Dataset: Pima Indians Diabetes Dataset (9 columns, 768 rows)

Columns:

0 Number of times pregnant

1 Plasma glucose concentration a 2 hours in an oral glucose tolerance test

2 Diastolic blood pressure (mm Hg)

3 Triceps skinfold thickness (mm)

4 2-Hour serum insulin (mu U/ml)

5 Body mass index (weight in kg/(height in m)^2)

6 Diabetes pedigree function

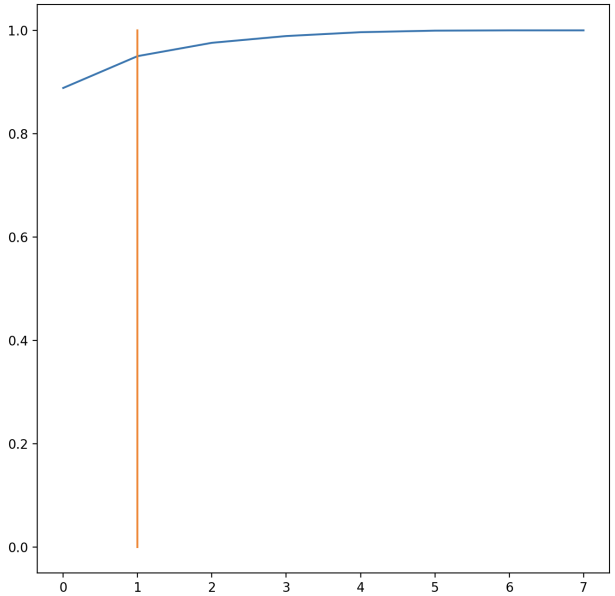
7 Age (years)

8 Class variable (0 or 1)

Model: LogisticRegressionCV

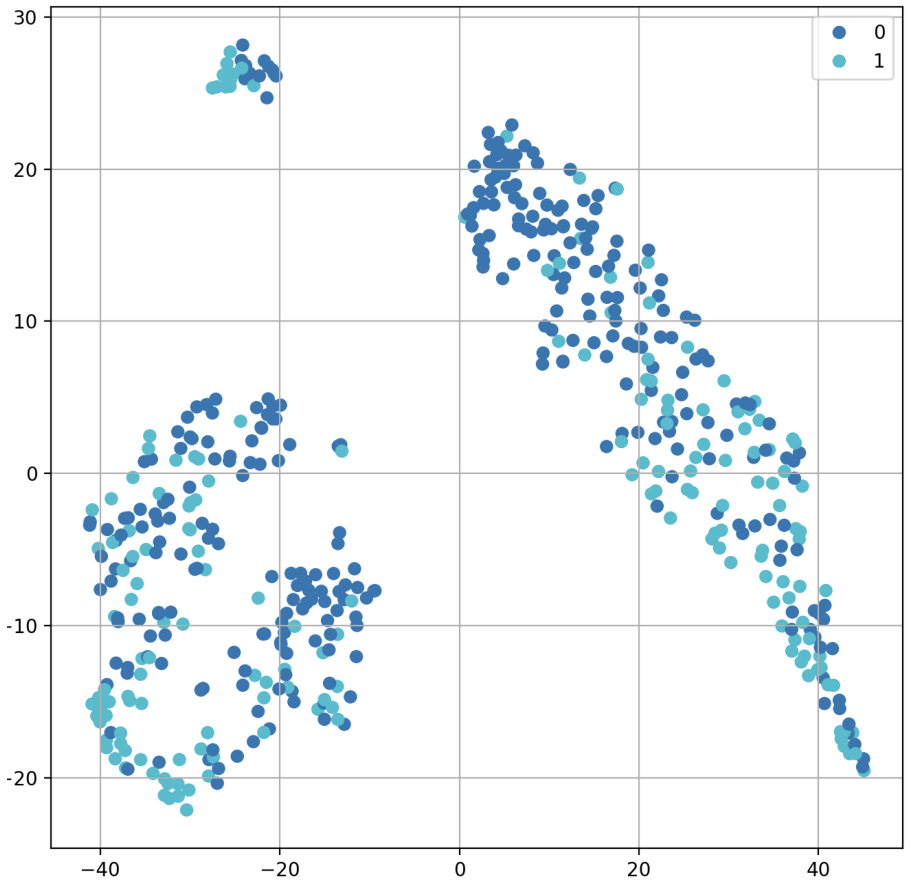
Знайдемо кількість головних компонент, яку можна використовувати для редукції розмірності даних так, щоб зберегти більше 95% варіації вихідних даних.

Графік explained variance:



Кількість головних компонент рівна 2.

Візуалізація датасету в 2D просторі:



Оцінимо час, витрачений на навчання моделі, та метрику оцінки якості моделі при 8 і 2 головних компонентах :

	8 components	2 components
ROC AUC, %	73,30827	64,76906
time, s	0,004069	0,002000

Бачимо, що застосування 2 головних компонентів призвело до гіршої метрики ROC AUC (65% проти 73%), але водночас зменшило час навчання майже вдвічі (0.002 с проти 0.004 с).

Цей датасет пов'язаний з діагностикою діабету, і тому, за можливості не рекомендується зменшувати час навчання шляхом втрати точності моделі.