MALIS
Eelis Kostiainen, 49971
Khristen Thornburg, 49696

Project proposal
Detecting Type II Diabetes in Patients

2020-10-19
eelis-juhani.kostiainen@eurecom.fr
khristen.thornburg@eurecom.fr

Motivation

Type II diabetes usual appears in patients later in life and is based on lifestyle and genetics. In order to help patients understand the factors that influence the development and prevention of Type II diabetes, we propose to analyze a data set of females over the age of 21 of Pima Indian decent to learn how accurately we can predict the probability of the onset of diabetes.
Knowing what factors can lead to the onset of a disease is a very practical application. As we are both beginners in the field of Machine Learning, our work does not aim to be groundbreaking or revolutionary as many similar studies have already been completed. However, we hope that what we learning over the course of this project can lead to a more general understanding of machine learning and how to predict other medical uncertainties. Additionally, the learnings from this course open a possibility for creating medical applications for treating diabetes. For Eelis, as a type I diabetic, this could prove insightful.

Method

For tackling this problem, we're planning to implement binary classification using logistic regression in order to be able to predict the probability of an individual having Type II diabetes. Classes C will describe whether the patient has type II diabetes or not. The predictors in our study are: number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, the diabetes pedigree function (scores likelihood of diabetes based on family history) and age. We plan to apply and compare the results of multiple approaches to see which works best with our data. We believe that some good options include the k-nearest neighbor analysis, Linear Discriminant Analysis, logistic regression, gradient boosting machine and perhaps more complex models like decision trees and neural nets. Based on a few articles that we referenced, many of these methods were prominent and produced good results. [1] [2]

Intended experiments and evaluation

Once we have build a machine learning algorithm using the training data from the females over 21 years of age of Pima Indian decent, we would like to test our algorithm against different demographics (eg, females of German decent [3]). In order to evaluate our machine learning algorithm, we will aim to be able to predict if a female over 21 of Pima Indian decent will develop Type II diabetes with 80% accuracy. Our data-set is found on kaggle. [4]

[1] https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6
[2] https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full
[3] https://www.kaggle.com/johndasilva/diabetes
[4] https://www.kaggle.com/uciml/pima-indians-diabetes-database