Predicting Diabetes in Pima Indian Women

Nik Počuča

McMaster University

Data Science 780

November 26, 2018

Overview

1. Introduction
2. Methodology
3. Results

Future of Diabetes in Canada

What is diabetes?

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.

Rate of Diabetes

By 2025 the estimated prevalence of diabetes in Canada will increase to 5 million individuals.

Cost of Diabetes

Cost alone is said to increase by 25% putting a strain on the system.

Introduction to the Pima Indian dataset

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 (Smith et al.1988).

- 532 complete records
- The population was tested for diabetes according to World Health Organization criteria.
- The dataset contains 7 covariates and 1 response variable

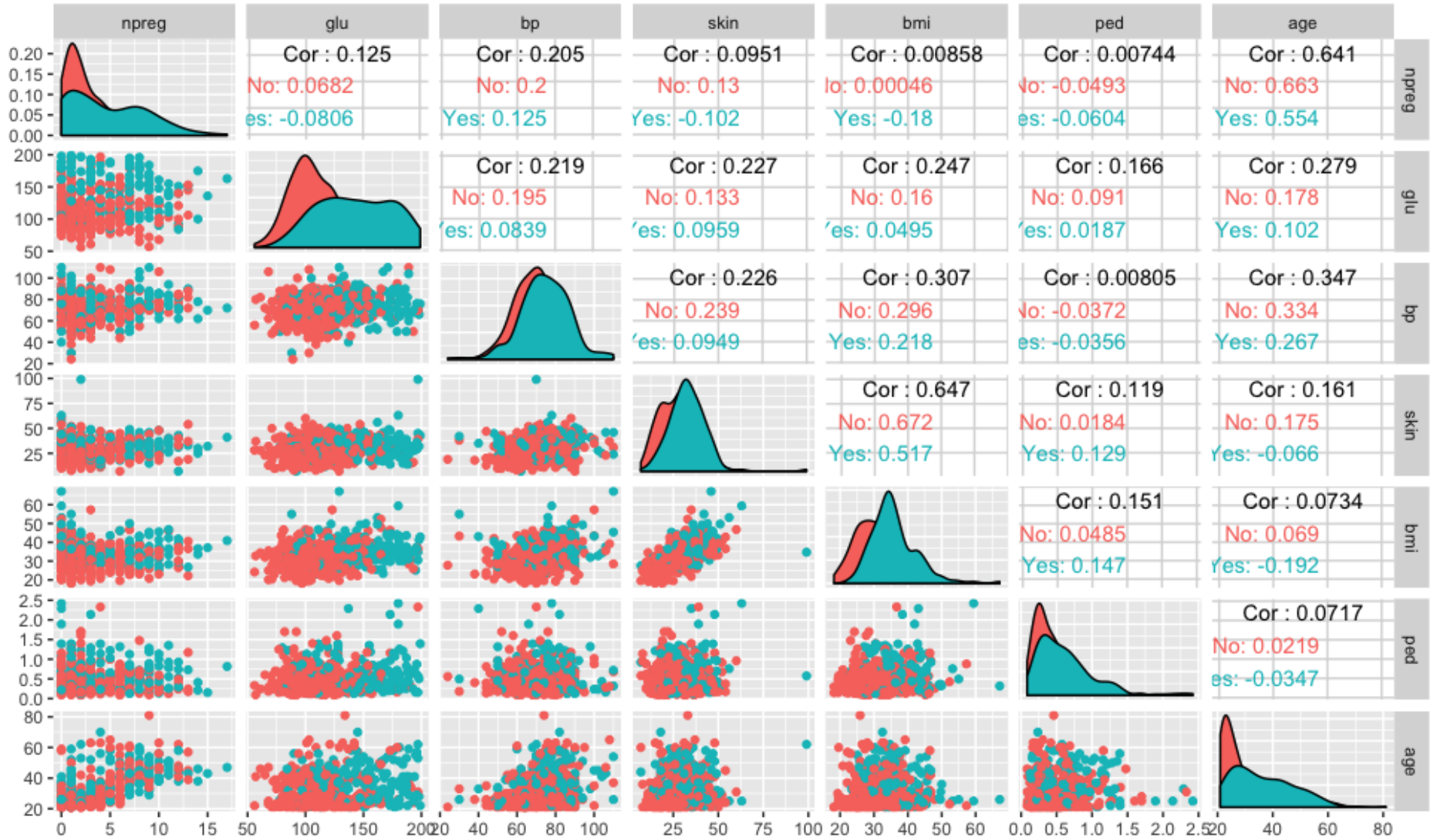
Description of covariates and response

Variable	Description	μ	σ	min	max
npreg	number of pregnancies (integer)	3.52	3.31	0.00	17.00
glu	plasma glucose concentration. (integer)[a]	121.03	30.97	56.00	199.00
bp	diastolic blood pressure (mm Hg).	71.51	12.30	24.00	110.00
skin	triceps skin fold thickness (mm).	29.18	10.51	7.00	99.00
bmi	body mass index (weight in kg/height in m^2).	32.89	6.87	18.20	67.10
ped	diabetes pedigree function [b]	0.50	0.34	0.09	2.42
age	age in years. (integer)	31.61	10.75	21.00	81.00
type	Yes or No (binary response)	-	-	-	-

