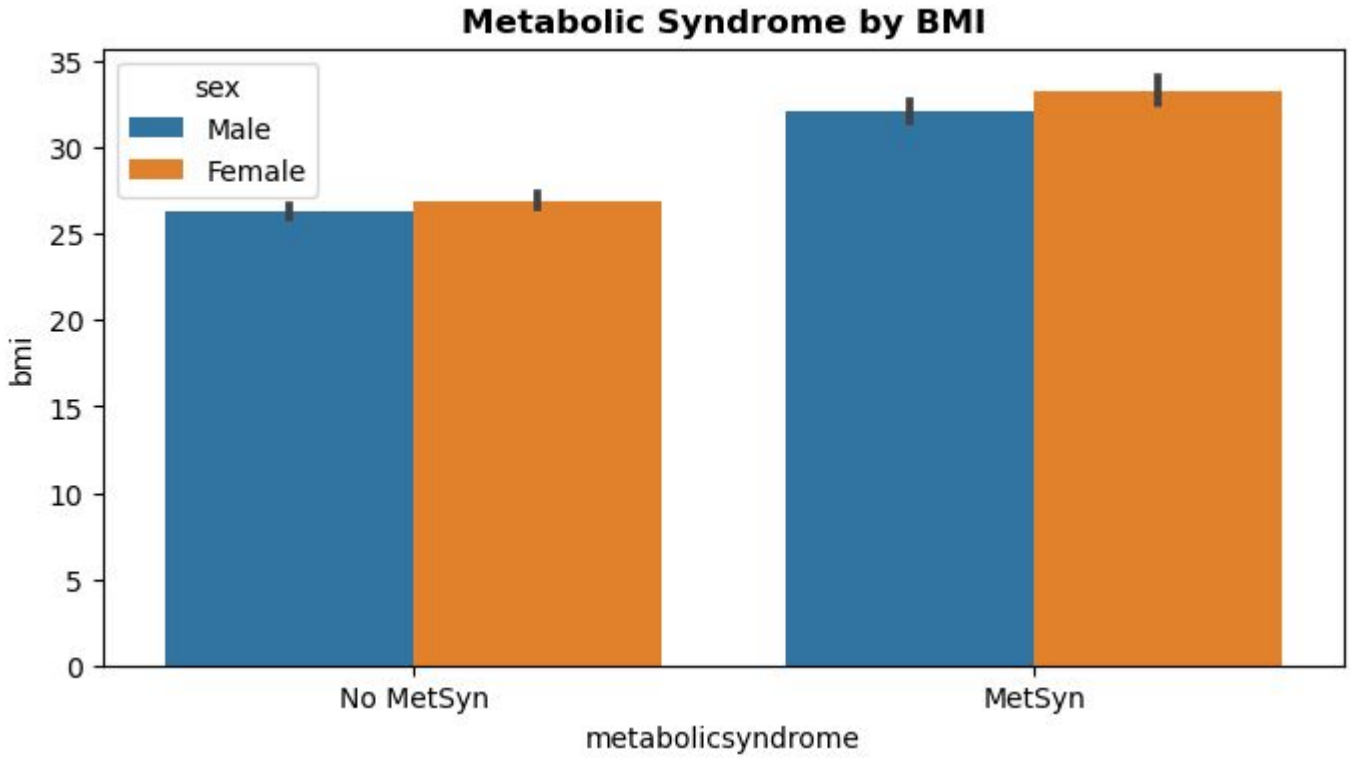
Objective:

1. Build predictive models
2. Develop all models and select the best performed model.
3. Gain insights into the relationships between different variables and their contribution to metabolic syndrome risk prediction.

