

Diabetes risk classifier

Jordi Hako Nintcheu · 2020-08-05

A KNN classifier predicting diabetes risk from basic health measurements.

Problem

Predict whether a patient is likely diabetic from a small set of routine health measurements (glucose, blood pressure, BMI, age, and a few others), using a well-known clinical dataset.

Approach

Cleaned the dataset first: several fields (glucose, blood pressure, skin thickness, insulin, BMI) had zero values recorded where zero isn't physiologically possible, those get treated as missing and filled with the column mean or median rather than left in place to quietly distort the model. Standardized the features, then tuned a k-nearest-neighbors classifier across a range of neighbor counts using cross-validation rather than picking one arbitrarily.

Result

Best model settled on $k=11$, with 76.6% test accuracy and a ROC AUC of 0.82. Precision and recall are weaker on the positive (diabetic) class, which is a known property of this dataset given its class imbalance, not something the tuning fully overcame.

What I'd do differently

Optimize directly for recall on the positive class instead of overall accuracy, missing an actual diabetes case matters more than a false alarm in a real screening context, and raw accuracy doesn't reflect that. Would also compare KNN against a couple of other baselines (logistic regression, random forest) rather than committing to one model family from the start.

Part of an ongoing series of applied AI and engineering projects. Full write-up and live links at theinternetengineer.com.