Alejandro Naranjo

Assignment 7 Report

Table of Contents

Introduction

For this report there will be two different sections; the first section will be addressing the data analytics for both datasets, and the second section will address the different models that were built. At the end of the report there will be a conclusion that goes over the results and what on these models.

For the two different data sets, I will be using the dataset which contains obesity data based on eating and physical conditions (UCI Machine Learning Repository), and the second dataset that was chosen is the dataset with a decades worth of data on Diabetes (UCI). The reason for picking these two datasets is because I am interested in looking at health data and how we can use it to make predictions about someone's health. It's important to track this information so we can make early detections of illnesses or improve the type of health treatment people can receive by analyzing the data that has been collected.

Data Analysis

Diabetes Dataset

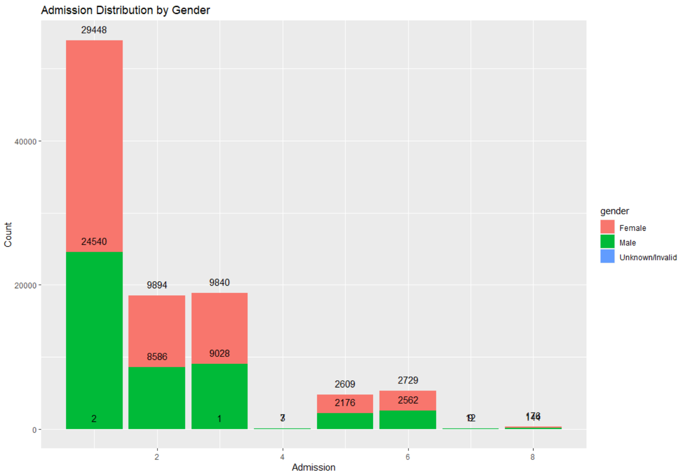


Figure 1 Distribution of Admission Id By Gender

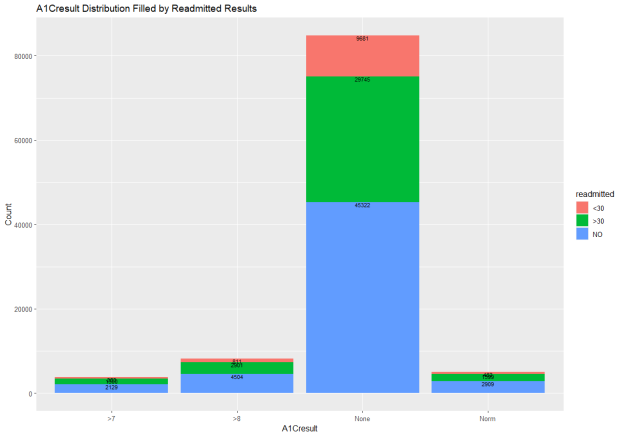


Figure 2 Distribution of A1CResult With Readmission Results

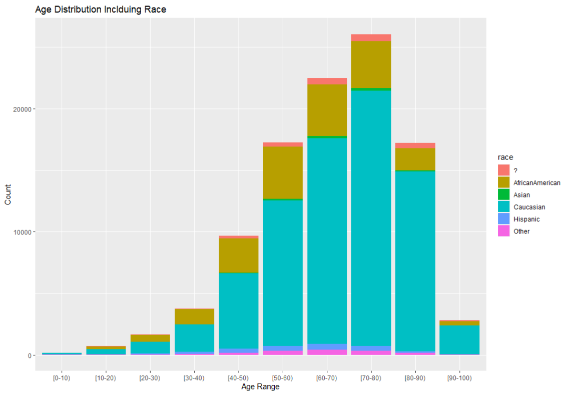


Figure 3 Age Distribution Including Race

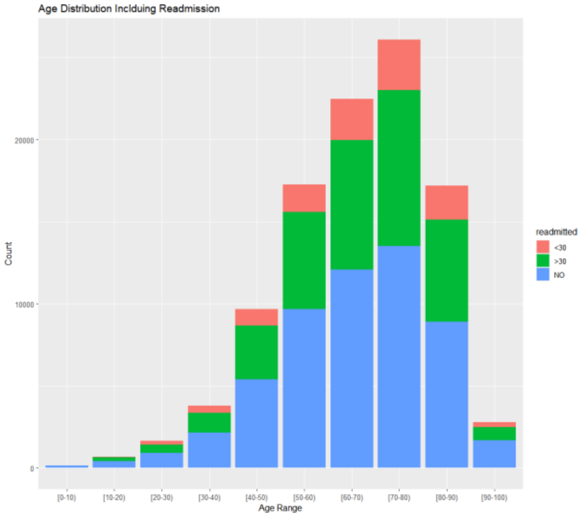
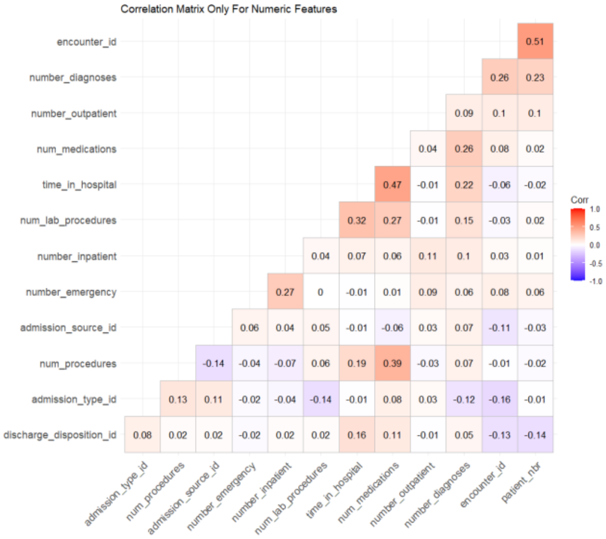


Figure 4 Age Distribution With Readmission



	admission_type_id	num_procedures	admission_source_id	number_emergency	number_inpatient	num_lab_procedures	time_in_hospital	num_medications	number_outpatient	number_diagnoses	encounter_id	patient_nbr
encounter_id												0.51
number_diagnoses											0.26	0.23
number_outpatient										0.09	0.1	0.1
num_medications									0.04	0.26	0.08	0.02
time_in_hospital								0.47	-0.01	0.22	-0.06	-0.02
num_lab_procedures							0.32	0.27	-0.01	0.15	-0.03	0.02
number_inpatient						0.04	0.07	0.06	0.11	0.1	0.03	0.01
number_emergency					0.27	0	-0.01	0.01	0.09	0.06	0.08	0.06
admission_source_id				0.06	0.04	0.05	-0.01	-0.06	0.03	0.07	-0.11	-0.03
num_procedures			-0.14	-0.04	-0.07	0.06	0.19	0.39	-0.03	0.07	-0.01	-0.02
admission_type_id		0.13	0.11	-0.02	-0.04	-0.14	-0.01	0.08	0.03	-0.12	-0.16	-0.01
discharge_disposition_id	0.08	0.02	0.02	-0.02	0.02	0.02	0.16	0.11	-0.01	0.05	-0.13	-0.14

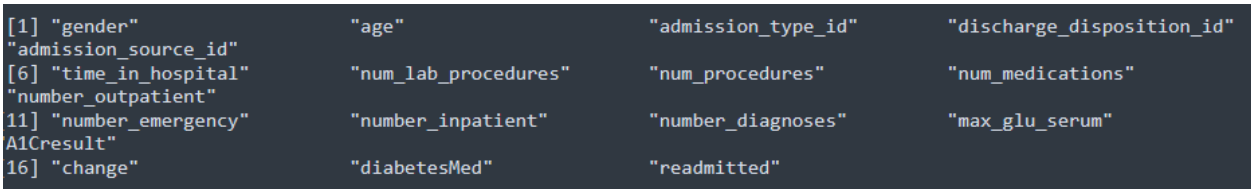
Figure 5 Correlation Matrix of Numeric Data Points

For the Diabetes data set there was a large data set which had over 50 different feature values. For these values we didn't need all of them and we could get rid of many. The first part of the process was to visualize and see what the data looked like. Looking at Figure 1, the purpose of this was took at what type of data entry made up the majority of the data set. The Admission Id for a patient was mapped to a numerical value and the following is what each represents; 1 is an emergency, 2 is urgent, 3 is elective, 4 is newborn, 5 is Not available, 6 is null, and 7 is trauma center. The data set was largely made up of emergency visits and the good part was neither sex group dominated in the visits for each ID group. The second graph, Figure 2, we are looking at the distribution from an *A1CResult* . What this feature represents is if a test was conducted to understand a person's glucose levels. Look at the distribution of the results we have a large number of "*None"* indicating the test was never conducted. This might play into effect when we are creating the models. Looking at the research paper associated to the data (Beata Strack, 2014) , the research goal was to see the impact of HbA1c results and how it affects a patient if they are readmitted. To us this is something important to consider and how it plays into our models.

For the next following figure, Figure 3, we look at the distribution of patients by age, but we also fill in by race. The reason for doing this is to get a better understand of the demographic of people and if race plays a critical role in determining who is getting admitted. Age is a clear indicator when it comes to patients being admitted, as we look at the histogram it is skewed to the right. This indicates those who are younger are not prone to having to make a medical trip that is diabetes related. Those who are older and have been diagnosed with diabetes are more prone to coming to the hospital or being readmitted. If we break it down even farther for our age groups, we can fill in our histogram based off readmittance. Looking at Figure 4 we can see that as the age gets older we see more and more readmittance, that is something we don't see as much with the younger groups.

For the last figure we look at the correlation between the numerical values so we can have an understanding which variable might affect our models. There isn't much which sticks out to us, there are a few features which have a positive correlation, but those variables will be removed.

For the dataset that we will be working with much of the metadata about the patient has been removed as well as the diagnosis the patient received, and the list of medications a patient is taking. I believe those features won't play a huge role for us. A few other features that have been removed were *Weight, Payer Code, and Medical Specialty;* the reason behind this was because each feature contained more than 50% of missing values. The next step in preparing the data was to convert the categorical values to numeric values. I did this by mapping each unique value to an increasing number starting at 1. I will be trying to predict if a patient is going to be readmitted in the future. The following figure depicts which feature variables I will be using for the next three models.



```
[1] "gender"                   "age"                      "admission_type_id"        "discharge_disposition_id"
"admission_source_id"
[6] "time_in_hospital"         "num_lab_procedures"       "num_procedures"           "num_medications"
"number_outpatient"
11] "number_emergency"         "number_inpatient"         "number_diagnoses"         "max_glu_serum"
A1Cresult"
16] "change"                   "diabetesMed"              "readmitted"
```

Figure 6 Reduced Features

Obesity Dataset

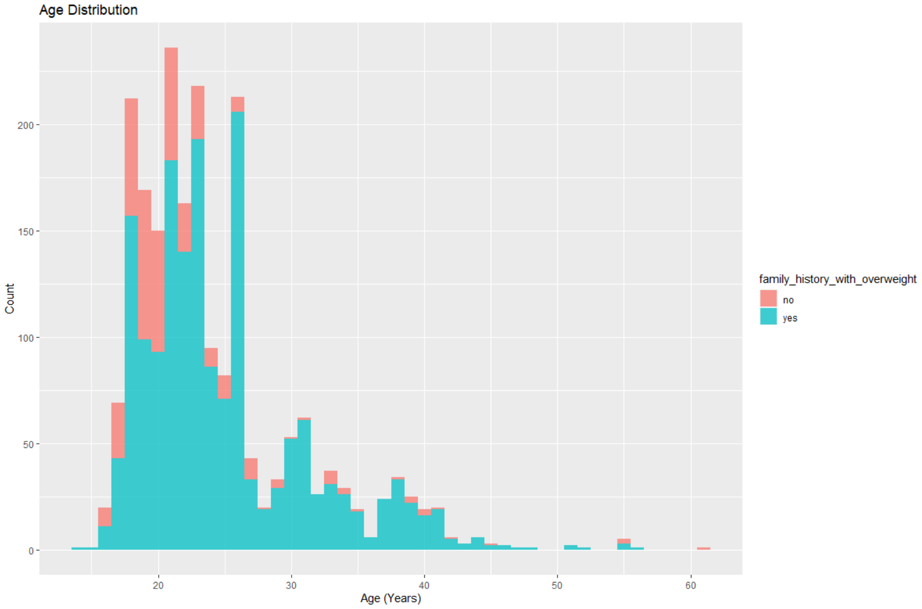


Figure 7 Age Distribution with History of Family Obesity

Looking at the Figure above we are looking at the age distribution and we can see the distribution is skewed to the left side. Our data consists of younger people are or ranging from children to younger adults. To have a better understanding of the data I filled in the histogram based off Family history to understand what effect that might have on those whose family has a history of obesity. Age and family history can have a large impact on the results of the model.

