Final Project
Sean Chapman
0959653

Abstract:

The pima dataset contains information which can be used to predict whether an individual is showing signs of diabetes. The data set contains information on 768 adult women, 245 of these records contain impossible values for one or more covariates. Various statistical learning models were built on the data with the aim of predicting the response of showing signs of diabetes. Generalized linear models, generalized additive models, decision trees and neural networks were built using the data to compare the predictive ability of different modelling approaches. The neural network with 5 hidden nodes had the lowest misclassification rate at 6%, while the optimal binary decision tree had a misclassification rate of 9% on an out-of-bag test sample. While the neural network has a better predictive accuracy, it is a complicated model to interpret as the classification algorithm used by the hidden nodes is unknown. Therefore, the optimized binary tree provides the highest predictive accuracy while remaining interpretable.

Introduction:

The pima dataset includes information recorded from a study conducted on 768 adult women by the National Institute of Diabetes and Digestive and Kidney Diseases. All individuals are Pima Indians living near Phoenix. The dataset contains information which experts believe to be related to instances of diabetes.

Table 1: Variables measured in pima diabetes data set. Information for each variable includes the name in the data frame, the unit of measurement and a brief description what the measurement represents. Descriptions found from CRAN of package faraway https://cran.r-project.org/web/packages/faraway/faraway.pdf.

Name	Unit of measurement	Description
pregnant	Count	Number of times pregnant
glucose	mmol/L	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
diastolic	Mm Hg	Diastolic blood pressure
triceps	Mm	Skin fold thickness of the triceps
insulin	Mu U/mm	2-hour serum insulin

bmi	Kg/m^2	Body mass index
diabetes		Diabetes pedigree function
age	Years	Individual's age in years
test	0 if negative, 1 if positive	Test indicating signs of diabetes if positive

The range of possible values which can be observed for each variable is well understood by medical experts. We can use this knowledge to assess the data set for impossible values of certain covariates. Removing these impossible values and replacing with NA values ensured a more accurate analysis of the data. For some models, NA values were not permitted in the data object which trains the models in R. For these models the set of points with complete data was used.

The response variable from this dataset is "test". Statistical learning models built using this dataset had the goal of classifying new individuals as positive or negative using the information contained within the other variables. The response variable is binary, which means that the goal of a predictive model is to assign a probability to each observation as being classified as positive or negative. From this probability we can assign a classification threshold to assign points as being either positive or negative for the test value. Since the models cannot be tested on additional data, we will use cross validation to train and test the various models. The test set was constructed using 100 observations from the subset of data not containing impossible values. The training set was either the remaining observations or the remaining observations not containing missing values. Models which did not include missing values in the training set were the reduced GLM, the reduced GAM and the neural networks. Additionally, most of the neural networks used standardized data as the input. Data was standardized using the scale() function in R.

Methods (couple pages):

The goal of the analysis is to find a model which can accurately assign individuals as testing positive or negative for signs of diabetes using the available information. There were several different statistical learning approaches used to assign the data. These methods were tested and compared to decide which model has the best predictive capability under the ideal conditions. The first method which was explored was the generalized linear model. This method is appropriate for the analysis of the pima data as the response variable is binary. This method is the simplest approach from a statistical perspective presented in this paper.

The next method which was explored using the pima dataset was the generalized additive model. This modelling approach does not operate under the assumption of some linear relationship between the explanatory variables and the response but allows the data to

indicate some function which best fits the data. This approach can lead to a better fitting model but is less parsimonious. The generalized additive model used in this analysis contains fitted smooths for all covariates.

Another method which was explored in the analysis of the data was classification trees and random forest models. The default tree contains 13 decision nodes to classify the test class of the observation. The cp scores of different sized trees were compared to find the tree with the lowest cross validated error scaled by the residual sum of squares (RSS) of the null tree. This tree was selected as the optimal tree size and plotted.

The final method which was explored in the analysis of the pima data set was neural networks. This modelling approach builds a neural network of input, output, and hidden nodes and connects them using different weights. The input nodes include each of the covariates. The hidden nodes contain some unknown algorithm which processes the inputted covariates using different weights. The output node contains a predicted value for the probability of the response which we can use to assess accuracy. The length of the output prediction matches the inputted training data. Therefore, the performance of the neural nets was assessed using the observed values of the response from the neural network training data.

Results section:

The distributions of the covariates were visualized in order to find impossible values. The plotted distributions and their relationship to the test variable are shown in figures 1-8. Covariates containing impossible values were re-plotted and included to observe changes in the distribution. The first covariate indicates the number of times pregnant for each observed woman in the data set. The median number of times pregnant is slightly higher among those who tested positive for diabetes, although there is a lot of overlap between the distributions. There is no obvious effect between the number of times pregnant and having diabetes.

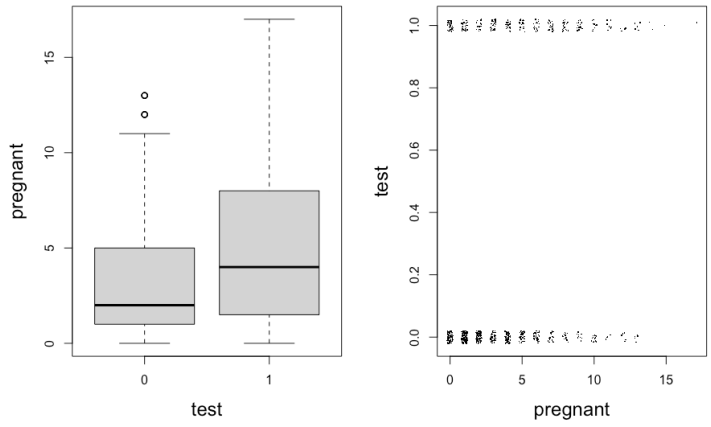

Figure 1: Distribution of the number of times pregnant among different test results.

The distribution of blood glucose concentrations at 2 hours during an oral glucose concentration test is plotted in figure 2. Patients with diabetes are expected to have a higher blood glucose concentration than those without diabetes as one of the symptoms of diabetes is an inability to uptake glucose due to a resistance to insulin. Additionally, blood glucose levels of zero are impossible and are seen as severe outliers on the left side of figure 2. While the median blood glucose concentration is higher at 2 hours, there is still overlap between diabetes positive and negative patients.

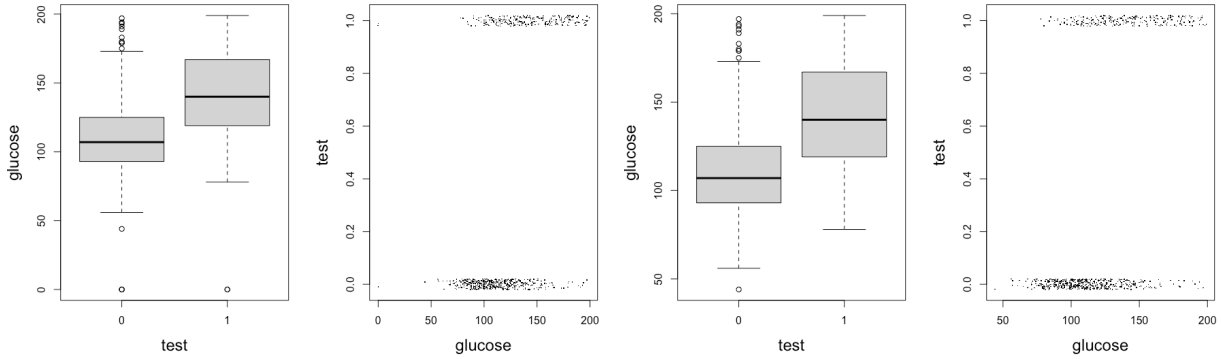

Figure 2: Distribution of plasma glucose concentration at 2 hours among different test results. Impossible values were observed (left) and replaced with NA values (right).

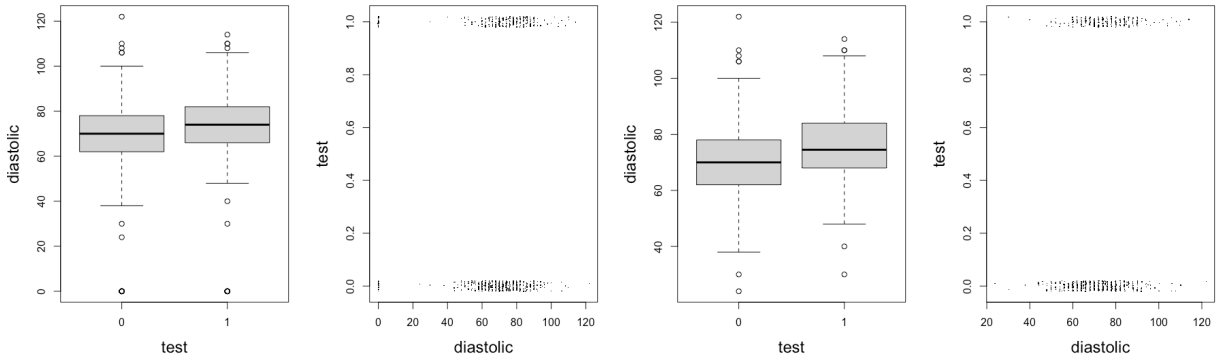

Figure 3: Distribution of diastolic blood pressure among different test results. Impossible values were observed (left) and replaced with NA values (right).

The distribution of diastolic blood pressure also contained impossible values as there were several patients having an observed diastolic blood pressure of 0. This is not possible in a living patient and are extreme outliers in the left graphic of figure 3. After replacing with NA values, the median blood pressure of diabetes positive patients is observed to be slightly higher than diabetes negative patients although there is still overlap between the two groups of patients.

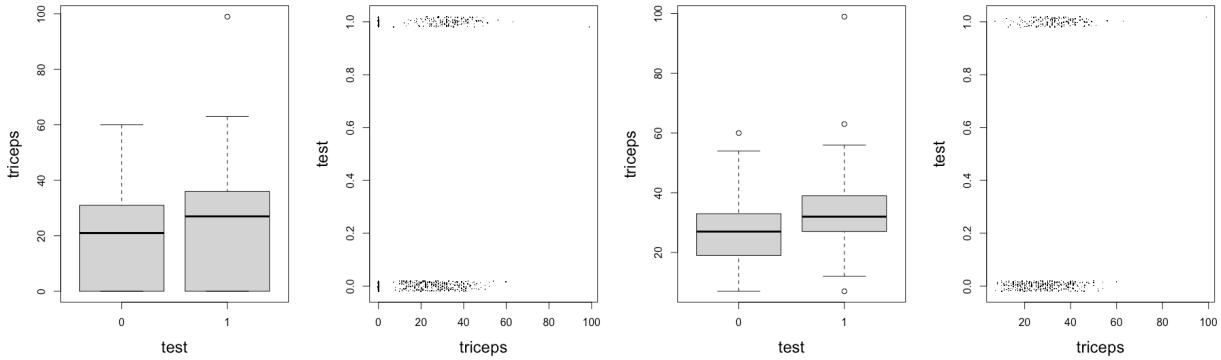

Figure 4: Distribution of triceps skin fold thickness among different test results. Impossible values were observed (left) and replaced with NA values (right).

The distribution of triceps skin fold thickness is shown in figure 4. A skin fold thickness of 0 is impossible and observations with this value were replaced with NA values. The distribution of triceps skin fold observations seems to be higher among patients positive with diabetes. These values are representative of body fat, with higher skin fold thickness correlating to a higher body fat percentage.

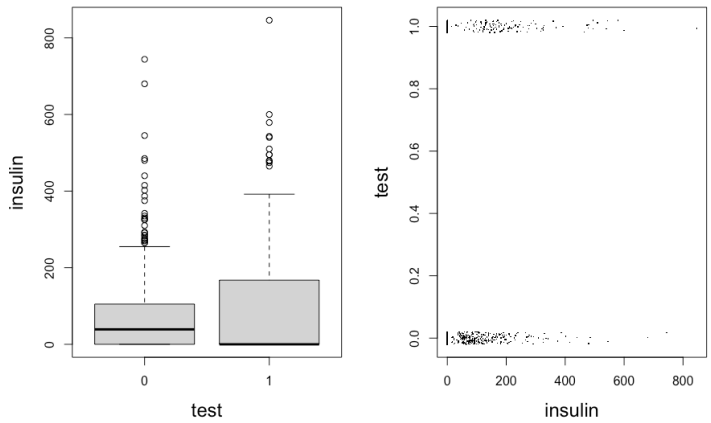

Figure 5: Distribution of serum insulin at 2 hours among different test results.

The distribution of serum insulin at 2 hours is shown in figure 5. While the jitter plot shows many observations with a value of 0, these are expected observations for serum insulin when observing diabetic patients. Since type 1 diabetic patients cannot produce insulin, it would be expected that the diabetic patients have a serum insulin value of 0. The median value of insulin among diabetic patients is 0. Type 2 diabetic patients can produce insulin but are resistant to it. Therefore, the remaining observations have a higher value for insulin than the non-diabetic patients.

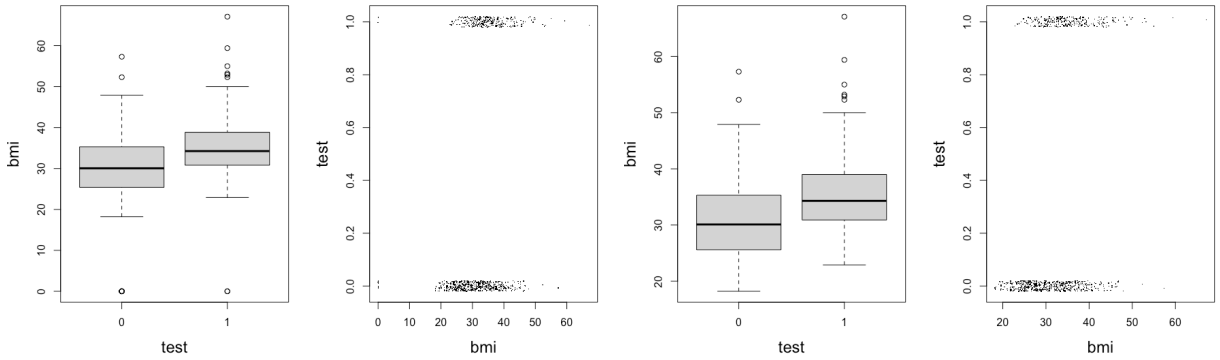

Figure 6: Distribution of body mass index (BMI) among different test results. Impossible values were observed (left) and replaced with NA values (right).

The distribution of body mass index is shown in Figure 6. BMI represents the ratio of weight to height of an individual. Individuals with a higher BMI are heavier than other individuals with the same height. A BMI of zero is impossible as this would indicate the patient has no weight. These points were replaced with NA values. The distribution of diabetic patients shows a higher median BMI with a lot of overlapping values between diabetic and non-diabetic patients.

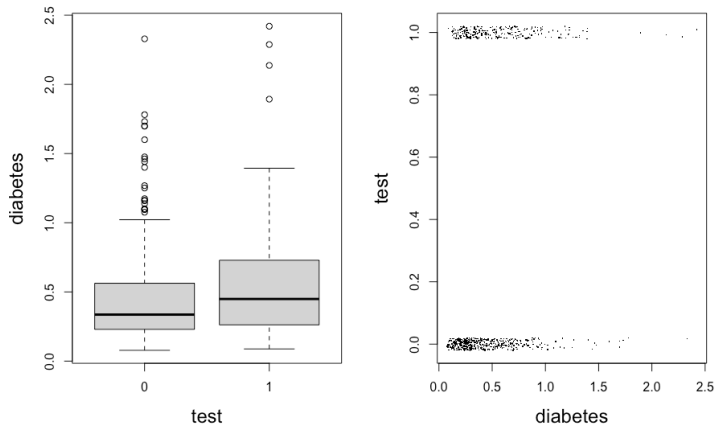

Figure 7: Distribution of the diabetes pedigree function among different test results.

The distribution of the diabetes pedigree function is shown in figure 7. This function represents the proposed hereditary component of diabetes. A value from the function represents the family history of the disease specific to the patient. A higher value indicates more family history of having diabetes. The distribution of diabetic patients shows a higher median for the diabetes pedigree function suggesting some relationship between the diabetes pedigree and the risk of being diabetic.

The distribution of age is shown in figure 8. All the individuals participating in the study are over the age of 20. The distributions suggest that diabetic patients have a higher median age. While type 1 diabetes is typically developed at a young age, type 2 diabetes is developed over time and would be more prevalent in older individuals.

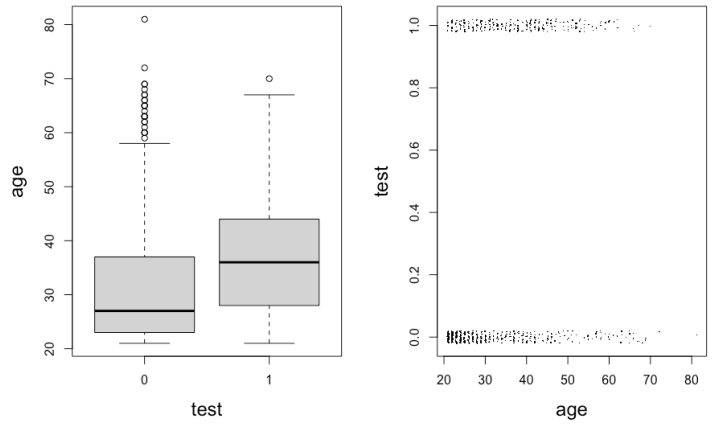

Figure 8: Distribution of age among different test results.

Results:

GLM Analysis

The first model constructed to predict the patients which are diabetic is the GLM. The full GLM was built using the training data and all covariates using test as the response. The estimates for the parameters of the model are included in table 2. From these estimates, the influence of the number of pregnancies, glucose concentration, bmi, diabetes pedigree and age are all significant in predicting the test result of a given patient. Of these, glucose concentration had the largest effect on the outcome of the model.

Table 2: Summary of the full GLM. The model has an Akaike Information Criterion (AIC) of 421.42, a null deviance of 105.849 on 469 df and a residual deviance of 64.647 on 461 df.

Parameter	Estimate	Std Error	P-Value
Intercept	-1.091e+00	1.247e-01	< 2e-16
pregnant	1.822e-02	6.787e-03	0.007515
glucose	6.694e-03	6.680e-04	< 2e-16
diastolic	-1.867e-03	1.624e-03	0.250780
triceps	-4.216e-04	2.202e-03	0.848251
insulin	1.025e-05	1.639e-04	0.950163
bmi	1.316e-02	3.529e-03	0.000216
diabetes	2.069e-01	5.295e-02	0.000107
age	5.152e-03	2.267e-03	0.023515

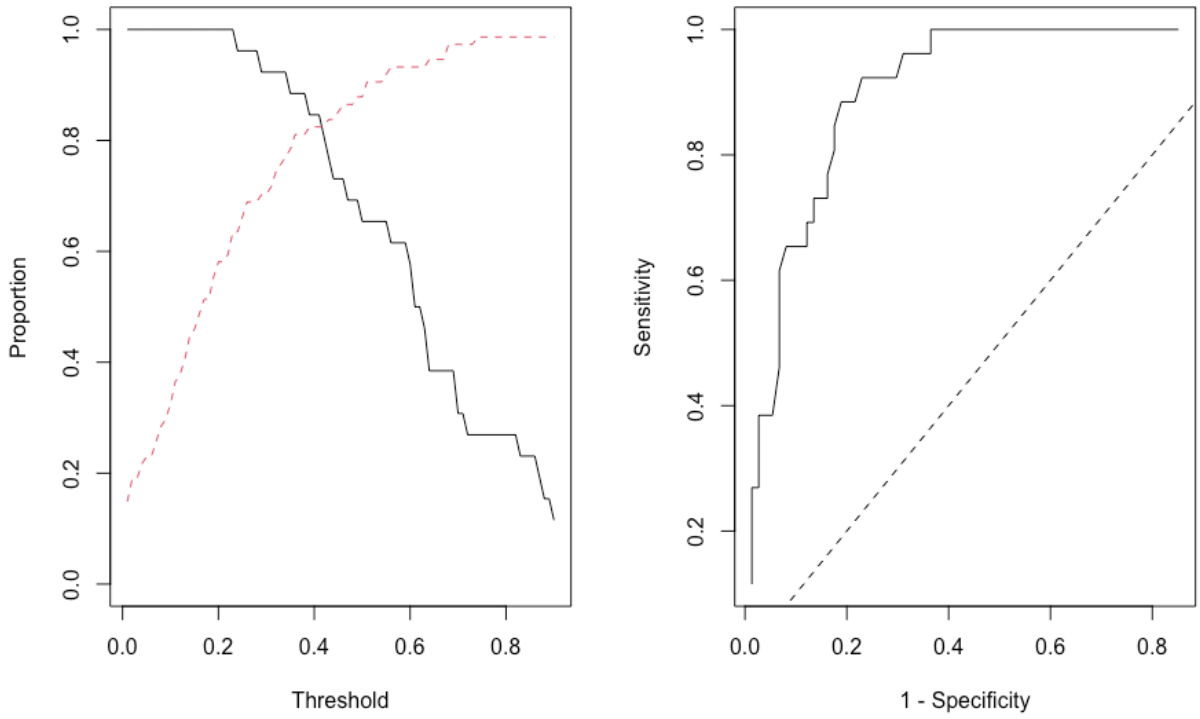

Figure 11: Sensitivity and Specificity of the full GLM predicting test results plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.43 was chosen to maximize both Sensitivity and Specificity of the model.

Once the model was built, the next step was to use the model to predict the test outcome of the observations in the test set. In order to do this, a classification threshold needed to be chosen which would minimize the misclassification rate of the model. The classification threshold of 0.43 was chosen from figure 11 as the model which maximizes the proportion of correct predictions. Using this classification threshold, the results of the prediction of the test set using the full GLM are shown in table 3.

Table 3: Confusion matrix for the full GLM. The full GLM has a misclassification rate of 18%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	62	12
	Positive	6	20

The full GLM was able to predict the outcome of the test results with 82% accuracy. The next step in the analysis is to reduce the model and test for accuracy with the goal of finding a more interpretable model with fewer parameters that maintains a high accuracy.

Reduced GLM:

Building a reduced GLM using the step() function in R yields a model which includes fewer parameters. The parameters in this model which are significant at the 95% level include the number of pregnancies, glucose concentration, BMI and the diabetes pedigree function. Of these, the diabetes pedigree function had the largest effect on the outcome of the model.

Table 4: AIC: 421.81, null deviance: 98.22 on 431 df, residual deviance: 65.01 on 426 df

Parameter	Estimate	Std. Error	P-Value
(Intercept)	-1.0769162	0.1149217	< 2e-16
pregnant	0.0231448	0.0075604	0.002343
glucose	0.0055239	0.0006448	< 2e-16
bmi	0.0133766	0.0029722	8.76e-06
diabetes	0.2074843	0.0570181	8.76e-06
age	0.0040464	0.0024018	0.092772

