Data Mining II  
Final Project  
Diabetes in Pima Indian Females

Group 3 Members:

Arasu, Vignesh

Aziz, Ali

Babar, Amir

Kumar, Tathagat

**Abstract**

Being able to predict Type II diabetes in advance can allow at-risk individuals to take preventive measures. For this study predictive models were built using the Pima Indian females population. A key challenge with this data set was dealing with a large amount of missing values. Multiple imputation and EM imputation were utilized to better understand the data. Using a dataset with EM imputation, 5 different classification models were built: a logistic regression model, a classification tree, a general additive model, a neural network and an adaptive boosting model. Based on both in-sample and out-of-sample AUCs and Misclassification Rates, the GAM model performed the best as it had an AUC of 0.88 and MR equal to 0.19 on the testing data. However, for more conclusive results further testing such as cross-validation is necessary.

**Executive summary**

The purpose of this classification study was to be able to predict the onset of Type II diabetes based on certain predictor variables using various data mining methods from supervised learning. The models were fit to a data set for 768 Pima Indian females who tested negative for diabetes but then developed the disease within 1 to 5 years. There were 8 continuous predictor variables, namely number of pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age (all measured at the time of the negative test). A major issue with the data was the number of missing values for the variables Insulin and Skin Thickness (48.7% and 29.6% respectively). Instead of using a trivial imputation method such as median imputation, we looked at 2 sophisticated imputation techniques: Expectation Maximization and Predictive Mean Matching. The latter was chosen as the final imputation method. These allowed us to make better use of the data set. After that we proceeded to fitting 5 different models: a logistic regression model, a classification tree, a generalized additive model, a neural network and an adaptive boosting model. Since the response variable was binary, the performance metrics were the Area Under the Curve (AUC) and Misclassification Rate (MR) i.e. symmetric cost. Both in-sample and out-of-sample performance were measured by randomly dividing the data into training (75%) and testing data sets (25%). The same seed was used to ensure fair comparison of models. Parameter tuning was done using training data only. For neural networks, the training data was further divided into a training-testing split to avoid overfitting. A comparison of all the models built in R is presented in the **Executive Table 1** below. As expected, generally performance on testing data was not as good as the performance on training data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **In-sample** | | **Out-of-sample** | |
| **Model** | **MR** | **AUC** | **MR** | **AUC** |
| GLM | 0.22 | 0.84 | 0.23 | 0.87 |
| Classification Tree | 0.23 | 0.88 | 0.24 | 0.84 |
| AdaBoost | 0.22 | 0.85 | 0.23 | 0.87 |
| GAM | 0.21 | 0.86 | 0.19 | 0.88 |
| Neural Network | 0.22 | 0.85 | 0.23 | 0.86 |

**Executive Table 1.**  Performance of the models in R on training and testing data

Based on the training-testing split of data used, the best model was GAM as it performed better than the other models both in-sample and out-of-sample. Neural networks were difficult to tune and offered similar results to other models. Adaptive Boosting gave only marginal improvement over the Classification Tree. Wherever possible the models were also fitted in SPSS and the results were seen to be comparable.

It should be stressed that the conclusions are based on a specific training-testing division. For more reliable results cross-validation should be used. From this study we learnt about the challenging nature of imputation and how different models give different importance to predictors even when the data was the same.

**Introduction**

Type 2 diabetes mellitus is a metabolic disorder involving impaired glucose uptake that has become increasingly prevalent worldwide, affecting close to 400 million people1. After meal consumption, insulin is released in response to increased levels of blood glucose, in turn normalizing levels of blood glucose. In type 2 diabetes, decreased responsiveness to insulin results in prolonged periods of high blood glucose. Sugars at this high concentration bind to proteins, inhibiting their function, and damage vessels, reducing blood flow and contributing to nerve damage. As a result, people with type 2 diabetes are at increased risk for heart disease, stroke, kidney failure, blindness, and limb amputation. The risk of dying among people with diabetes is at least double the risk of peers without it2,3.

Major contributing factors to the onset of type 2 diabetes include obesity and lack of exercise, although a genetic component is also involved4. The ability to predict who is likely to develop type 2 diabetes (prior to onset) is of great importance for a variety of reasons5. Reliable forecasting can enhance preventative care. At risk individuals should be educated and encouraged to take preventative measures, importantly, before the occurrence of the many terrible outcomes associated with the disease. In addition to improving patient outcomes, such measures would also help to reduce healthcare costs, an issue of ever increasing concern.

Our work involves utilizing a data set previously used by researchers with the goal of constructing reliable models to predict the onset of type 2 diabetes6,7. This data set involves individuals from the Pima Indian community, a native American community located in central and southern Arizona. This group has the highest prevalence of type 2 diabetes worldwide, in part due to a genetic component that is highly shared in the homogenous community, as well as due to lifestyle changes brought about by recent Westernization8,9,10.

