Diabetes Prediction in the Pima Indian Population

MATH-493 Applied Biostatistics

Zacharie Legault

May 28, 2019

1 Introduction

The task of classification is central to the modern practice of statistics. Given a set of explanatory variables, we want to model the data such that we can both gain new insights into the underlying mechanisms and make useful predictions given previously unseen data.

1.1 Dataset Presentation

The dataset used in this study is the well known Pima Indians Diabetes Dataset, which is often used as an introduction to statistical and machine learning classification methods. This data was originally presented by Smith et al. (1988) to investigate the use of neural networks for this task [1]. The dataset consists of 8 explanatory variables and the diagnosis for 768 patients from the Pima Indian population. This group is known for a high prevalence of diabetes (Bennett et al., 1971) and as such has been followed for many years. The observed variables are the following:

- Number of pregnancies
- Plasma glucose concentration 2 hours after a glucose tolerance test (mg/dL)
- Diastolic blood pressure (mmHg)
- Triceps skin fold thickness (mm)
- Insulin concentration 2 hours after a glucose tolerance test (µU/mL)
- Body mass index (BMI; kg/m^2)
- Diabetes pedigree function (DPF): this indicator was developped by Smith et al. in order to give a general idea of the history of diabetes in the patient's family [2]. The forumla is presented in Appendix A.
- Age (years)

The authors defined a positive diabetes diagnosis as a plasma glucose concentration of 200 mg/dL two hours after ingesting 75 g of a carbohydrate solution. Only the subjects where the diagnosis was made between one and five years after the initial examination were kept in the dataset, and labeled as a positive outcome. If diabetes still had not been diagnosed after five years, the subjects were labeled as a negative outcome. Out of the 768 subjects, 268 were given a positive diagnosis and 500 were given a negative diagnosis. All patients were female, and at least 21 years old.

[1] For the current study, data was obtained as a CSV file through Kaggle (UCI Machine Learning, 2016).

[2] Note that at the time of publication of the original paper, the function had not been validated and still does not seem to have been.

1.1.1 Data Preprocessing: Incomplete Data and Standardization

One important aspect of this dataset is that it is not complete, as some entries are impossible. There are respectively 11, 5, 227, 374 and 35 subjects where the recorded BMI, blood glucose concentration, triceps skin fold thickness, insulin concentration and blood pressure is 0, which is obviously wrong. The histograms of the raw data are presented in Figure 1. Subjects where 2 or more of these variables were missing are removed from the dataset, which leaves 534 subjects (177 of whom with a positive diagnosis and 357 with a negative diagnosis). However, 26.2% of the subjects still have missing data on insulin concentration, whereas the proportion is less than 1% for the other variables. Any imputation technique to fill in the missing values for insulin concentration would introduce significant bias in the analysis when such a high proportion the data is absent. As such, the variable is removed altogether from the dataset. Mean substitution is performed to impute the missing items in BMI, blood glucose concentration, triceps skin fold thickness, and blood pressure. The histograms of the processed data are presented in Figure 2. Finally, in order to limit distortions due to the different scales of the features, each of them is stardardized (by substracting the mean and dividing by the standard deviation) such that they all have a mean of 0 and a standard deviation of 1. From now on, this preprocessed dataset will be referred to simply as *the dataset*.