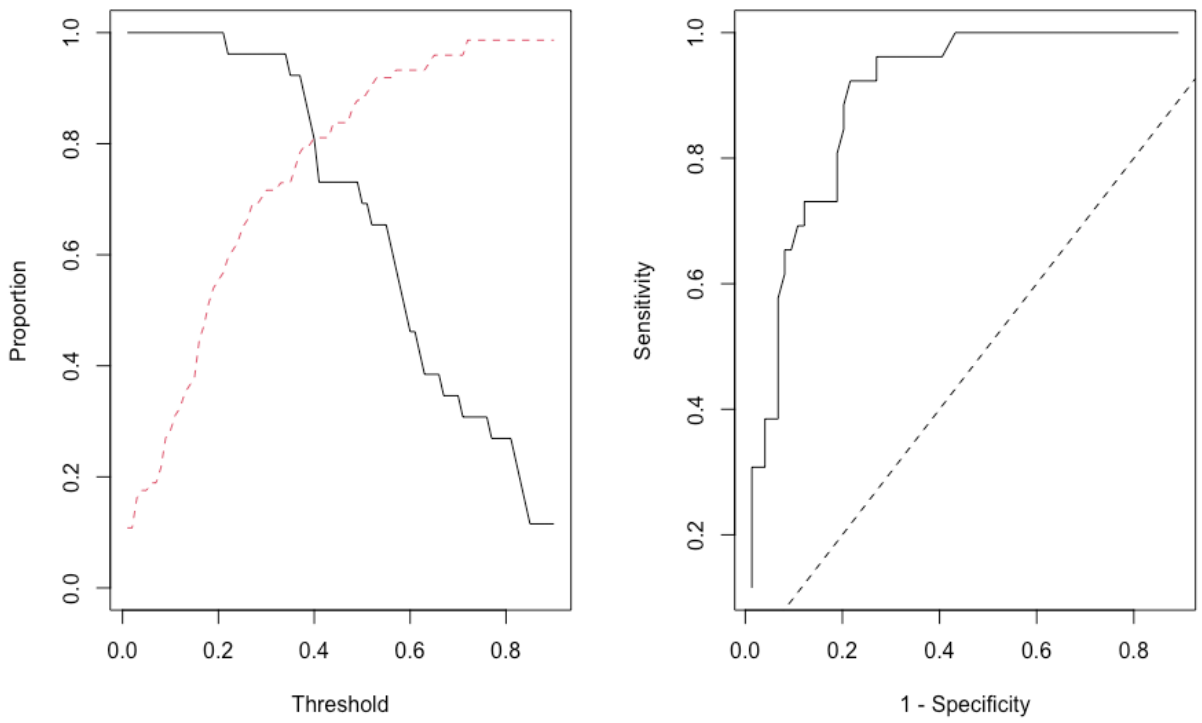

Figure 12: Sensitivity and Specificity of the reduced GLM predicting test results plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.40 was chosen to maximize both Sensitivity and Specificity of the model.

Once the reduced model was built, the next step was to use the model to predict the test outcome of the observations in the test set. The classification threshold of 0.43 was chosen from figure 12 as the model which maximizes the proportion of correct predictions. Using this classification threshold, the results of the prediction of the test set using the reduced GLM are shown in table 4.

Table 4: Confusion matrix for the reduced GLM. The reduced GLM has a misclassification rate of 19%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	60	14
	Positive	5	21

The full GLM was able to predict the outcome of the test results with 82% accuracy. The next step in the analysis is to reduce the model and test for accuracy with the goal of finding a more interpretable model with fewer parameters that maintains a high accuracy.

Full GAM:

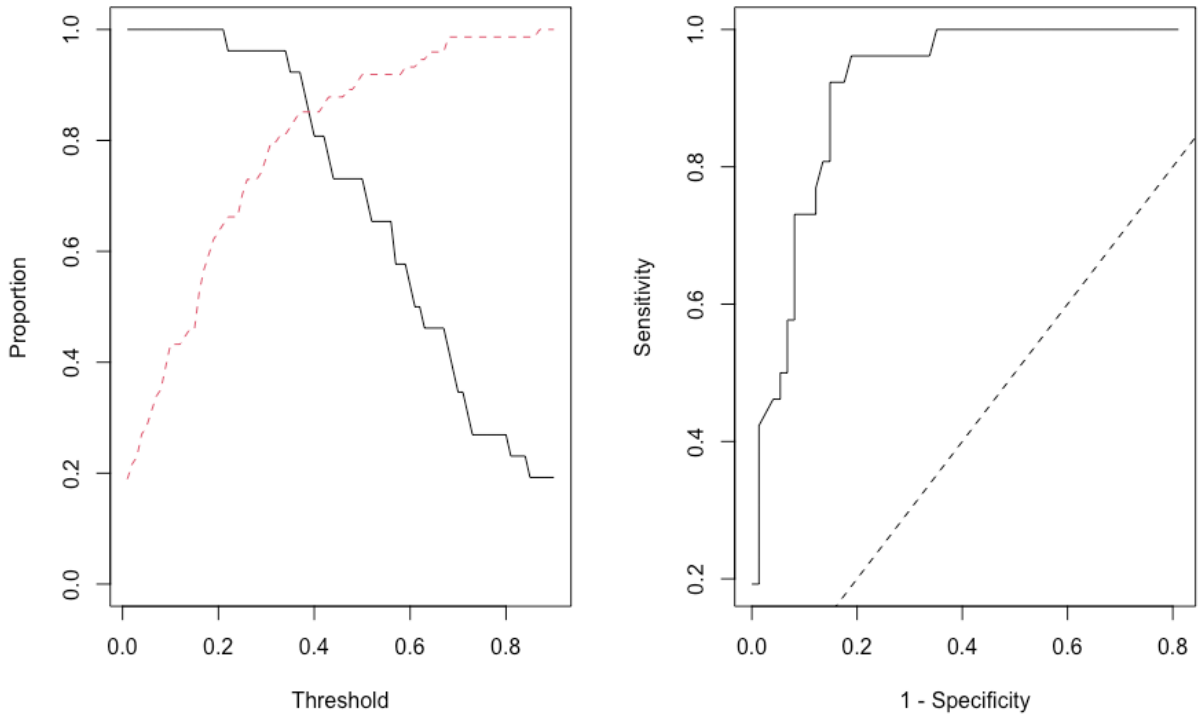

Figure 12: Sensitivity and Specificity of the full GAM predicting test results plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.37 was chosen to maximize both Sensitivity and Specificity of the model.

The next model constructed was the full GAM. This model was built using smooth fits on all the covariates in the model. The model was then used to classify the observations of the test set using an optimized classification threshold of 0.37. The results of the predictions are shown in table 6. The model has an accuracy of 87%, which is significantly better than either the full or reduced GLMs.

Table 6: Confusion matrix for the full GAM. The full GAM has a misclassification rate of 13%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	63	11
	Positive	2	24

Reduced GAM:

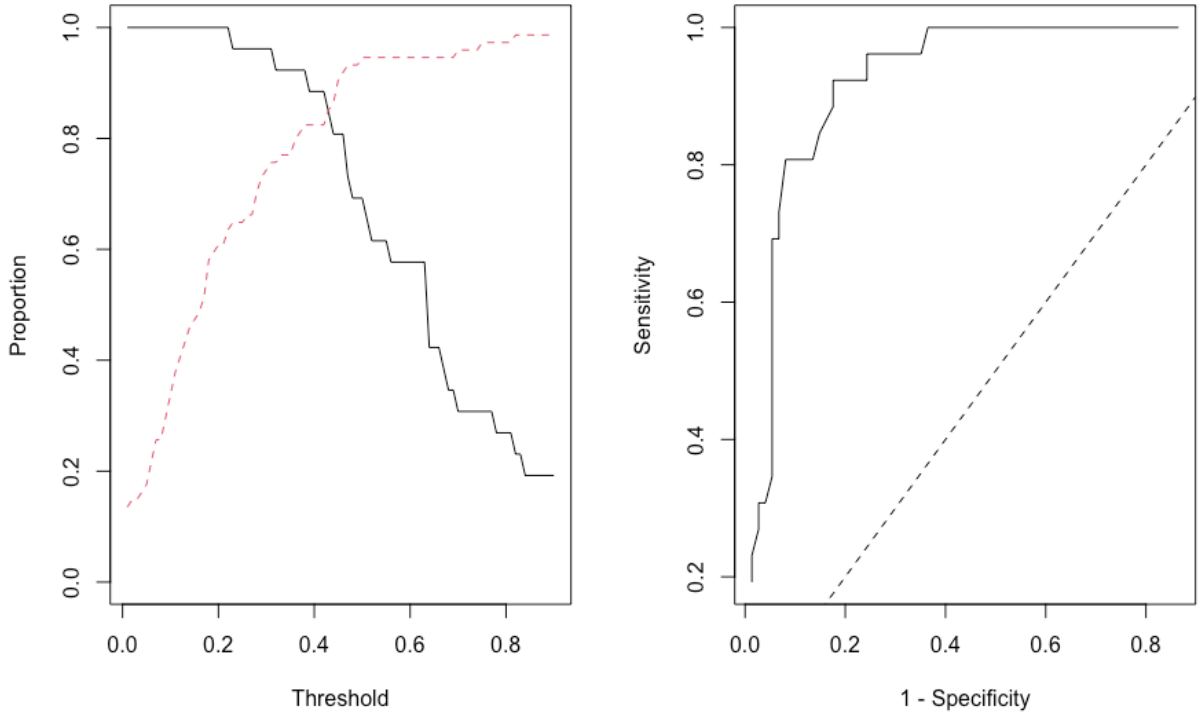

Figure 12: Sensitivity and Specificity of the reduced GAM predicting test results plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.46 was chosen to maximize both Sensitivity and Specificity of the model.

The full GAM was reduced using the step() function in R to create a reduced GAM. This model was built using smooth fits on select covariates in the model. The model was then used to classify the observations of the test set using an optimized classification threshold of 0.46. The

results of the predictions are shown in table 7. The model has an accuracy of 89%, which is slightly better than the full GAM.

Table 5: Confusion matrix for the reduced GAM. The reduced GAM has a misclassification rate of 11%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	68	6
	Positive	5	21

Default Binary Tree:

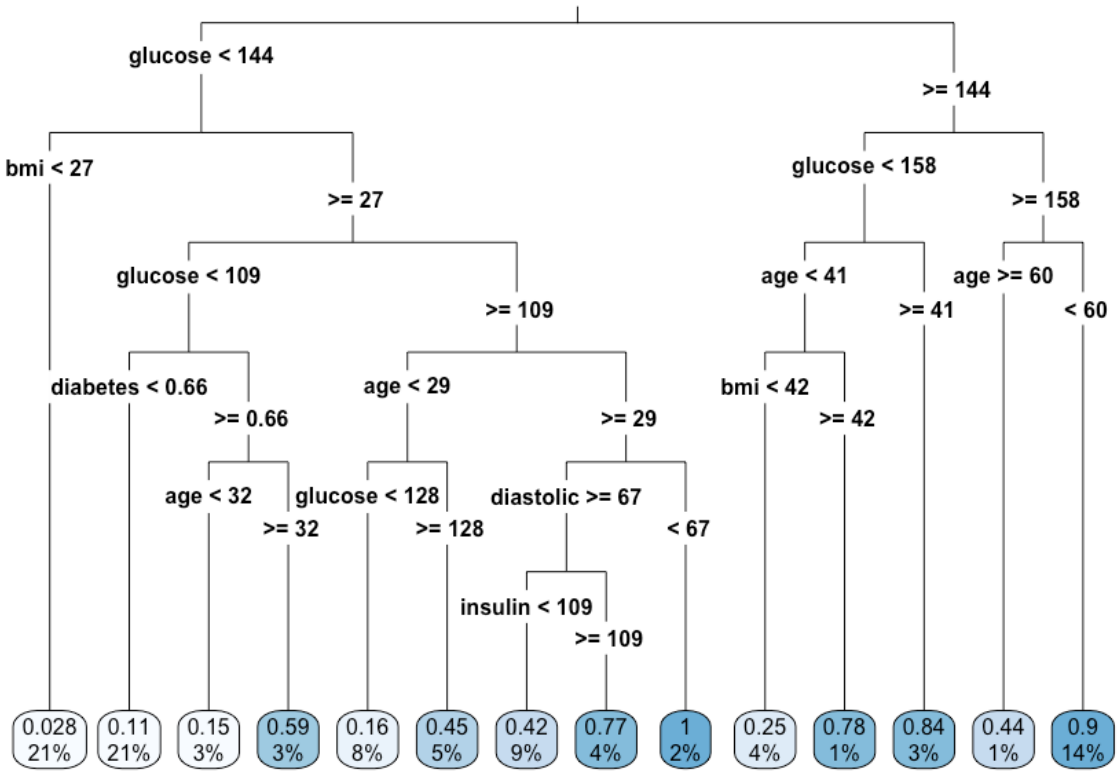

Figure 13: The default binary decision tree.

A binary decision tree was constructed using the analysis techniques from Faraway (2016). The decision nodes at the bottom of figures 13 and 14 represent the probability of this group of data to be positive for diabetes, as well as the proportion of data which is represented in the node. Decisions which are more likely to be classified as diabetic are darker blue. The tree in figure 13 was used to classify the test set as positive or negative and performed with 89% accuracy. A classification threshold of 0.43 was chosen as the ideal threshold.

Table 7: Confusion matrix for the default binary tree. This model has a misclassification rate of 11%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	68	6
	Positive	5	21

Optimal Tree:

In order to build the optimal tree, the cp-scores and cross validated error of different trees. These trees contain a different number of decision nodes and are shown in table 8. From these the tree with the lowest cross validated error was selected as the ideal tree and was constructed using the corresponding cp-score.

Table 8: Cp-score and cross validated error of trees with differing decision nodes.