**Data Description**

For the data set, women who tested negative for diabetes, as indicated via a glucose tolerance test (GTT), were selected. In a GTT, a sugary solution is ingested and blood sugar is measured after 2 hours. Impaired glucose tolerance results in high blood sugar levels even after this time period, and a level of at least 200 mg/dL results in a diagnosis of type 2 diabetes. The binary outcome of interest was whether or not the women were diagnosed with diabetes within five years of their initial appointment. Women diagnosed within one year of their initial appointment were excluded from the data set. This resulted in the selection of 768 subjects. Eight variables were utilized as predictors, based on previous identification as risk factors for type 2 diabetes. These are explained together with the response variable in **Table 1.**

|  |  |  |
| --- | --- | --- |
| Variable | Interpretation | Comment |
| Pregnancies | Number of times pregnant | Women often develop gestational diabetes during pregnancy due to increased loads. Those women are at higher risk for developing TIIDM. |
| Glucose | Plasma glucose concentration after 2 hours in an oral glucose tolerance test | Elevated in people developing TIIDM |
| Blood Pressure | Diastolic blood pressure  (mm Hg) | Believed to be associated with diabetes risk |
| Skin Thickness | Triceps skin fold thickness  (mm) | Related to obesity/body fat |
| Insulin | 2-Hour serum insulin  (mu U/ml) | Elevated in people developing TIIDM |
| BMI | Body mass index  (weight in kg/(height in m)^2) | Related to obesity/body fat |
| Diabetes Pedigree Function | Construct developed to take into account family history and genetics | Combines different risk factors for TIIDM to obtain value |
| Age | Age in year | Risk increases with Age |
| Outcome | Class variable (0 or 1) | 1 if diabetes developed within 1-5 years after initial test, 0 otherwise |

**Table 1**. Data dictionary for the variables in the dataset

**Exploratory Data Analysis**

**Missing Values Analysis**

The original dataset contains 768 complete observations. On closer inspection, however, we noted that a number of variables, such as glucose, blood pressure, skin thickness, and BMI had values of 0 in certain observations. Such observations are not physiologically possible, leading us to conclude that those values were actually missing in the dataset. After recoding the values, we examined the nature of the missing data. Five of our eight variables had some level of missingness, and roughly half of our total observations were incomplete at some level (**Figure 1**). The variables with the most missing values are insulin and skin thickness, missing in 48.7% and 29.6% of observations, respectively (**Table 2**). All other variables had less than five percent missing values.

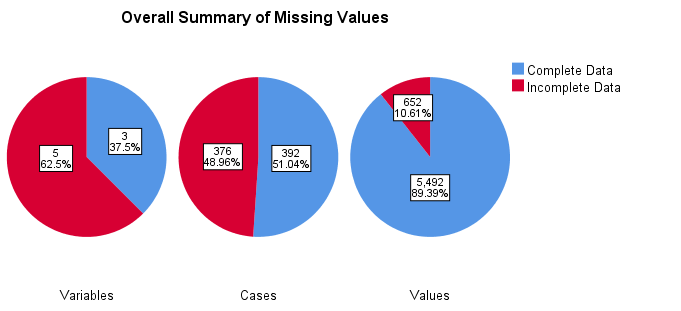


Figure 1: Summary of missing values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
|  | Missing | | Valid N | Mean | Std. Deviation |
| N | Percent |
| Insulin | 374 | 48.7% | 394 | 155.55 | 118.776 |
| SkinThickness | 227 | 29.6% | 541 | 29.15 | 10.477 |

Table 2. Variable summary. Minimum percentage of missing values for variable to be included: 5.0%

In general, missing values in a dataset can arise for a number of reasons. Three terms commonly utilized to describe the nature of missingness are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)11. With MCAR, the pattern of missingness is random, as the term random is commonly understood. With MAR, the pattern of missingness may be dependent upon other observed data. Lastly, in MNAR, the missingness depends on data that is not observed. Analyses of MCAR and MAR missingness patterns allow for unbiased parameter estimates to be obtained, while MNAR yields biased estimates.

To help determine the pattern of missingness in our dataset, a plot of the different patterns of missingness, as well as their frequency, was constructed (**Figures 2 &3**). While not appearing to be MCAR, the exact pattern of missingness was difficult to determine.

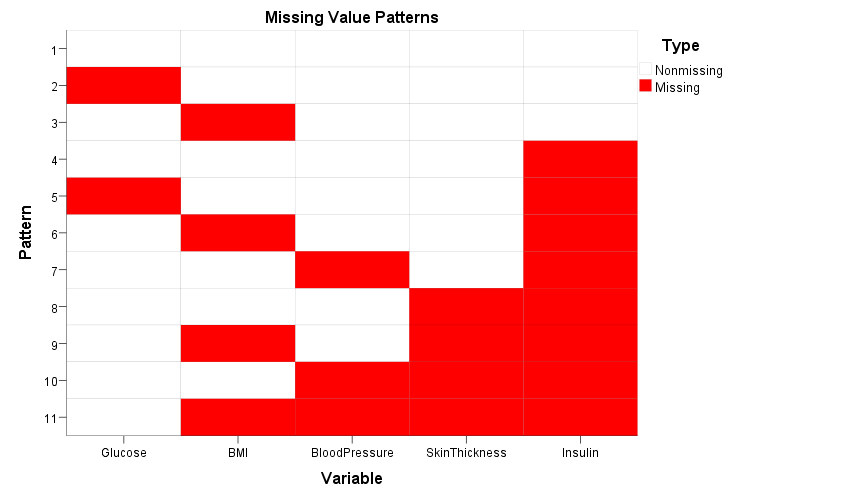


Figure 2: Patterns of missingness in the dataset.

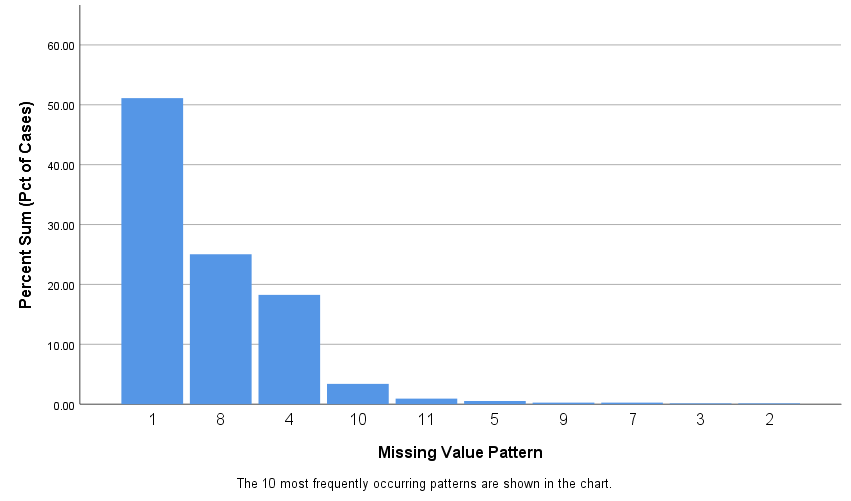


Figure 3. Frequency of missing value patterns within data set.

**Other Exploratory Data Analysis**

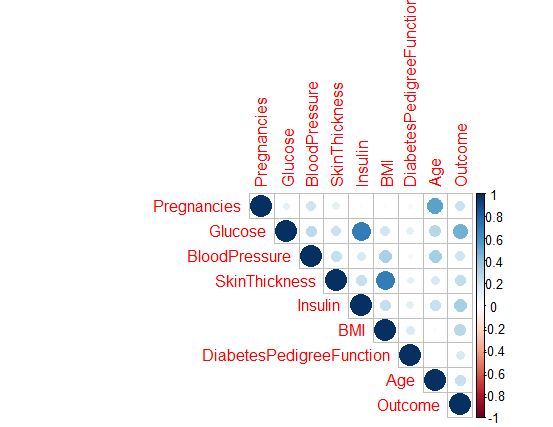
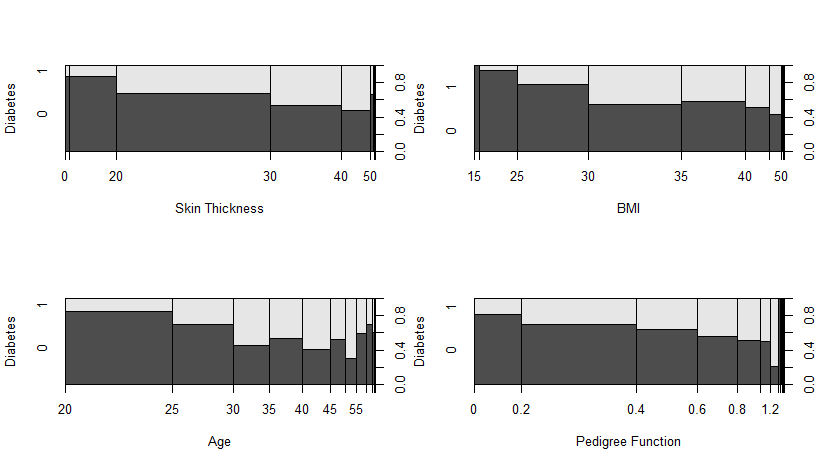
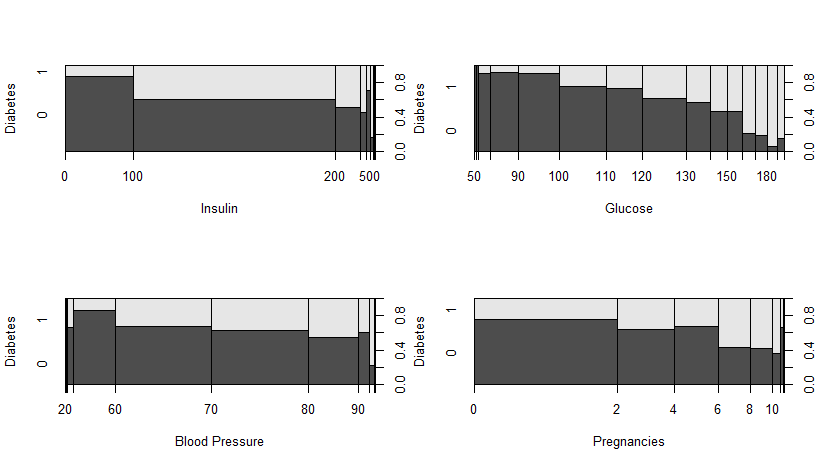


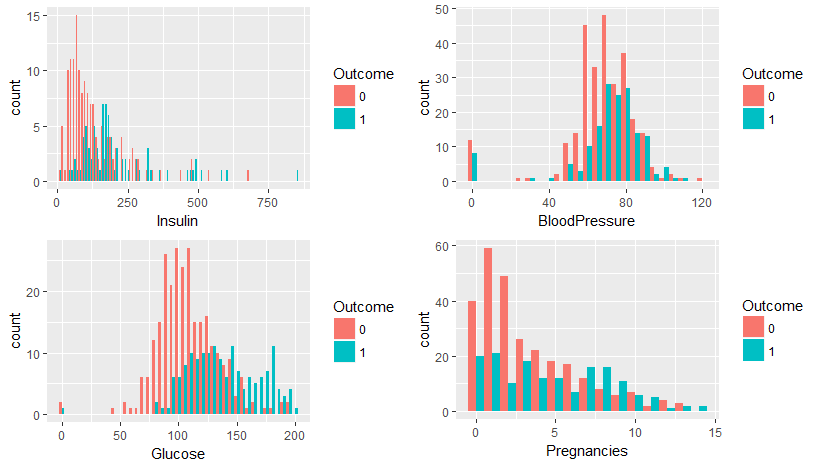
Figure 4: Correlation plot of variables in the data set. Pairwise complete observations were utilized.



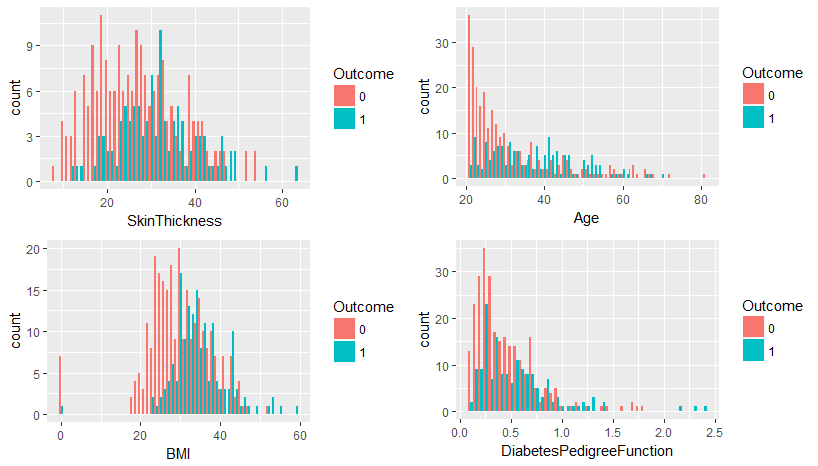
**Figure 5 :** Density plots of Skin Thickness, BMI, Age, and Pedigree Function variables



**Figure 6:** Density Plot of Insulin, Glucose, Blood Pressure, and Pregnancies



**Figure 7:** Histogram of Insulin, Blood Pressure, Glucose, and Pregnancy Variables



**Figure 8:** Histogram of Skin Thickness, Age BMI, and Diabetes Pedigree Function

**Figures 5 through 8** show density plots and histograms showing the relationships between the binary outcome of having and not having diabetes against each of the numerical predictor variables**. Figures 5 and 6** show that BMI, Diabetes Pedigree Function, Glucose, and Age seem to have a large effect on the presence of TIIDM in Pima Indian Females**. Figure 7** shows that glucose levels and the number of pregnancies is higher in females who were found to have TIIDM **while Figure 8** shows that females with TIIDM tended to be older than those that didn’t have TIIDM and that females who had TIIDM tended to have higher BMI values. Some females who had TIIDM had BMI values close to 60 and diabetes pedigree function values greater than 2.0 (**Figure 8**).

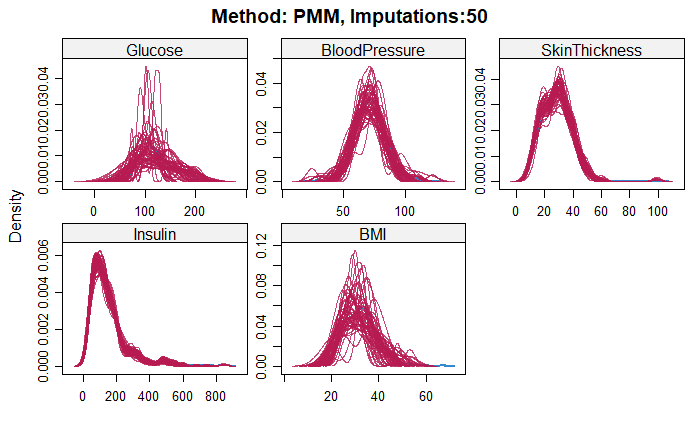
**Imputation for Missing Values**

The best way to deal with missing data is to avoid having any in the first place. In our case, this is not possible. Although it may seem reasonable to remove the variables with a high percentage of missing values (insulin and skin thickness), doing so up front does not allow us to assess the potential usefulness of those variables in our modeling. Imputation is one technique that allows for dealing with missing data. The goal is not so much to replicate as best as possible the exact values of the missing data, but rather to help ensure parameter estimates are unbiased. Two widely used methods are expectation maximization (EM) and multiple imputation (MI).

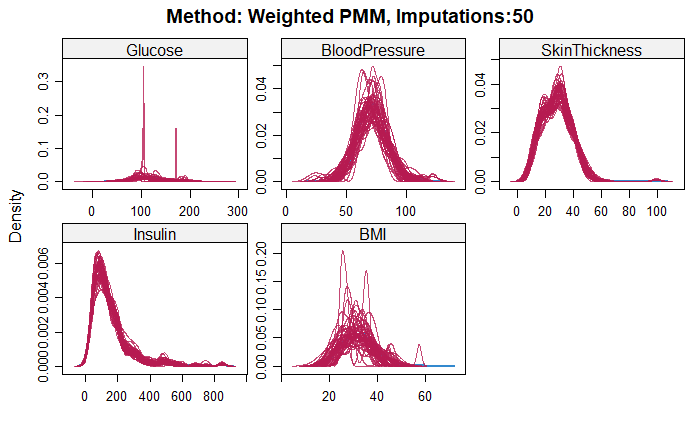
EM involves an iterative approach to obtaining maximum-likelihood estimates of parameters. While EM provides good estimates, it does not provide standard errors, making analyses such as hypothesis testing less feasible. We utilized the PROC MI function in SAS to obtain a single data set with EM imputed values. This was ultimately used for the rest of our analyses due to the simplicity of dealing with one data set.

In MI, values are determined by regressing on other variables in the data set. Rather than providing values which lie directly on the regression line, MI adds random normal error to adjust for error lost from the missing values. Several imputations are performed, and results are analyzed and pooled to obtain errors. For MI, we used the default settings in SPSS with 25 imputations. Furthermore, we utilized the R package MICE to do imputation by predictive mean matching (PMM), weighted PMM, and CART based imputation. 50 imputations were utilized based on recommendation from the literature12. Density plots of the imputed data sets are shown below **(Figures 9 to 12).**

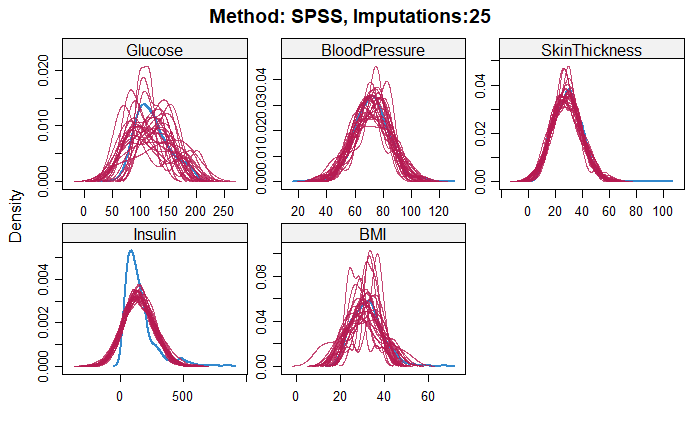
Logistic regression was performed on the imputed data sets, which for each method involved conducting regression on each of the 50 imputed data sets and then pooling the results. Starting with a model that included all the variables as predictors, backwards step-wise selection was manually performed by removing the predictor with the highest p-value after each round. Results from all four MI methods (PMM, weighted PMM, CART imputation, and SPSS) agreed, identifying the following significant predictors: glucose, BMI, pregnancies, and DPF. Logistic regression utilizing the EM imputed data set also had such results. Of note, insulin and skin thickness were not identified as significant in any of the models.



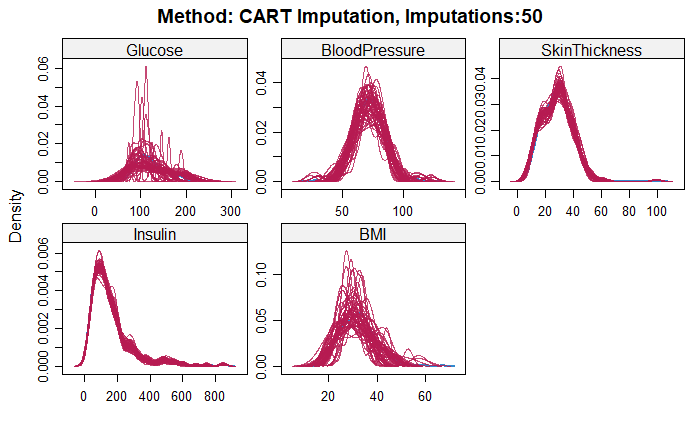
**Figure 9**: PMM Imputation Plot(50 Iterations)



**Figure 10:** Weighted PMM Imputation(50 Iterations)



**Figure 11: SPSS Imputation (25 Iterations)**



**Figure 12: CART Imputation (50 Iterations)**

**Modeling**

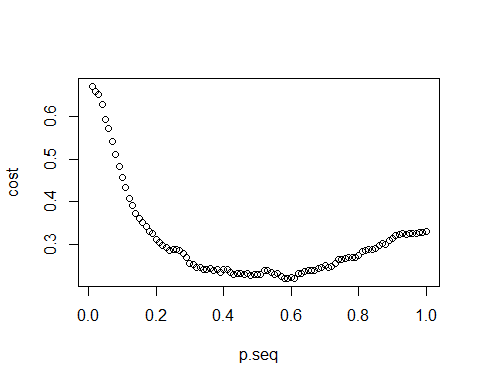
Before modeling we set a seed and split our data into 75% training and 25% testing data. We then proceeded to fit 5 different models to our data.

**GLM Analysis**

Variable selection was done using the stepwise AIC algorithm. The final model had an AIC of 551 and was as follows:

***Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction***

It was interesting to note that the variables insulin and skin thickness were not chosen by the model. Before using the estimated probabilities to do classification, we did a grid search using a symmetric cost function to verify the value of the probability threshold. As expected, the lowest symmetric cost was at a threshold of approximately 0.5 (**Figure 13**).

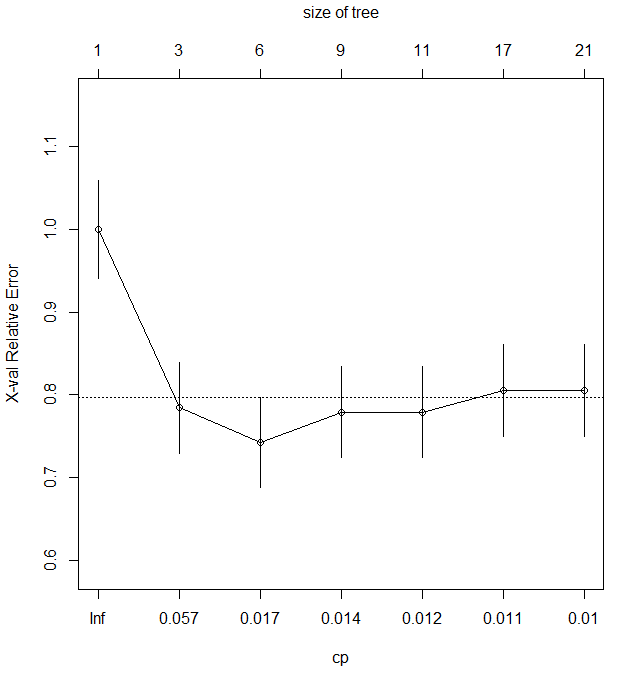


**Figure 13**: Optimal Threshold Value Plot

The in sample and out of sample misclassification rates were 0.22 and 0.23 respectively while the in sample and out of sample AUC values were 0.84 and 0.87 respectively **(Table 3).**

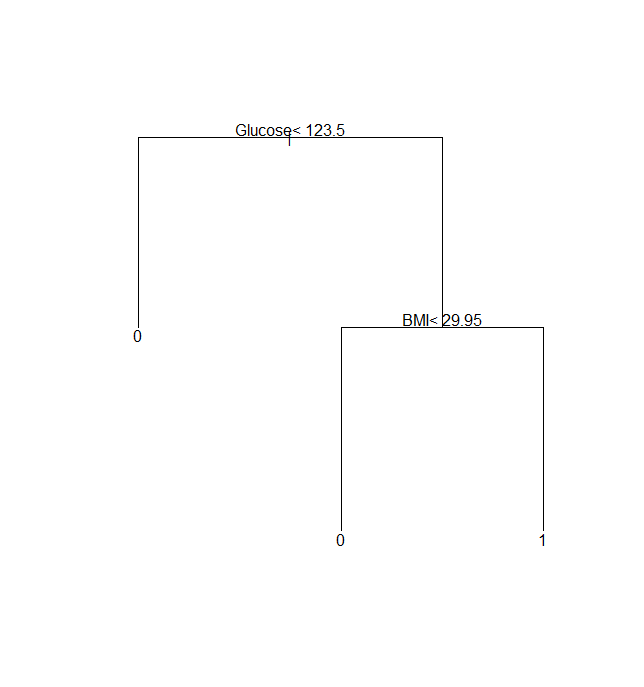
**CART**

To determine tree complexity (cp parameter), the cross-validation error metric was used. As can be seen in **Figure 14** a value of 0.057 seemed to be a good choice.