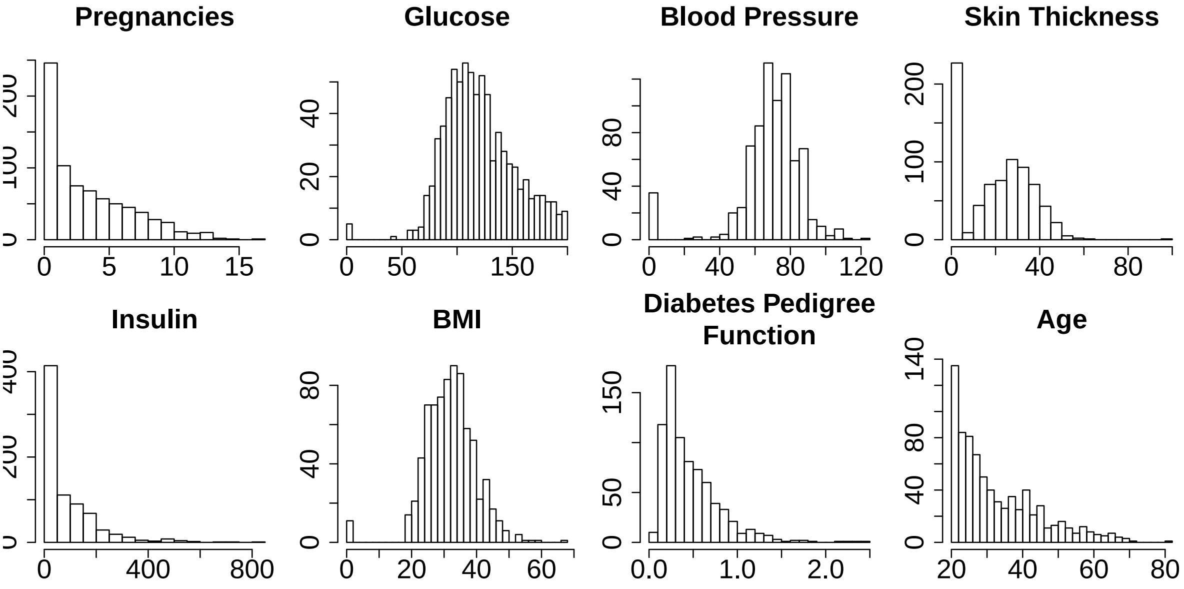

Figure 1: Histograms of raw explanatory variables, with missing data

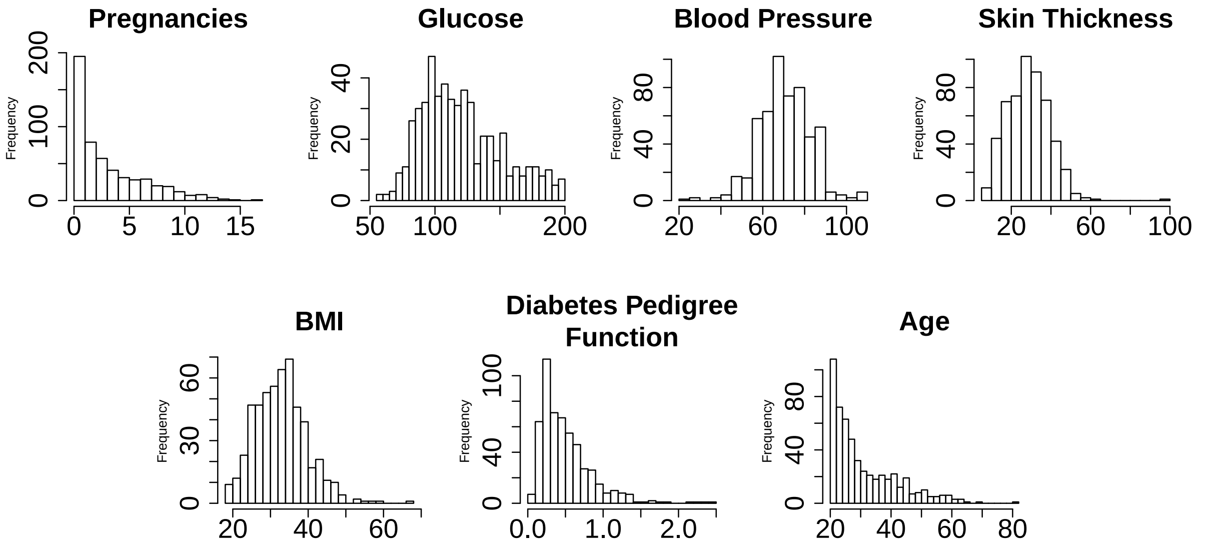

Figure 2: Histograms of selected explanatory variables, with imputed data

1.1.2 Exploratory Data Analysis

As a first step to explore the dataset, a scatter matrix (Figure 4) and a correlation plot (Figure 3) are prepared to vizualise any clear relationships between pairs of variables. In general there does not seem to be any clear relationship between any of them. There is however a moderate correlation berween BMI and skin thickness ($r \approx 0.648$), and age and the number of pregnancies of the subjects ($r \approx 0.642$). In both case, these are relationships we can expect. Since no pairs of variables has a disproportionate correlation, none of them have to be removed in order to avoid unwanted multicollinearity.