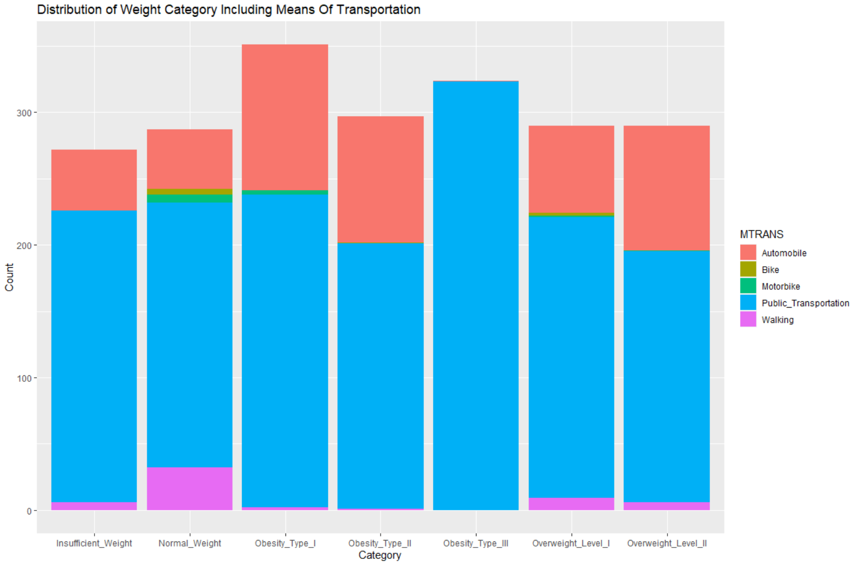


Figure 8 Distribution of Weight with Means of Transportation

Continuing to look at the data I wanted to see if there was a larger number of data points that have been categorized in one body weight category over the other. Across all 7 categories there isn't one that sticks out or contains most of the data points. The only group that sticks out is *Obesity Type 1,* this could be a small cutoff between *Normal Weight* and *Obesity Type II.* Another interesting thing while looking at the chart is as the *Obesity* categories increase, we see a decrease in *Walking, Bike usage, and Motorbike Usage.* To us this is to take note of as it can play a role in helping determine which category each person falls into.

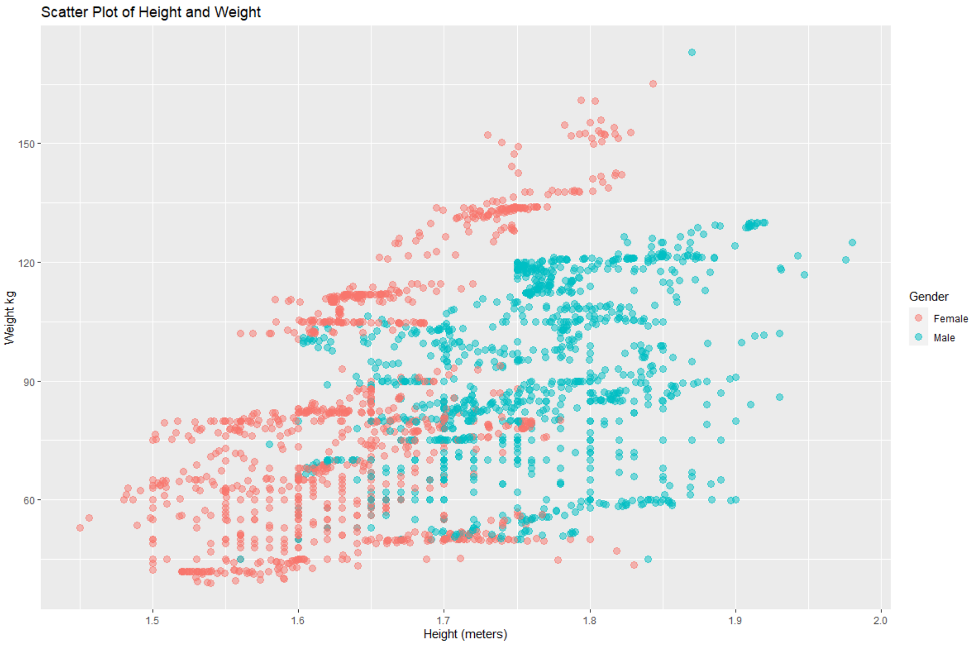


Figure 9 Scatter Plot of Weight and Height, Filled by Sex

Looking at the scatter plot we can see two different distinct groups based off of sex; however, there is a large enough overlap that will make it hard to cluster. One of the models that will be used is a clustering algorithm to see if we can separate the two and predict a person Sex based off only *Height and Weight.*

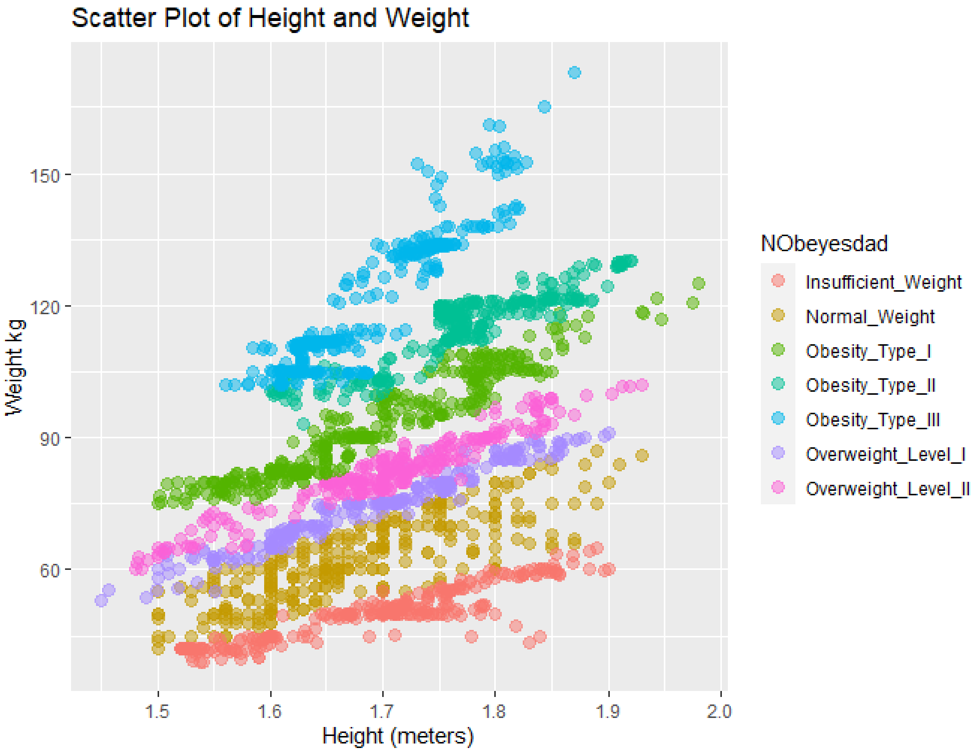


Figure 10 Scatter Plot Height and Weight Filled by Obesity Category

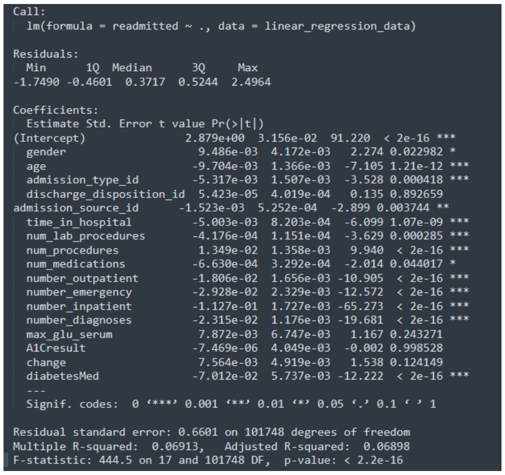
The same as above we can see the distinct cut offs across each section and as weight increases so does the *NObeyesdad* categorization. An important thing to note is how the cut off for certain categories increase based off height. Height will play an important role for the last model that will take in all the data to predict where a person lies. As for the second model I will be building a clustering model to see if we could cluster the data into 7 clusters and predict a person's category based off height and weight. Of course, we know the number of labels so we know what to expect but that is why we are also building the Random Forest Model.

Modeling

*** Each Model was built using a 70/30 train/test split. ***

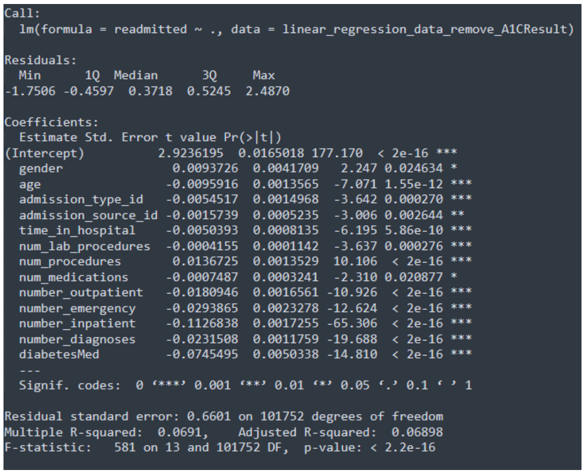
Diabetes Dataset

Linear Regression: Predicting Readmittance



```
 Call:
   lm(formula = readmitted ~ ., data = linear_regression_data)

 Residuals:
   Min      1Q  Median      3Q     Max
 -1.7490 -0.4601  0.3717  0.5244  2.4964

 Coefficients:
   Estimate Std. Error t value Pr(>|t|)
 (Intercept)               2.879e+00  3.156e-02  91.220  < 2e-16 ***
   gender                    9.486e-03  4.172e-03   2.274 0.022982 *
   age                      -9.704e-03  1.366e-03  -7.105 1.21e-12 ***
   admission_type_id        -5.317e-03  1.507e-03  -3.528 0.000418 ***
   discharge_disposition_id  5.423e-05  4.019e-04   0.135 0.892659
 admission_source_id      -1.523e-03  5.252e-04  -2.899 0.003744 **
   time_in_hospital         -5.003e-03  8.203e-04  -6.099 1.07e-09 ***
   num_lab_procedures       -4.176e-04  1.151e-04  -3.629 0.000285 ***
   num_procedures            1.349e-02  1.358e-03   9.940  < 2e-16 ***
   num_medications          -6.630e-04  3.292e-04  -2.014 0.044017 *
   number_outpatient        -1.806e-02  1.656e-03 -10.905  < 2e-16 ***
   number_emergency         -2.928e-02  2.329e-03 -12.572  < 2e-16 ***
   number_inpatient         -1.127e-01  1.727e-03 -65.273  < 2e-16 ***
   number_diagnoses         -2.315e-02  1.176e-03 -19.681  < 2e-16 ***
   max_glu_serum             7.872e-03  6.747e-03   1.167 0.243271
   A1Cresult                -7.469e-06  4.049e-03  -0.002 0.998528
   change                    7.564e-03  4.919e-03   1.538 0.124149
   diabetesMed              -7.012e-02  5.737e-03 -12.222  < 2e-16 ***
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 0.6601 on 101748 degrees of freedom
 Multiple R-squared:  0.06913,   Adjusted R-squared:  0.06898
 F-statistic: 444.5 on 17 and 101748 DF,  p-value: < 2.2e-16
```