**Figure 14**: Complexity parameter Findings

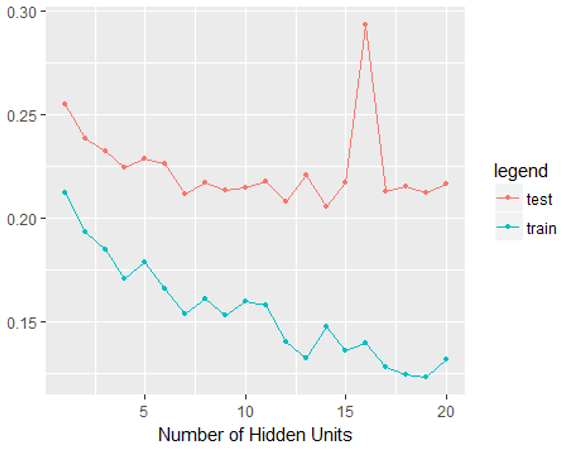
With this value of cp a simple tree with 2 variables, Glucose and BMI, with 2 splits was built as shown in **Figure 15.** The in sample and out of sample misclassification rates were 0.23 and 0.24 respectively while the in sample and out of sample AUC values were 0.88 and 0.84 respectively**(Table3).**



**Figure 15**: Glucose and BMI Classification Tree Splits

**Neural Network Analysis**

For the neural network the nnet package was used. To determine the number of hidden nodes the training data was further divided into a 90/10 split. The in-sample and out-of-sample misclassification rate for this 90/10 split was then plotted and is shown in **Figure 16**. Based on these results the number of hidden nodes was chosen to be 17 (the decay was kept at the default value of 0). The in sample and out of sample misclassification rates were 0.22 and 0.23 respectively while the in sample and out of sample AUC vales were 0.85 and 0.86 respectively **(Table 3).**



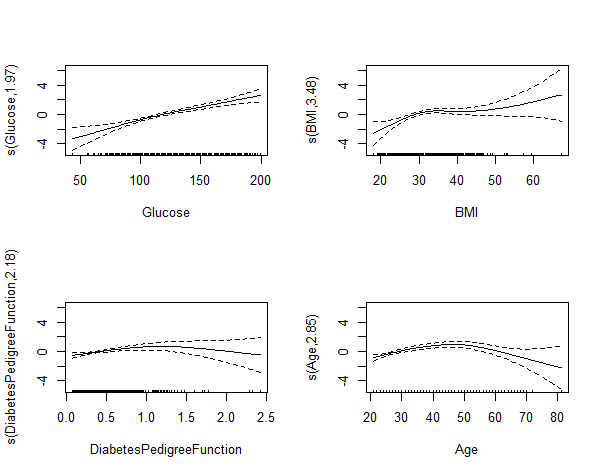
**Figure 16:** In Sample and Out of Sample Misclassification Rate Neural Network

**Adaptive Boosting (AdaBoost)**

Adaptive Boosting was the first really successful boosting algorithm developed for binary classification. It is commonly used with “decision stumps” i.e. binary trees with just one split. Of course the general idea behind boosting is to add weak models sequentially that complement each other’s strengths. This is done by using weighted training data, more weight is given to incorrectly predicted instances from previous models. The process continues until a pre-set number of weak learners created or there is no further improvement. Predictions are made by calculating the weighted average of the weak classifiers, more accurate models have more weight. For AdaBoost we used the boosting function from the *adabag* package in R. A sequence of 20 trees with a depth of 1 were used. In our case boosting only gave a marginal improvement over using a single decision tree. The variables Glucose, Insulin, BMI and Age were identified as the most important predictors. The in sample and out of sample misclassification rates for this method were 0.22 and 0.23 respectively while the in sample and out of sample AUC value curves were 0.85 and 0.87 **(Table 3).**

**GAM Analysis**

Generalized additive models are performed on univariate nonparametric functions in which model assumptions aren’t known. These models capture the non-linear relationship of a function and replace the linear components of generalized linear models by a sum of nonparametric functions over each component of a predictor variable X. Smoothing numerical predictor variables is necessary to find underlying true relationships between the response variables and predictor variables for generalized additive models. Since all the response variables for the diabetes dataset are numerical, they all are smoothed in the full model analysis: **GAM.Diabetes<gam(Outcome~s(Pregnancies)+s(Glucose)+s(BloodPressure)+s(SkinThickness)+s(Insulin)+s(BMI)+s(DiabetesPedigreeFunction)+s(Age),data=Diabetes.train).** From this full model, the glucose, BMI, diabetes pedigree function, and age variables were all found to be significant with p-values less than 0.05. This led to the development of the final model: **GAMBestFit<-gam(Outcome~s(Glucose)+s(BMI)+s(DiabetesPedigreeFunction)+s(Age), data = Diabetes.train).**  The mean residual deviance of this best fit model found by dividing the deviance by the degree of freedom of residuals was 0.883. The edf values for the four significant variables were 1.968, 3.477, 2.183, and 2.848 respectively and the plots were all non-linear indicating that there was no potential for switching any of the variables to a linear term. **Figure 17** below shows the plots for all the significant variables.



**Figure 17**: Plots for Significant Variable for GAM Analysis

The in sample and out of sample misclassification rates for generalized additive modeling method were 0.207 and 0.193 respectively (**Table 3**). The AUC values for the in and out of sample models were 0.860 and 0.878 respectively (**Table 3**). The fact that the in sample and out of sample misclassification rates and AUC values were close to each other indicates a good fit model.

**Table Comparison of Findings from R and SPSS**

GLM, CART, Adaboost, GAM, and Neural network methods were run in R software while GLM, CART, and Neural network modeling methods were run in SPSS. The two tables shown below will compare the in and out of sample misclassification rates along with the in and out of sample AUC values for each of the methods using R (**Table 3**) and SPSS (**Table 4**) software respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **In – Sample**  **AUC** | **Out of Sample**  **AUC** | **In Sample**  **MR** | **Out of Sample**  **MR** |
| GLM | 0.84 | 0.87 | 0.22 | 0.23 |
| Classification Tree | 0.88 | 0.84 | 0.23 | 0.24 |
| AdaBoost | 0.85 | 0.87 | 0.22 | 0.23 |
| **GAM** | **0.86** | **0.88** | **0.21** | **0.19** |
| Neural Network | 0.85 | 0.86 | 0.22 | 0.23 |

**Table 3**: Model Comparison Results using R software

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **In – Sample**  **AUC** | **Out of Sample**  **AUC** | **In Sample**  **MR** | **Out of Sample**  **MR** |
| GLM | 0.86 | 0.81 | 0.21 | 0.25 |
| Classification Tree | 0.86 | 0.80 | 0.18 | 0.26 |
| Neural Network | 0.87 | 0.82 | 0.22 | 0.27 |

**Table 4**: Model Comparison Results using SPSS Software

**Conclusion**

Our goal for this analysis was to find the best prediction of TIIDM onset on our sample of Pima Indian females. We used imputation methodology (expectation maximization and predictive mean matching) to improve the effect of the significant predictor variables. We ran different models on the predictor variables and found out their significance on prediction of the binary outcome value of whether Pima Indian females had TIIDM. Considering the generalized linear model, the variables: Pregancies, Glucose, BMI, and Diabetes Pedigree Function were significantly related to the outcome. The nonlinear generalized additive model had Glucose, BMI, Diabetes Pedigree Function, and Age as significant predictor variables. By looking at our results from the GLM, CART, GAM, AdaBoost, and Neural network analysis, we decided to use GAM as our best model because the in sample and out of sample misclassification rates and AUC results were closest to each other**(Tables 3 and 4).** Cross-validation methodology would give more reliable results/comparisons and can be considered for future analysis.

**References**

1. Kahn SE, Cooper ME, Del Prato S. PATHOPHYSIOLOGY AND TREATMENT OF TYPE 2 DIABETES: PERSPECTIVES ON THE PAST, PRESENT AND FUTURE. Lancet. 2014;383(9922):1068-1083. doi:10.1016/S0140-6736(13)62154-6.

2. Afable A, Karingula NS. Evidence based review of type 2 diabetes prevention and management in low and middle income countries. World Journal of Diabetes. 2016;7(10):209-229. doi:10.4239/wjd.v7.i10.209.

3. Olokoba AB, Obateru OA, Olokoba LB. Type 2 Diabetes Mellitus: A Review of Current Trends. Oman Medical Journal. 2012;27(4):269-273. doi:10.5001/omj.2012.68.

4. Asif M. The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern. Journal of Education and Health Promotion. 2014;3:1. doi:10.4103/2277-9531.127541.

5. Soyiri IN, Reidpath DD. An overview of health forecasting. Environmental Health and Preventive Medicine. 2013;18(1):1-9. doi:10.1007/s12199-012-0294-6.

6. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care. November 1988:261-265.

7. Kaggle, “Pima Indians Diabetes Database”. Accessed April 3,2018. https://www.kaggle.com/uciml/pima-indians-diabetes-database

8. Esparza-Romero J, Valencia ME, Urquidez-Romero R, et al. Environmentally Driven Increases in Type 2 Diabetes and Obesity in Pima Indians and Non-Pimas in Mexico Over a 15-Year Period: The Maycoba Project. Diabetes Care. 2015;38(11):2075-2082. doi:10.2337/dc15-0089.

9. Esparza-Romero J, Valencia ME, Martinez ME, Ravussin E, Schulz LO, Bennett PH. Differences in Insulin Resistance in Mexican and U.S. Pima Indians with Normal Glucose Tolerance. The Journal of Clinical Endocrinology and Metabolism. 2010;95(11):E358-E362. doi:10.1210/jc.2010-0297.

10. Melillo W. Why are the pima indians sick?: Studies on arizona tribe show excessive rates of diabetes, obesity and kidney disease. The Washington Post (1974-Current file). 1993:7.

11. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549-76. doi: 10.1146/annurev.psych.58.110405.085530.

12. Todd E. Bodner (2008) What Improves with Increased Missing Data Imputations?, Structural Equation Modeling: A Multidisciplinary Journal, 15:4, 651-675,DOI: [10.1080/10705510802339072](https://doi.org/10.1080/10705510802339072)