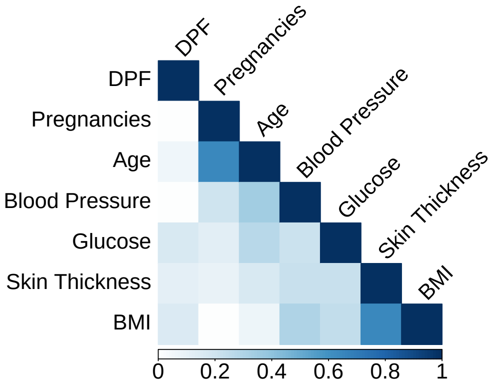

Figure 3: Correlation plot of the dataset. The strength of the correlation between each pair of variable is color coded with the scale on the right.

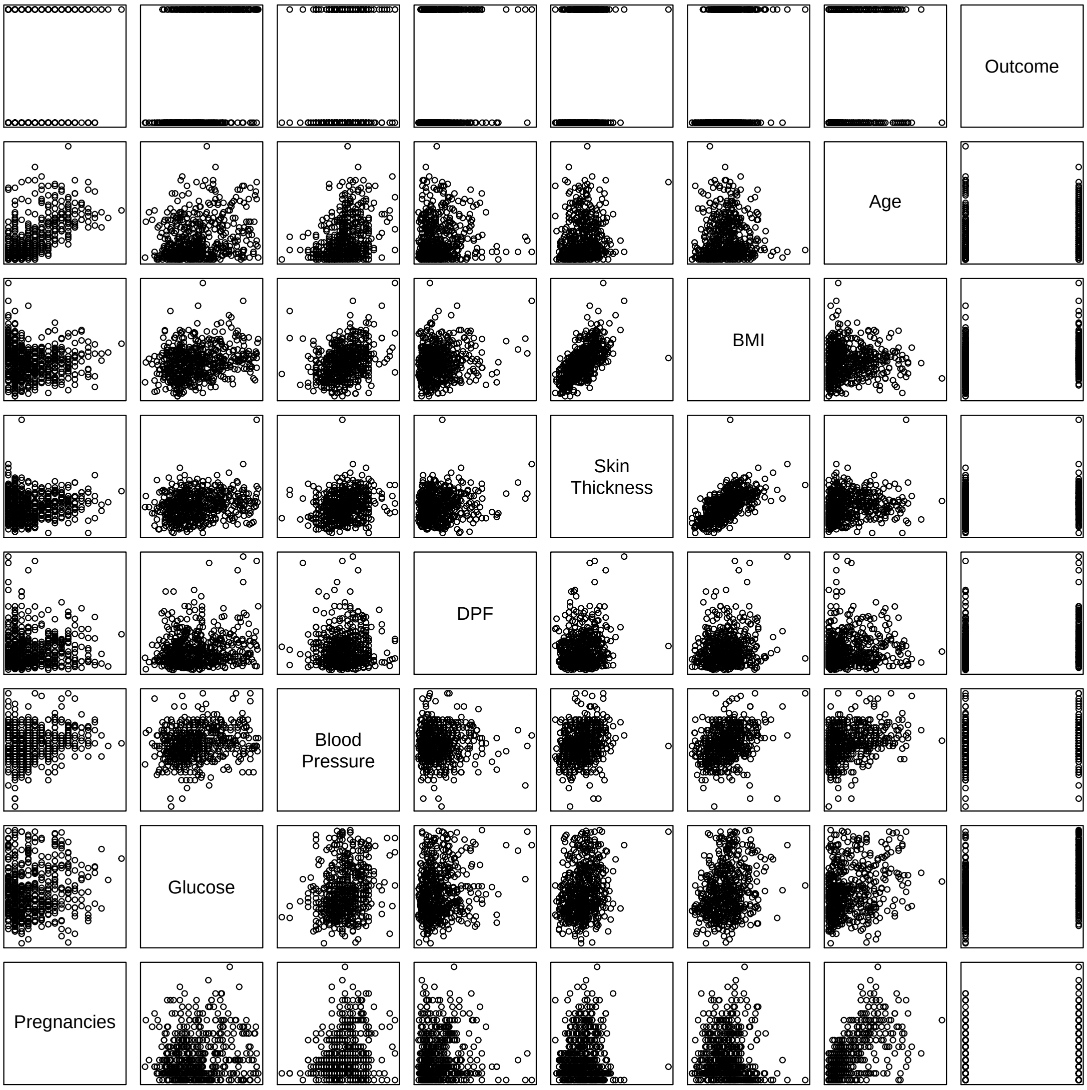

Figure 4: Scatter matrix of the dataset

1.2 Logistic Regression

Logistic regression is one of the most common methods for binary classification given a set of predictors. It is an instance of a generalized linear model (GLM) with the *logit* link function. The model will take as input the explatanory variables x_k and give a prediction p that the outcome is positive (i.e. 1), with parameters β_k. The logit transformation allows us to constrain the value of p to the interval $[0, 1]$.

