Diabetes risk predictor

Data Science Bootcamp

Sprint 2

Created By:
Diego Villanueva

Agenda

1. Problem overview
2. Proposed solution
3. EDA and pre-processing
4. Models and evaluation
5. Next steps

Diabetes has become one of the biggest epidemics in human history

It is estimated that 422 million people are living with diabetes in the world...

422 M cases

4.8 M cases

... and almost half of them have not been diagnosed

46% of global cases are undiagnosed

In the UK, up to **1M people**

Sources: https://www.who.int/health-topics/diabetes#tab=tab_1
https://www.diabetes.co.uk/diabetes-prevalence.html#:~:text=World%20diabetes%20prevalence,with%20diabetes%20worldwide%20by%202040.

A ML solution

Early warning system to help **doctors predict** the risk of a patient developing diabetes

EDA /Pre-processing:
The dataset used has 20 indicators

Dataset with 20 indicators

Demographic

Lifestyle

Medical

Issues addressed during pre-processing:

- Inbalanced data
  - Over sampling
- Different numerical scale
  - Scaling
- Multicollinearity
  - Detection & drop features

4 models were developed to find the one with the best balance of recall and precision scores

	Logistic Regression
	Pipeline / Grid Search
	Standard Scaler
Accuracy	0.85
Precision	0.54
Recall	0.16
F1-score	0.25

4 models were developed to find the one with the best balance of recall and precision scores

Logistic Regression
Pipeline / Grid Search
Standard Scaler

Accuracy	0.85	-> if we tested 100 people, the model classfied 85 of them correctly
Precision	0.54	-> out of all the ppl we predicted to have diabetes, only 54% of them have it
Recall	0.16	-> out of all the ppl who have diabetes, we only detected 16% of them
F1-score	0.25	-> this is the balance between precision and recall

4 models were developed to find the one with the best balance of recall and precision scores

	Logistic Regression	Decision Tree
	Pipeline / Grid Search	Pipeline / Grid Search
	Standard Scaler	Standard Scaler
Accuracy	0.85	0.85
Precision	0.54	0.59
Recall	0.16	0.12
F1-score	0.25	0.20

4 models were developed to find the one with the best balance of recall and precision scores

	Logistic Regression	Decision Tree	Logistic Regression
	Pipeline / Grid Search	Pipeline / Grid Search	Pipeline / Grid Search
	Standard Scaler	Standard Scaler	Scaling + Oversampling
Accuracy	0.85	0.85	0.73
Precision	0.54	0.59	0.34
Recall	0.16	0.12	0.74
F1-score	0.25	0.20	0.46

4 models were developed to find the one with the best balance of recall and precision scores

	Logistic Regression	Decision Tree	Logistic Regression	Decision Tree
	Pipeline / Grid Search	Pipeline / Grid Search	Pipeline / Grid Search	Pipeline / Grid Search
	Standard Scaler	Standard Scaler	Scaling + Oversampling	Scaling + Oversampling
Accuracy	0.85	0.85	0.73	0.72
Precision	0.54	0.59	0.34	0.31
Recall	0.16	0.12	0.74	0.60
F1-score	0.25	0.20	0.46	0.40

Learned

- Look for good balance of precision and recall -not only accuracy

Next steps

- Try different models
- Make solution accesible