



InClass Prediction Competition

MUDAC 2018 Prelims

Getting started with Kaggle.

3 teams · 2 days to go

Competition Data

test.csv

train.csv 17.37 KB

Download

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)[My Submissions](#)[Submit Predictions](#) `kaggle competitions download -c mudac-2018-prelims`

Data Description

Here is the data description that can be found on the [UCI website](#), which also hosts this data set.

1. Title: Pima Indians Diabetes Database

2. Sources:

(a) Original owners: National Institute of Diabetes and Digestive and Kidney Diseases

(b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)

Research Center, RMI Group Leader

Applied Physics Laboratory

The Johns Hopkins University

Johns Hopkins Road

Laurel, MD 20707

(301) 953-6231

(c) Date received: 9 May 1990

3. Past Usage:

1. Smith,~J.-W., Everhart,~J.-E., Dickson,~W.-C., Knowler,~W.-C., \& Johannes,~R.-S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *{it Proceedings of the Symposium on Computer Applications and Medical Care}* (pp. 261--265). IEEE Computer Society Press.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cutoff of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

4. Relevant Information:

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. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

5. Number of Instances: 768

6. Number of Attributes: 8 plus class

7. For Each Attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

8. Missing Attribute Values: Yes

9. Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value Number of instances

0 500

1 268

10. Brief statistical analysis:

Attribute number: Mean: Standard Deviation:

1. 3.8 3.4
2. 120.9 32.0
3. 69.1 19.4
4. 20.5 16.0
5. 79.8 115.2
6. 32.0 7.9
7. 0.5 0.3
8. 33.2 11.8