Figure 11 Linear Regression Before Adjusting Variables



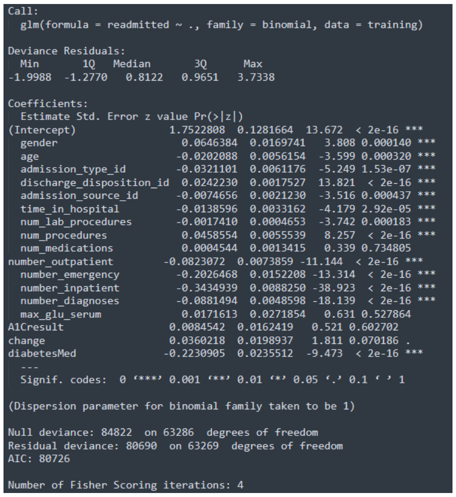
```
Call:
  lm(formula = readmitted ~ ., data = linear_regression_data_remove_A1CResult)

Residuals:
  Min      1Q  Median      3Q     Max
-1.7506 -0.4597  0.3718  0.5245  2.4870

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.9236195  0.0165018 177.170  < 2e-16 ***
  gender               0.0093726  0.0041709   2.247 0.024634 *
  age                 -0.0095916  0.0013565  -7.071 1.55e-12 ***
  admission_type_id   -0.0054517  0.0014968  -3.642 0.000270 ***
  admission_source_id -0.0015739  0.0005235  -3.006 0.002644 **
  time_in_hospital    -0.0050393  0.0008135  -6.195 5.86e-10 ***
  num_lab_procedures  -0.0004155  0.0001142  -3.637 0.000276 ***
  num_procedures       0.0136725  0.0013529  10.106  < 2e-16 ***
  num_medications     -0.0007487  0.0003241  -2.310 0.020877 *
  number_outpatient   -0.0180946  0.0016561 -10.926  < 2e-16 ***
  number_emergency    -0.0293865  0.0023278 -12.624  < 2e-16 ***
  number_inpatient    -0.1126838  0.0017255 -65.306  < 2e-16 ***
  number_diagnoses    -0.0231508  0.0011759 -19.688  < 2e-16 ***
  diabetesMed         -0.0745495  0.0050338 -14.810  < 2e-16 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6601 on 101752 degrees of freedom
Multiple R-squared:  0.0691,    Adjusted R-squared:  0.06898
F-statistic:   581 on 13 and 101752 DF,  p-value: < 2.2e-16
```

Figure 12 Linear Regression After Adjusting Features

For this model I will be using the feature variables described in the previous section to predict if someone were to get readmitted. The result of the first model built had a 0.069 R-Squared value which is not good at all. This means that the values used is difficult to predict the dependent variable. For the second model on the right, three feature variables were removed. Those variables *: max_glu_serum, change, and A1CResult;* one of the features I believed would have an impact on the model did not at all with a P value of 0.998. By removing the values some of the other features became more significant but the model was performing very badly.

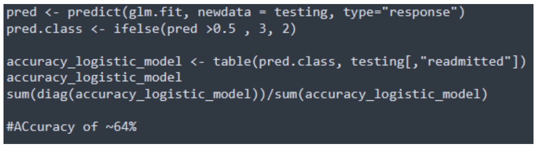
Looking at the linear regression for

Logistic Regression: Predicting Readmittance After 30 days



```
Call:
  glm(formula = readmitted ~ ., family = binomial, data = training)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-1.9988  -1.2770   0.8122   0.9651   3.7338

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.7522808  0.1281664  13.672  < 2e-16 ***
  gender                    0.0646384  0.0169741   3.808 0.000140 ***
  age                      -0.0202088  0.0056154  -3.599 0.000320 ***
  admission_type_id        -0.0321101  0.0061176  -5.249 1.53e-07 ***
  discharge_disposition_id  0.0242230  0.0017527  13.821  < 2e-16 ***
  admission_source_id      -0.0074656  0.0021230  -3.516 0.000437 ***
  time_in_hospital         -0.0138596  0.0033162  -4.179 2.92e-05 ***
  num_lab_procedures       -0.0017410  0.0004653  -3.742 0.000183 ***
  num_procedures            0.0458554  0.0055539   8.257  < 2e-16 ***
  num_medications           0.0004544  0.0013415   0.339 0.734805
number_outpatient        -0.0823072  0.0073859 -11.144  < 2e-16 ***
  number_emergency         -0.2026468  0.0152208 -13.314  < 2e-16 ***
  number_inpatient         -0.3434939  0.0088250 -38.923  < 2e-16 ***
  number_diagnoses         -0.0881494  0.0048598 -18.139  < 2e-16 ***
  max_glu_serum             0.0171613  0.0271854   0.631 0.527864
A1Cresult                 0.0084542  0.0162419   0.521 0.602702
change                    0.0360218  0.0198937   1.811 0.070186 .
diabetesMed              -0.2230905  0.0235512  -9.473  < 2e-16 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 84822  on 63286  degrees of freedom
Residual deviance: 80690  on 63269  degrees of freedom
AIC: 80726

Number of Fisher Scoring iterations: 4
```