[a] count of quantity of plasma after 2 hours

[b] diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient

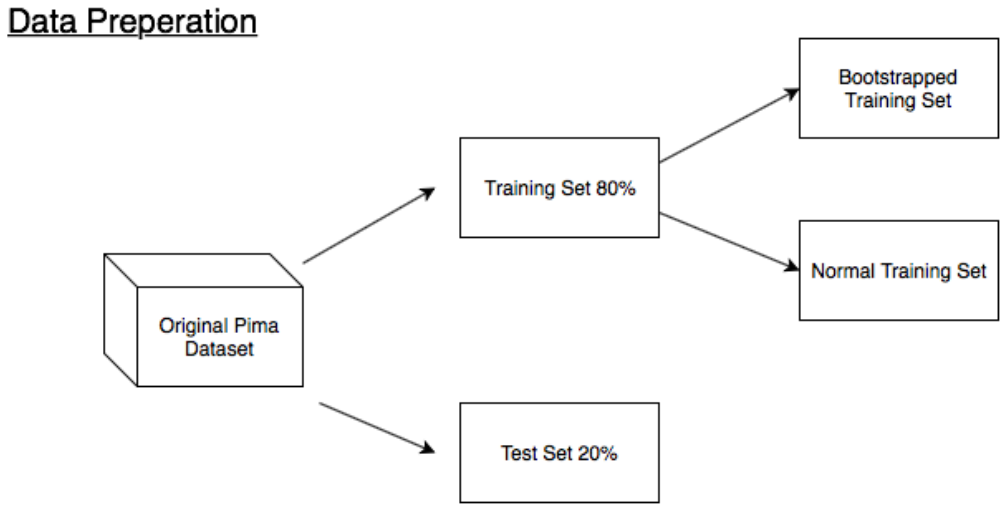
Pair plot of covariates

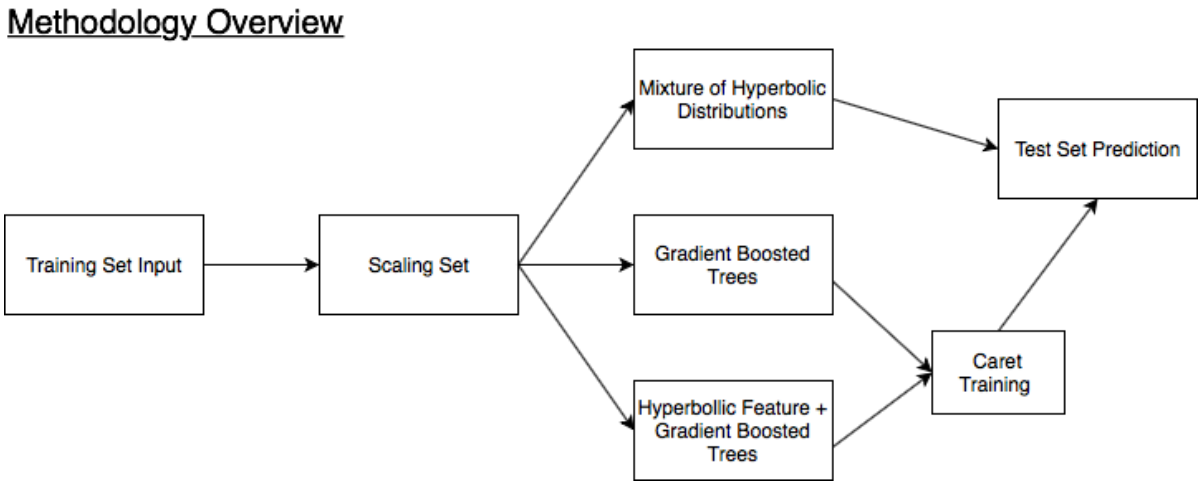


Data Preperation

- Due to differences in max and min, all of the covariates were scaled.
- An 80/20 - training/test split was constructed at random.
- A second training set was created by bootstrapping the first to have balanced diabetic and non-diabetic individuals.

