The Classification and Prediction of Pima-Indians-Diabetes.data

Group Members: ZijiaHe, ChangXu, HongyiHuang, WangzhihuiMei

1 Introduction

The data set we used was collected by the National Institute of Diabetes and Digestive and Kidney Diseases from a group of women and Pima Indian descendants at least 21 years old. The goal is to use patient information to predict whether a patient has diabetes. There are 8 features:
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. 2-Hour serum insulin (mu U/ml)
5. Triceps skin fold thickness (mm)
6. Body mass index (weight in kg/$(heightinm)^2$)
7. Diabetes pedigree function
8. Age (years)

And 1 label (response variable). These data are collected from actual patients and represent a task, usually performed by a human doctor, with the purpose of identifying the patients most likely to have diabetes in order to propose preventive measures.
Next, we conducted some comparative studies on these data, we can get a table:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

There are some problems from tables above: The lowest blood glucose, blood pressure, skin thickness, insulin, BMI are all 0. This seems suspicious because these physical quantities cannot be 0 (for living people). Therefore, this has told us that we need to estimate these five columns. The scope of other variables seems to be reasonable.
However,there are still some informationsfrom the dataset: Higher blood sugar correlates with a result of 1, which means that the patient has diabetes. Age also seems to be related to diabetes: younger patients have a lower risk of diabetes.

Appendix

A web portal for online purchase services

Group Members (5): ZijiaHe, ChangXu, HongyiHuang, ZhanpingZhou

Our division of the work:
HongyiHuang - Goods Services
ChangXu - Shopping cart
ZijiaHe - Order Services
ZhanpingZhou - Member Services