$$\operatorname{logit}(p(\boldsymbol{x};\boldsymbol{\beta})) = \ln\left(\frac{p(\boldsymbol{x};\boldsymbol{\beta})}{1-p(\boldsymbol{x};\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m = \boldsymbol{\beta}^\top \boldsymbol{x} \tag{1}$$

$$p(\boldsymbol{x};\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{\beta}^\top \boldsymbol{x})} \tag{2}$$

The parameter β_k represents the log-odds of feature x_k, telling us how much the logarithm of the odds of a positive outcome (i.e. the logit transform) increases when predictor x_k increases by 1. The odds are obtained by taking the exponential of these parameters.

Fitting is generaly done by maximum likelihood estimation. The likelihood of the model is given by

$$\ell(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}; \boldsymbol{\beta}) = \prod_{i=1}^{n} p(\boldsymbol{x}^{(i)})^{y^{(i)}} (1 - p(\boldsymbol{x}^{(i)}))^{1-y^{(i)}} \tag{3}$$

where $y^{(i)}$ is the outcome of subject i. Maximizing the likelihood is equivalent to maximizing the log-likelihood of the model.

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}; \boldsymbol{\beta}) &= \log(\ell(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}; \boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} \left[y^{(i)} \log(p(\boldsymbol{x}^{(i)})) + (1 - y^{(i)}) \log(1 - p(\boldsymbol{x}^{(i)})) \right] \end{aligned} \tag{4}$$

Because the derivative of Equation 4 is non-linear for logistic regression, its minimization is generally done numerically by iteratively re-weighted least-squares (IRLS).

One major advantage compared to other classification methods is that a logistic regression model is easily interpretable and linear in the paramters, each parameter β_k in the linear combination telling us the log-odds of each predictor.

k-fold Cross-Validation. A commonly used technique in statistics and machine learning to assess the performance of a given model on new data is k-fold cross-validation. The dataset is partitioned into k random subsets – or folds – assumed to be representative of the whole dataset, and k models are fitted, each of them with one of the subsets held out. The models are then evaluated on the subset they have not yet seen. A drop in performance on new data would indicate that the model is overfitting its training data.

2 Results

A simple model is first fitted using the whole dataset, with all variables combined without any interactions: pregnancies, glucose concentration, blood pressure, skin thickness, BMI, DPF, and age are all assigned a single coefficient, with an additionnal intercept parameter.

The resulting model is then used as baseline for a stepwise selection phase. The `stepAIC` from the `MASS` package is used, which utilizes the Akaike Information Criterion (AIC) to find the best model given a baseline model and a parameter scope. A scope of all second degree interactions (i.e. squaring any single variable or pairwise multiplicative interaction between variables) is provided to the function. The procedure drops blood pressure and skin thickness as predictors, and introduces cross-interactions between glucose concentration and DPF, and between pregnancies and age. Since the models are not nested, we cannot do a likelihhod ratio test, but we can see that the new model reduces the deviance from 468 to 460 (bothe with a null deviance of 678). The selected model also increases McFadden's pseudo-R^2 from 0.311 to 0.323 which indicates a better fit [3]. Because of the improved fit of the stepwise-selected model, this is the one that is chosen to go forward.

[3] McFadden's pseudo-R^2 is given by $R^2_{\text{McFadden}} = 1 - \mathcal{L}/\mathcal{L}_0$ where $\mathcal{L}$ and $\mathcal{L}_0$ are respectively the log-likelihood of the fitted and null models (i.e. only an intercept term).

The coefficient estimates and their standard error, 95% confidence interval and p-value for the selected model are presented in Table 1, and odds ratios are in Figure 5 as a forest plot. In all the significant predictors, blood glucose concentration is by far the most important one, with the odds increased by 3.09 times for every increase of one standard deviation (30.9 mg/dL). The number of pregnancies, BMI and DPF follow, all with a similar risk increase (odds ratios of 1.77, 1.87, and 1.65 respectively). There is also a small but significant risk reduction when considering the cross-interaction between glucose concentration and DPF (0.760). This is surprising given that each factor individually increases the probability of developping diabetes, but together they decrease it.