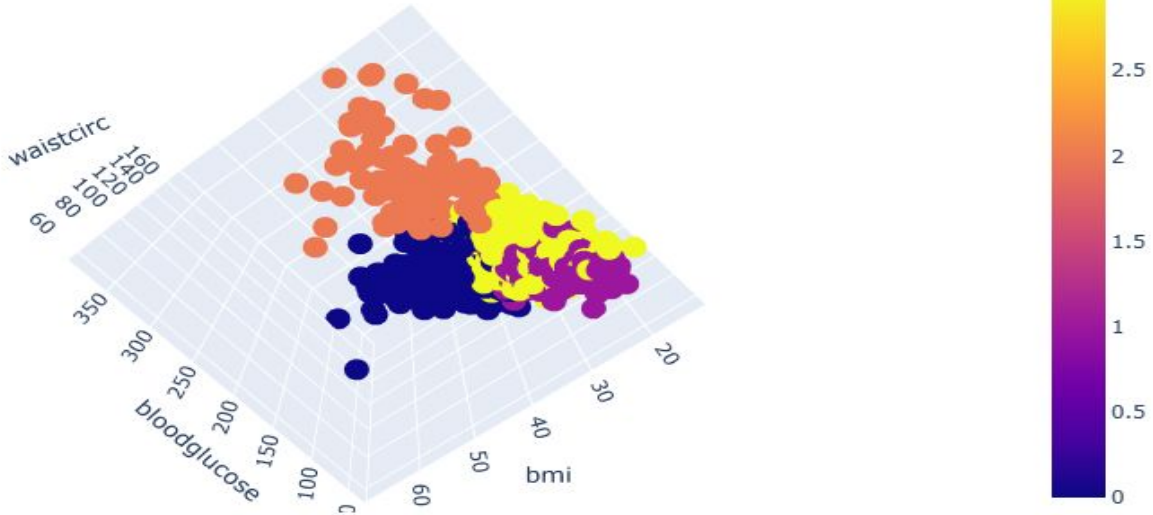
Methodology

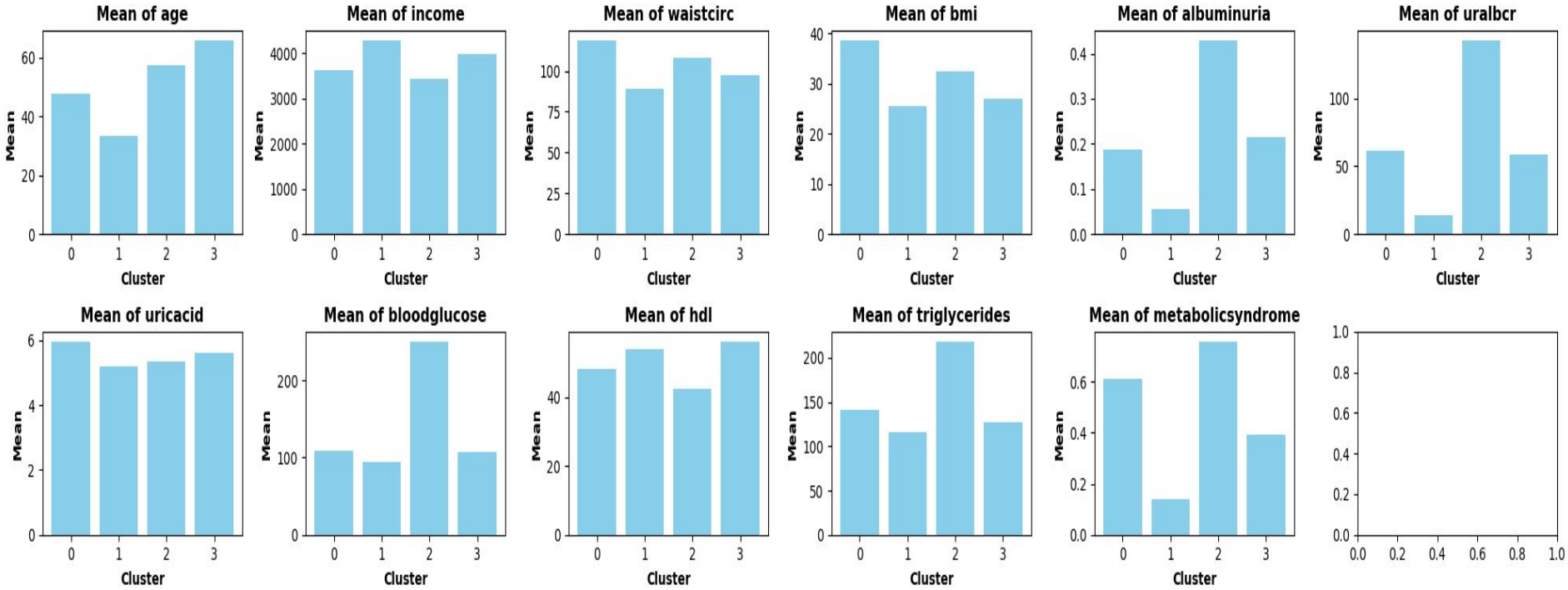
1. Data Preprocessing and Visualizations.
2. Model training.
3. Feature Selection (PCA) and Engineering.
4. Clustering.
5. Model Evaluation.
6. Best model interpretation and Insights.