Methodological Overview





Mixtures of Generalized Hyperbolic Distributions (McNicholas P.D. 2015)

Consider a random vector $\boldsymbol{X}$ that emanates from a finite mixture model with $\boldsymbol{x} \in \boldsymbol{X}$ with $\boldsymbol{x}$ as a realization of $\boldsymbol{X}$. Furthermore assume that for the sample space Ω, where $X \in \Omega$ one can partition Ω into G subgroups $\Omega = \{\Omega_1, \ldots, \Omega_G\}$. Stemming from this the joint distribution of $\boldsymbol{X}$ is written as

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f(\boldsymbol{x}|\boldsymbol{\theta}_g),$$

Mixtures of Generalized Hyperbolic Distributions

Given an assumption that $\boldsymbol{X}$ is of a generalized hyperbolic distribution (Tortora et. al 2014), then $f(\boldsymbol{x}|\boldsymbol{\theta}_g)$ is formulated as

$$f(\boldsymbol{x}|\boldsymbol{\theta}_g) = \left[\frac{\omega_g + \delta(\boldsymbol{x}, \boldsymbol{\mu_g}|\boldsymbol{\Sigma}_g)}{\omega_g + \alpha_g' \boldsymbol{\Sigma_g^{-1}} \alpha_g}\right]^{(\lambda - p/2)/2} \frac{K_{\lambda - p/2}\left(\sqrt{(\omega_g + \alpha_g' \boldsymbol{\Sigma_g^{-1}} \alpha_g)(\omega_g + \delta(\boldsymbol{x}, \boldsymbol{\mu_g}|\boldsymbol{\Sigma}_g))}\right)}{(2\pi)^{p/2}|\Sigma|^{1/2} K_\lambda(\omega_g) \exp\{-(\boldsymbol{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma_g^{-1}} \boldsymbol{\alpha}_g\}}.$$

Mixture Discriminant Analysis

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{k} \sum_{g=1}^{G} \left[\pi_g f(\boldsymbol{x}|\boldsymbol{\theta}_g)\right]^{z_{ig}} \times \prod_{j=k+1}^{n} \sum_{h=1}^{H} \left[\pi_h \phi(\boldsymbol{x}_j|\boldsymbol{\theta}_h)\right].$$

$$z_{jh} = \arg\max_{h \in H} \tau_{ih}, \quad \tau_{ih} = \frac{\hat{\pi}_{ih} f(\boldsymbol{x}|\boldsymbol{\theta}_h)}{\sum_{h=1}^{H} \hat{\pi}_{ih} f(\boldsymbol{x}|\boldsymbol{\theta}_h)}.$$

Gradient Boosted Trees

Given a set of data with n observations and m covariates $(\boldsymbol{x_i})$ on some set $D = \{(x_i, y_i)\}$. A tree ensemble uses K additive functions to predict the output of (y_i) written as follows

$$\hat{y}_i = \sum_{k=1}^{K} \lambda f_k(\boldsymbol{x}_i),$$

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_k(\boldsymbol{x}_i)) + \phi(f_k)$$

Evaluation Methodology

$$\mathbf{BIC} = \log(n)k - 2\log(\hat{L}),$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

Training Results

Table: Model performance results with ARI and class error over all six variations.

Method	ARI	Class Error
DA Hyperbolic	0.1974	0.2710
DA Hyperbolic + Boostrap	0.2203	0.2617
Boosted Trees	0.3532	0.1963
Boosted Trees + Boostrap	0.3532	0.1963
Boosted Trees + Hyperbolic Feature	**0.3546**	**0.1963**
Boosted Trees + Hyperbolic Feature + Bootstrap	0.3546	0.1963

Boosted Tree with Hyperbolic Features

Table: Classification table for mixture of hyperbolic distributions (left) and gradient boosted trees (right), vertical (0,1) is positive for diabetes and horizontal is mixture class label.

Classes	1	2	3	4	5
0	8	93	76	34	77
1	17	20	54	31	15

True Values	Predicted 0	1
0	64	3
1	17	23

Table: Relative influence of each covariate on boosted tree prediction in percetages.

Variable	glu	age	bmi	ped	npreg	skin	hyper	bp
Relative Influence (%)	51.480	18.203	13.484	9.989	5.107	1.253	0.484	0.000

Conclusions

Limitations

Gradient boosted trees are limited in their predictive power when covariates do not necessarily predict response.

Improvements

Slight improvements can be made to model accuracy by using mixture models to generate sub groups for prediction.

More Covariates

This gradient boosted trees are not sufficient in predicting diabetes accurately and other types of data should be collected to improve predictive power.

References

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. CoRR abs/1603.02754.

Scalable Tree Boosting

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

Clinical Practice Guideline for Kidney Disease

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society B 39, 1–38.

Expectation Maximization Algorithm

Max Kuhn, J. W., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt., T., 2018. caret: Classification and Regression Training. R package version 6.0-80.

Caret Package

McNicholas, P. D., 2015. Mixture Model-Based Classification. Chapman and Hall.

Mixture Model-Based

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.

Pima Indian Dataset

Tortora, C., Franczak, B. C., Browne, R. P., McNicholas, P. D., Mar. 2014. A Mixture of Coalesced Generalized Hyperbolic Distributions. ArXiv e-prints.

Mixture of Coalesced Generalized Hyperbolic

The End