Table 1: Coefficients of the logistic regression model. The p-values below the 5% significance level are highlighted in blue.

Coefficient	Estimate	Standard Error	Confidence interval (95%)	p-value
β_0	-0.910	0.144	$-1.20/-0.633$	2.40×10^{-10}
$\beta_{\text{pregnancies}}$	0.568	0.167	0.246/0.901	6.47×10^{-4}
β_{glucose}	1.13	0.135	0.873/1.40	5.59×10^{-17}
β_{BMI}	0.624	0.127	0.381/0.881	9.42×10^{-7}
β_{DPF}	0.500	0.117	0.271/0.733	2.10×10^{-5}
β_{age}	0.288	0.148	$-0.001\,24/0.582$	0.0518
$\beta_{\text{glucose-DPF}}$	-0.274	0.104	$-0.470/-0.0556$	8.27×10^{-3}
$\beta_{\text{pregnancies-age}}$	-0.204	0.126	$-0.457/0.0412$	0.106

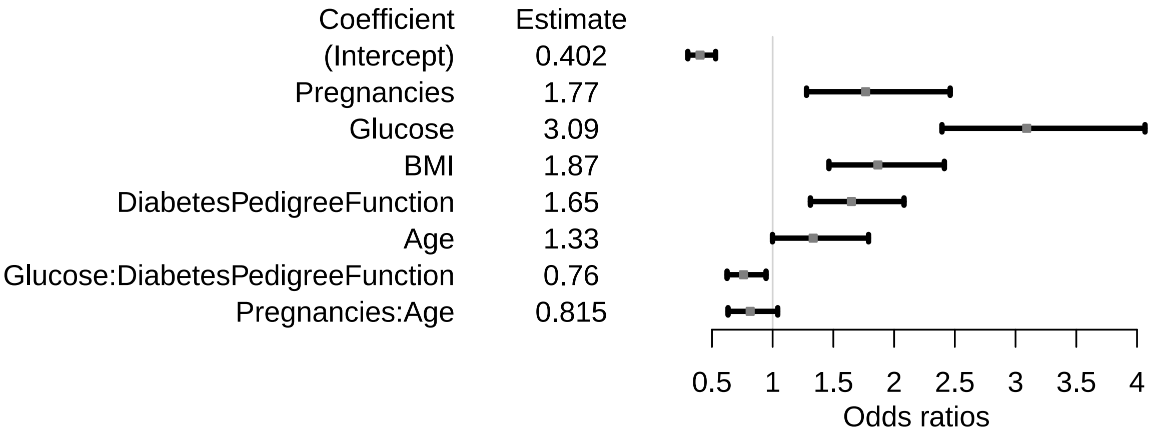

Figure 5: Forest plot if the odds ratios

To check for any multicolinearity between predictors, their variance inflation factor (VIF) are calculated (see Table 2). As expected, the highest values are those concerning age and the number of pregnancies. As mentioned previously, there is a noticeable correlation between those two variables, so a higher VIF is normal. All other VIF values are small and indicate little multicolinearity between the selected features.

Table 2: Variance inflation factors of the predictors

Predictor	Variance inflation factor
Pregnancies	2.19
Glucose concentration	1.11
BMI	1.06
DPF	1.12
Age	1.72
Glucose-DPF	1.22
Pregnancies-Age	1.78

To validate that the model is not overfitting, a 10-fold cross-validation is run using the same model architecture. The mean and standard deviation of the parameter estimates over the 10 folds is presented in Table 3. The mean values are nearly indentical as those obtained with the whole dataset, with very little variability. The mean validation accuracy (on unseen data, 0.792 ± 0.0646) is very slightly lower than the training accuracy (on the fitted data, 0.798 ± 0.006 53), indicating little to no overfitting.

Table 3: Mean and standard deviation of the parameter estimates for a 10-fold cross-validation

Coefficient	Mean estimate	Standard deviation
β_0	-0.912	0.0586
$\beta_{\text{pregnancies}}$	0.568	0.0538
β_{glucose}	1.13	0.0475
β_{BMI}	0.626	0.0369
β_{DPF}	0.502	0.0374
β_{age}	0.289	0.0469
$\beta_{\text{glucose-DPF}}$	-0.271	0.0564
$\beta_{\text{pregnancies-age}}$	-0.204	0.0532

3 Discussion

The dataset explored in the current classification task is the Pima Indians Diabetes dataset which documents the diagnosis of diabetes in Pima Indian women. Logistic regression is used to model the standardized data, with the number of pregnancies, blood glucose concentration, blood pressure, tricep skin fold thickness, BMI, DPF and age as predictors. After fitting a baseline model including linear contributions of all of these features, stepwise selection yields a model that drops blood pressure and skin thickness as predictors, and introduces cross-interactions between glucose concentration and DPF, and between pregnancies and age.