Data Visualization for target column



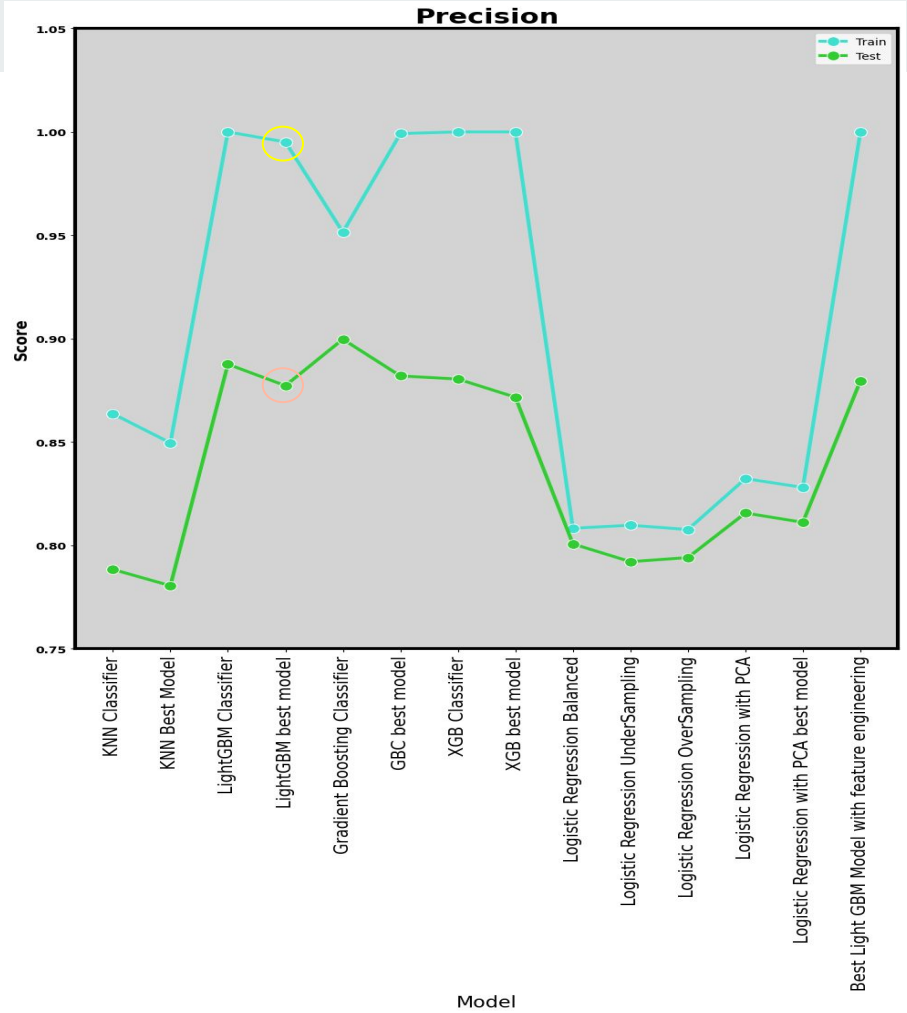
Clustering

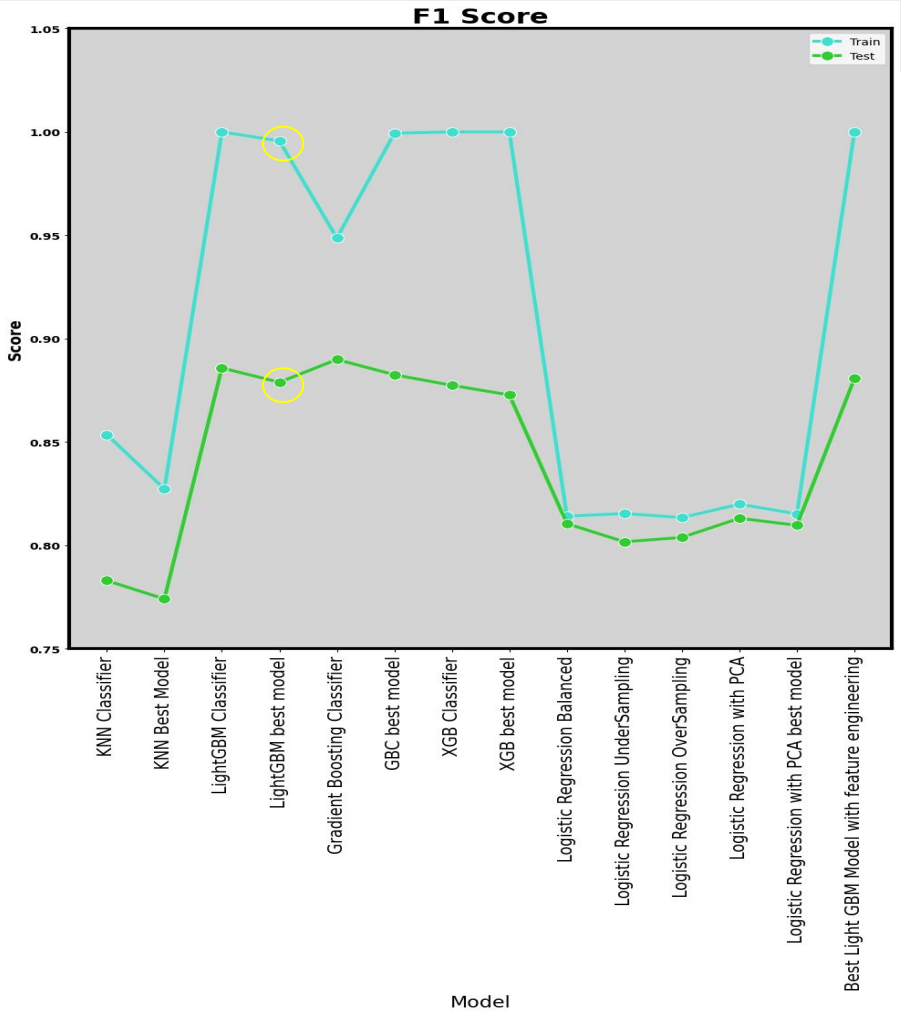

waistcirc
60
80
100
120
140
160
bloodglucose
350
300
250
200
150
100
bmi
60
50
40
30
20
cluster
3
2.5
2
1.5
1
0.5
0


Mean of age
Mean of income
Mean of waistcirc
Mean of bmi
Mean of albuminuria
Mean of uralbcr
Mean of uricacid
Mean of bloodglucose
Mean of hdl
Mean of triglycerides
Mean of metabolicsyndrome
Mean
Cluster

