Diabetes Project Report

Manuel Alexis Mena Nieves

6/23/2020

1. Introduction

This project report is about creating a model prediction system for the HarvardX Data Science professional certificate program, using the Indians Pima Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

This dataset consists of eight medical predictor variables and one target variable, which shows if the patient has diabetes or not. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, skin thickness, glucose level, blood pressure and computed value called Diabetes Pedrigree Function.

The goal of the project is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset using all the tools shown throughout the courses in this series. To accomplish this an exploratory analysis was done, in order to understand the data and summarize their main characteristics with tables and visual methods. After this, a machine learning model and an ensemble model was created to predict whether or not the patients in the dataset have diabetes.

2. Getting the data

The following code will be used to download the dataset. We begin loading the libraries tidyverse, caret and data.table: