MCMASTER UNIVERSITY

FINAL YEAR PROJECT FOR DATASCIENCE 780

Predicting Diabetes in Pima Indian Women

by Nik Počuča

Instructor:
Dr. Sharon McNicholas

November 20, 2018

1 Introduction

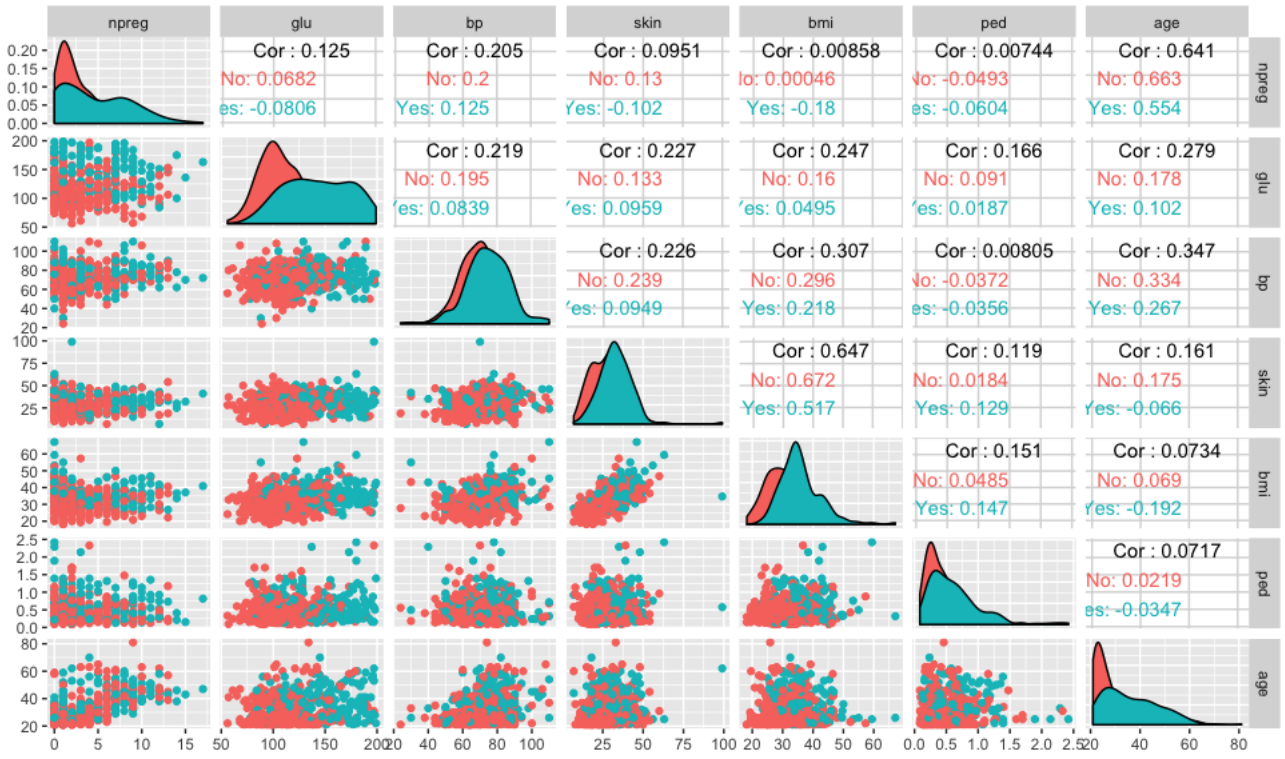
Diabetes is an ongoing chronic illness that causes the body to have an inability to process glucose(sugar) by restricting or eliminating the kidney's ability to produce insulin. By 2025 the estimated prevalence of diabetes in Canada will increase to 5 million or 12.5 of Canadians everywhere, not only putting Canadian lives at risk of premature death but cost alone is estimated to increase by 25 by 2025 (Chubbs, 2017). Therefore the need to accurately predict diabetes is growing with data analysis at the forefront. The Pima Indians dataset contains a population of women who were at least 21 years old of Pima Indian heritage and living near Phoenix, Arizona. The population was tested for diabetes according to World Health Organization criteria and data were collected by the US National Institute of Diabetes(Smith et al., 1988). This report contains a thorough set of analyses for the purposes of classifying diabetes based on covariates. First a look at each covariate and description of measurements is provided. Secondly, model methodology and description of model use is discussed. Furthermore, results from both training models and criteria for optimal model selection is provided. Finally, conclusions regarding the feasibility of the utilization of these models are discussed.

2 Pima Dataset

The Pima dataset (Diabetes of Pima Indian Women) taken from the MASS package (Venables and Ripley, 2002) in the programming langauge R (R Core Team, 2018) contains 532 complete records after dropping the (mainly missing) data on serum insulin. The dataset contains 7 covariates decribed in Table 2.

Table 1: Description of covariates in Pima dataset with summary statistics.

Covariate	Description	μ	σ	max	min
$npreg$	Number of pregnancies	1	1	1	1



References

Chubbs, D. O., 2017. Primary Care Provider Adherence to the Canadian Diabetes Association Clinical Practice Guideline for Chronic Kidney Disease. ScholarWorks@UMass Amherst.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org/`

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: Proceedings of the Annual Symposium on Computer Application in Medical Care. pp. 261–265.

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S, 4th Edition. Springer, New York, iSBN 0-387-95457-0.
URL `http://www.stats.ox.ac.uk/pub/MASS4`