Models Comparison and Recommendation

	Train Accuracy	Train Recall	Train Precision	Train F1-Score	Test Accuracy	Test Recall	Test Precision	Test F1-Score
KNN Classifier	0.868946	0.846083	0.863689	0.853552	0.817094	0.778359	0.788393	0.782958
KNN Best Model	0.849003	0.814746	0.849587	0.827311	0.810256	0.768876	0.780510	0.774115
LightGBM Classifier	1.000000	1.000000	1.000000	1.000000	0.902564	0.884084	0.887635	0.885824
LightGBM best model	0.996011	0.996186	0.995121	0.995650	0.895726	0.880627	0.877264	0.878911
Gradient Boosting Classifier	0.953276	0.946098	0.951476	0.948684	0.907692	0.881780	0.899709	0.889930
GBC best model	0.999430	0.999558	0.999199	0.999378	0.899145	0.883109	0.881963	0.882532
XGB Classifier	1.000000	1.000000	1.000000	1.000000	0.895726	0.874601	0.880463	0.877434
XGB best model	1.000000	1.000000	1.000000	1.000000	0.890598	0.873892	0.871673	0.872767
Logistic Regression Balanced	0.824501	0.825344	0.808291	0.814107	0.829060	0.829227	0.800716	0.810518
Logistic Regression UnderSampling	0.826211	0.825587	0.809757	0.815438	0.820513	0.821517	0.792124	0.801743
Logistic Regression OverSampling	0.823932	0.824542	0.807664	0.813450	0.822222	0.824264	0.794117	0.803858
Logistic Regression with PCA	0.839886	0.811648	0.832285	0.819949	0.841026	0.810794	0.815657	0.813137
Logistic Regression with PCA best model	0.835897	0.806752	0.828050	0.815226	0.837607	0.808313	0.811164	0.809707
Best Light GBM Model with feature engineering	1.000000	1.000000	1.000000	1.000000	0.897436	0.881868	0.879601	0.880719


Precision
Train
Test
Score
1.05
1.00
0.95
0.90
0.85
0.80
0.75
KNN Classifier
KNN Best Model
LightGBM Classifier
LightGBM best model
Gradient Boosting Classifier
GBC best model
XGB Classifier
XGB best model
Logistic Regression Balanced
Logistic Regression UnderSampling
Logistic Regression OverSampling
Logistic Regression with PCA
Logistic Regression with PCA best model
Best Light GBM Model with feature engineering
Model


F1 Score
Train
Test
Score
1.05
1.00
0.95
0.90
0.85
0.80
0.75
KNN Classifier
KNN Best Model
LightGBM Classifier
LightGBM best model
Gradient Boosting Classifier
GBC best model
XGB Classifier
XGB best model
Logistic Regression Balanced
Logistic Regression UnderSampling
Logistic Regression OverSampling
Logistic Regression with PCA
Logistic Regression with PCA best model
Best Light GBM Model with feature engineering
Model

Personal Data Insight:

- **Detecting Metabolic Syndrome Early**
  - **Benefits for Individuals and Public Health**
  - **Early Intervention: Lifestyle modifications, medication, and targeted interventions**
- **Benefits of Early Treatment:**
  - **Manage blood pressure, blood sugar, cholesterol levels, and weight**
  - **Reduce risk of cardiovascular diseases, diabetes, and related complications**
  - **Improve overall health outcomes and quality of life**
  - **Potentially reduce healthcare costs**

Data Insights for Healthcare Professionals

- **Benefits for Healthcare Professionals:**
  - **Streamline diagnostic process with data-driven approach**
  - **Save valuable time for personalized care and interventions**
- **Machine Learning and Metabolic Syndrome Dataset:**
  - **Analyze patterns and relationships for accurate syndrome detection**
- **Enhancing Clinical Decision-Making:**
  - **Incorporate variables and risk factors**
  - **Gain insights into underlying mechanisms and predictors**

Please give me some feedbacks!