Number of decision nodes	Cp score	Cross Validated error
0	0.2105485	1.00153
1	0.0521500	0.85705
2	0.0475619	0.81349
3	0.0309329	0.77825
4	0.0251328	0.73631
5	0.0177567	0.72676
6	0.0156413	0.72943
8	0.0124788	0.73238
9	0.0114810	0.73849
10	0.0113032	0.73961
12	0.0112979	0.73852
13	0.0082938	0.73535
14	0.0071623	0.73305
15	0.0065432	0.72907
17	0.0057999	0.72193
18	0.0050198	0.70730
19	0.0048329	0.71518
20	0.0043558	0.71480

The tree with the optimal cross validated error score was selected as the tree containing 18 decision nodes and was built using the rounded cp-score of 0.005. This tree is shown in figure 14.

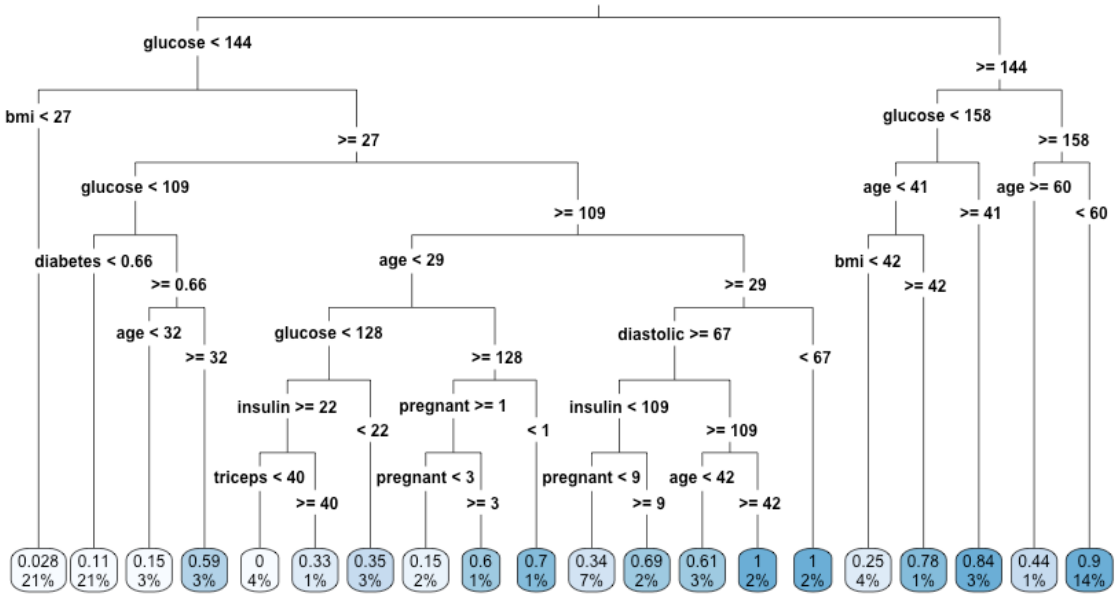

Figure 14: Optimal binary decision tree.

This tree was then used to classify the test outcomes of the observations from the test set. The results of this classification are shown in table 9. The optimal tree performs with 91% accuracy. This is the highest accuracy of the models tested to this point. The model uses a classification threshold of 0.6 to classify the observations in the test set.

Table 9: Confusion matrix for the optimal binary tree. This model has a misclassification rate of 9%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	69	5
	Positive	4	22

Random Forest

A default random forest model was constructed using the data and tested for performance. This model used a classification threshold of 0.48 and was able to classify the results of the test set with 87% accuracy. This performs more accurately than the GLMs and GAMs, but not as well as the binary decision trees.

Table 10: Confusion matrix for the default random forest. This model has a misclassification rate of 13%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	67	7
	Positive	6	20

A subset random forest model was created using the subset of covariates with the lowest MSE using cross validation tests to reduce the number of covariates used in the forest. The subset random forest model was then used to predict the results of the test set, with an 87% accuracy. A classification threshold of 0.6 was used for the greatest accuracy.

Table 11: Confusion matrix for the subset random forest. This model has a misclassification rate of 13%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	70	9
	Positive	4	17

NN Analysis

Several neural networks were constructed to predict the outcome of the input data. The networks used were all feed forward neural networks with one hidden layer. The initial neural networks contain 2 hidden nodes while others were constructed using 8 hidden nodes. The neural networks were not assessed using cross validation but assessed using their ability to predict the input data which was used as an input.

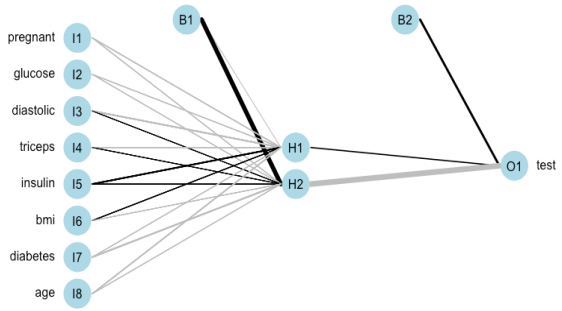

Figure 16: Feed forward neural network with 2 hidden nodes. Model fitted on the full data set.

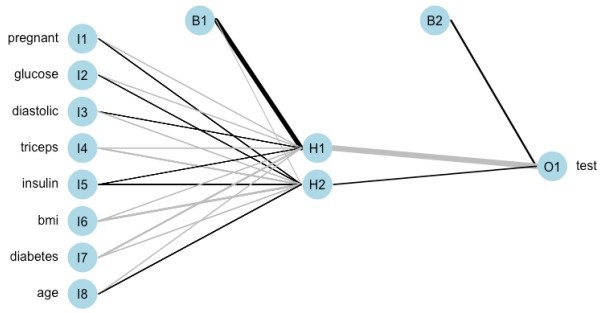

Figure 17: Feed forward neural network with 2 hidden nodes. Model fitted on the full standardized data set.

The length of the output probabilities are the same as the length of the data used to train the data, which can create an issue when the test set has fewer observations than the training set. There are 768 data points in the predicted values of the neural networks trained using the full data set. This is the same as the rows in the data frame used to train the model. The first neural network was constructed using the raw data. As such, the model output is the same for each predicted value. These results are not very informative but can still predict the outcome with 79% accuracy as shown in table 12. The data was then converted to standard units and used to train a new network shown in figure 17. The outputs of this model were much more informative as they differed depending on the data point. Each value represented the probability of a given point to test positive for diabetes. It was important to include the response variable as a factor and not include the response in the standardization of the data set. Once the data was normalized, a neural network trained on the full standardized data set performed with a 84% accuracy.

Table 12: Confusion matrix for the neural net trained on the full data set. This model has a misclassification rate of 21%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	312	43
	Positive	68	109

After testing the accuracy of the neural network shown in figure 17, the same approach was iterated 100 times and the best model of these was selected using the residual sum of squares. This network, shown in figure 18 was constructed because neural networks trained on the same data can have slightly different weights. There is a randomized component to the neural network which can affect the performance of the model. This model was able to classify the data set with an accuracy of 84%

Table 13: Confusion matrix for the neural net trained on the full standardized data set. This model has a misclassification rate of 16%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	323	32
	Positive	51	126

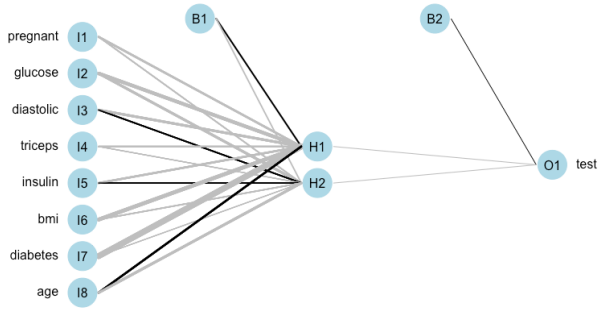

Figure 18: Feed forward neural network with 2 hidden nodes. Model selected from 100 networks fitted on the full standardized data set.

Table 14: Confusion matrix for the selected best neural net trained on the full standardized data set. This model has a misclassification rate of 16%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	224	57
	Positive	14	137

For the next neural network, the number of hidden nodes was increased to improve the predictive capability of the model. The model shown in figure 19 was trained on the standardized training data set, while the network in figure 20 was trained using the full standardized data set. The performances of these models are shown in tables 15 and 16.

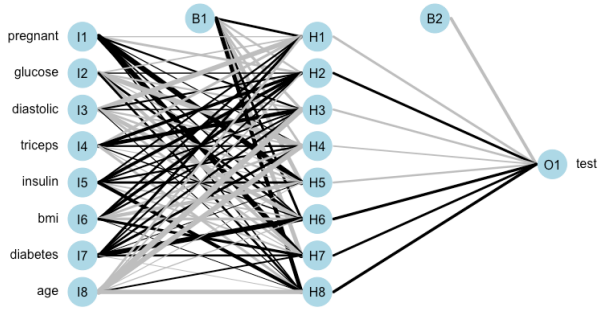

Figure 19: Feed forward neural network with 8 hidden nodes. Model selected from 100 networks fitted on the standardized training data set.

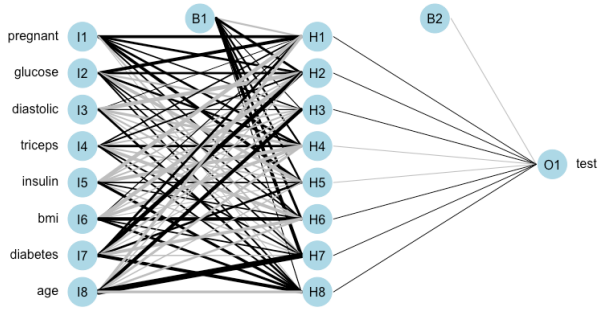

Figure 20: Feed forward neural network with 8 hidden nodes. Model selected from 100 networks fitted on the full standardized data set.

Table 15: Confusion matrix for the selected best neural net trained on the standardized training data set. This model has a misclassification rate of 6%. This neural net is shown in figure 19.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	269	12
	Positive	14	137

The feed forward neural network using the full standardized data set had an accuracy of 94%, while the neural network trained using the standardized training set as input had an accuracy of 93%. These models had the highest accuracy of any model in the analysis.

Table 16: Confusion matrix for the selected best neural net trained on the full standardized data set. This model has a misclassification rate of 7%. This neural net is shown in figure 20.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	332	23
	Positive	13	164

Discussion:

Table 17: Comparison of the accuracy of the models built in the analysis of the pima dataset

Model	Accuracy
Full GLM	82%
Reduced GLM	81%
Full GAM	87%
Reduced GAM	89%
Default Binary Tree	89%
Optimal Binary Tree	91%
Default Random Forest	87%
Subsample Random Forest	87%
Full Data NN	79%
Full Standardized Data NN	84%
Best Standardized Data NN	84%
Full Data NN 8 nodes	93%
Training Data NN 8 nodes	94%

Of the models trained in the analysis, the model which performed with the greatest accuracy was the neural network with 8 hidden nodes, trained using the standardized training data set. The model did not calculate the out-of-bag error rate, but rather the observed results of the input data used to train the model. This is not an accurate representation of how these models would be deployed in practice. Additionally, the neural network does not offer interpretability, but rather is used to gain the highest predictive ability. There are no parameters in the model to interpret. Rather, there are a set of hidden nodes which apply an unknown model to the data, with differing weights applied to these nodes. This model does not allow an understanding of the nature of the data.

The model which performed with the highest accuracy on the test set was the optimized binary decision tree, which classified the observations in the test set with an accuracy of 91%. This model performs quite well and could be used in practice to predict the diabetic status of a pima woman in Arizona with the information included in the model. This model does not provide a very good understanding of the nature of the data either. We can see which variables are

important as they will appear higher in the tree, but we do not know the estimated effect that a unit change in the predictor will have on the response as we do in other models.

The reduced GLM provides the greatest interpretability of the model, as all the estimated coefficients are significant at the 10% level, and only one is insignificant at the 95% level. In exchange for this understanding of the nature of the data, the model sacrifices predictive capability.

The choice for which model to use depends on the objective of the user. If predictive capability is the end goal, then the neural network or binary decision tree is the best option. If interpretability and an understanding of the different relationships between the covariates and the response is important, then the GLMs are a better option. There are other modelling approaches which were not explored in this analysis. For example, a support vector machine may provide some insight on the data that these models did not.

References:

Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press.