Machine Learning Engineer Nanodegree

Capstone Proposal

Ashwinder Singh Arora
November 20th, 2019

Proposal

Domain Background

Diabetes is one of the deadliest diseases which causes an imbalance of blood sugar. As of 2019, an estimated 463 million people have diabetes worldwide.[1] This represents around 10 percent of the adult population, and according to recent studies the trend suggests that rates will continue to rise. Diabetes at least doubles a person's risk of early death. According to the International Diabetes Federation (IDF) 1 in 2 adults with diabetes are undiagnosed. In 2017 alone diabetes resulted in approximately 3.2 to 5.0 million deaths worldwide. Pre-screening of the people with diabetes can help in preventing the disease from worsening and will save many lives.

The goal of this project is to be able to predict if a person has diabetes or not using the medical attributes provided in the UCI PIMA Indian Diabetes Database[2].

Problem Statement

The problem to be solved here is to predict if a person has diabetes with accuracy. Although the research carried in this project will not be a replacement for a lab test, it can surely help in pre-screening people with diabetes. The problem can be solved with the help of supervised learning algorithms to classify the population into two classes- positive and negative.

Dataset and Inputs

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from

a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
The dataset is released in the public domain. There are 768 entries in the dataset of which 500 are negative (non-diabeteic) and 268 are positive (diabetic). The predictor variables include the number of pregnancies the patient has had, their glucose level, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age. The outcome variable is the target class variable. The data needs to be processed before applying the learning algorithm.

Solution Statements

The proposed solution to this problem is to apply supervised learning techniques that have been known to perform well with binary classification problems. The data in its raw form is not suitable for accurate predictions and it requires some cleaning up. There are a few missing values replaced by zeros. The data will be tested with several algorithms to test which performs the best. An ideal solution would correctly predict all of the 500 negatives and 268 positives. Simply guessing all the patients to be diabetic would result in accuracy of 34.89%

Benchmark Model

The benchmark model[2] used the same dataset, applies preprocessing to it to obtain the comparative performance of classification algorithms. Experiments are performed using internal cross-validation 10-folds. The following table illustrates the results of the benchmark model.

	Precision	Recall	F1-Score	Accuracy %	ROC
Naive Bayes	0.759	0.763	0.760	76.30	0.819
SVM	0.424	0.651	0.513	65.10	0.500
Decision Trees	0.735	0.738	0.736	73.82	0.751

Evaluation Metrics

The evaluation metric for this problem would be the Accuracy Score as it determines the accuracy of the algorithm in predicting instances, and the F1-Score as we want the model to have a high recall while still being accurate.

Project Design

The project will be implemented in a iPython notebook running Python 3.7 and the numpy, pandas, scikit-learn and matplotlib libraries will be used. The data is imported from a csv file, following which the shape and size of the data will be

checked. Various statistical measures such as the mean, median, quartiles, minimum and maximum values will be evaluated. The missing values (if any) will then be addressed. There features will then be plotted and visualised to check for correlation between them. Scaling of data will be done as appropriate and the learning algorithms will then be applied.

Now that the data is processed, it will be divided into training and testing sets. Cross-validation will be performed and the models will be trained. We will use Random Forest, Logistic Regression, and Stochastic Gradient Descent algorithms. The performance will be evaluated by comparing their accuracy, and F1-scores. Stacking may be used in case the individual classifiers fail to perform on par with the benchmark model.

References

[1] International Diabetes Federation (2019). IDF Diabetes Atlas, 9th edn. Brussels, Belgium: International Diabetes Federation.

[2] UCI Machine Learning Repository, Pima Indian Diabetes Database

[3] Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, Volume 132, 2018, Pages 1578-1585, ISSN 1877-0509