Since the diagnostic threshold for a positive was set as a glucose concentration of 200 mg/dL by Smith et al. (1988) when creating the dataset, it is not surprising that the predictor with the highest odds ratio is the glucose concentration at the initial examination of the subjects. This feature is by far the most important risk factor; we have to go to the lower bound of the confidence interval (2.39) and to the upper bound for the following features (2.46 for pregnancies, 2.41) for BMI) to have comparable values. Having more pregnancies, a higher BMI and a higher DPF score all significantly increase the risk of receiving a positive diabetes diagnosis. Surprisingly, an combined increase in both glucose concentration and DPF results in a small but significant risk decrease. Being older increases the odds, while the combination of age and pregnancies decreases them, but the effect of these predictors is not statistically significant.

There is little multicolinearity between the predictors as indicated by varianc inflation factors close to 1, with perhaps the exception of age and the number of pregnancies. These variables are somewhat correlated, which is expected as older women have had more time to have children. Since all VIF values smaller than 2.19, multicolinearity can be neglected.

Finally, the select model is stable as it converges to similar parameter values even when fitted to a subset of all the data. It's predictive power is also virtually as good for new datapoints as for the ones used for the fitting procedure (0.792 ± 0.0646 validation accuracy and 0.798 ± 0.006 53 training accuracy over a 10-fold cross-validation), indicating little to no overfitting.

4 Conclusion

Classification is a very common task in modern statistics and machine learning, with various approaches developped over the years to tackle this problem. Logistic regression on a set of features is one of the most basic of these techniques, but its simplicity and interpretability make it still relevant today. Logisitic regression was applied in this study to model the onset of diabetes in Pima Indian women. The best performing model was selected by a stepwse selection procedure, with the selected predictors being the number of pregnancies, blood glucose concentration, BMI, DPF and age as well as the combined effect of glucose and DPF, and pregnancies and age. Blood glucose concentration is by far the most important predictor. No major multicolinearity was observed, and the model has virtually the same predictive accuracy on new data compared to its accuracy on its training data.

Further exploration of different modeling techniques such as tree-based classification, support vector machines, and neural networks, would be the next step. Given its simplicity and interpretability, logistic regression could be used as a proper baseline to compare the performance of these more complex methods.

References

Peter H. Bennett, Thomas A. Burch, and Max Miller. Diabetes Mellitus in American (Pima) Indians. *The Lancet*, 298(7716):125–128, 1971.

Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

UCI Machine Learning. Pima Indians Diabetes Database, October 2016. URL `https://www.kaggle.com/uciml/pima-indians-diabetes-database`. Version 1.

Appendix A Diabetes Pedigree Function

The diabetes pedigree function proposed by Smith et al. (1988) is computed as follows:

$$\text{DPF} = \frac{\sum_{d \in D} K_d(88 - ADM_d) + 20}{\sum_{n \in N} K_n(ALC_n - 15) + 50} \tag{5}$$

where

- D is the set of the subject's relatives who have a *positive* diabetes diagnosis at the time of the initial examination.
- N is the set of the subject's relatives who have a *negative* diabetes diagnosis at the time of the initial examination.
- K_x is the proportion of genes shared between relative x and the subject. A parent or sibling's K_x will be 0.5; a half-sibling, grand-parent, aunt, or uncle will have a K_x of 0.25; a half-aunt, half-uncle, or first cousin will have a K_x of 0.125.
- ADM_d is the age (in years) of relative d at the time of their diagnosis.
- ALC_n is the age (in years) of relative n at the time of their last examination.

The constants 88 and 14 are the maximum and minimum age at which the subject's relatives have been diagnosed with diabetes (with some rare exceptions). The constants 20 and 50 have been chosen, according to Smith et al., in order to (a) give a DPF value a little below average to subjects with no relatives, (b) limit distortions when young relatives (who are less likely to have a positive diagnosis) are added to the computation, and (c) rapidly increase the value when relatives receive a positive diagnosis.