Figure 13 Logistic Regression Summary

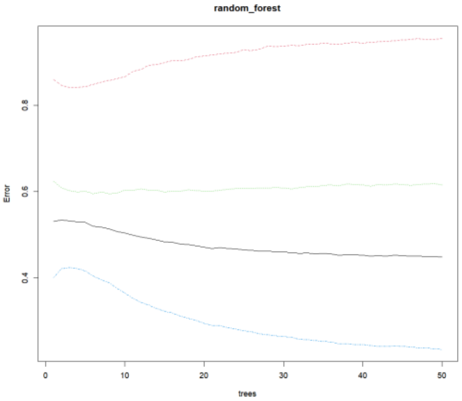


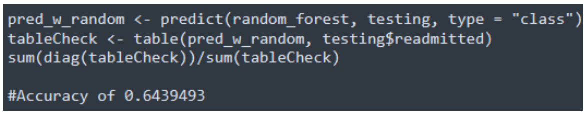
```
pred <- predict(glm.fit, newdata = testing, type="response")
pred.class <- ifelse(pred >0.5 , 3, 2)

accuracy_logistic_model <- table(pred.class, testing[,"readmitted"])
accuracy_logistic_model
sum(diag(accuracy_logistic_model))/sum(accuracy_logistic_model)

#ACcuracy of ~64%
```

Figure 14 Logistic Regression Predictions and Results

For the logistic regression because we are only able to achieve a binary classification, I removed one of the data points. The data points that were removed were those that were readmitted under 30 days. The reason for this is because it made up a small portion of the data; this can be seen in Figure 2. For this model we are only predicting if they would get readmitted after 30 days or if they would never come back. The results of the modeling building are like that of the linear regression as some of the features that were not significant for that model are also not significant for this model. Looking at the prediction results we can see that the model had a 64% model accuracy. Those results are much better than the linear regression model.

Random Forest: Predicting Readmittance





```
pred_w_random <- predict(random_forest, testing, type = "class")
tableCheck <- table(pred_w_random, testing$readmitted)
sum(diag(tableCheck))/sum(tableCheck)

#Accuracy of 0.6439493
```

Figure 15 Random Forest Error

Figure 16 Random Forest Prediction Results

With the random forest model, look at the error chart we can see that the error for the various model build is not going down. The only error value that was reduced was the light blue line at the bottom. This is an indication that the data was not good enough to build a model to successfully predict if a patient would be readmitted. From the prediction results my statement is support as we only get a 64% accuracy and this is including all three different results.

Obesity Dataset

KMeans Clustering: Predicting Sex Using Height and Weight

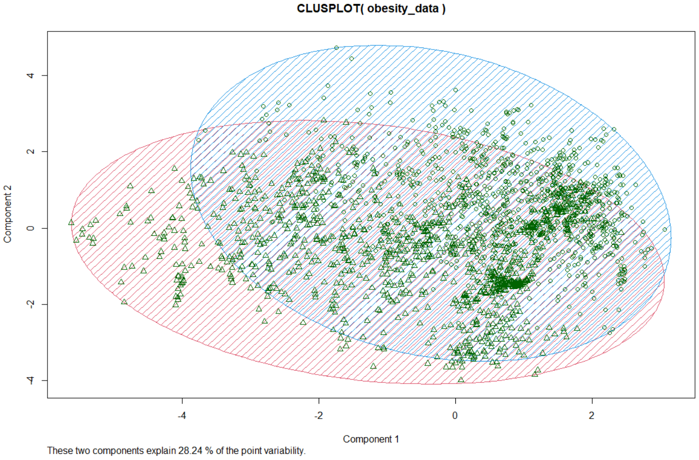
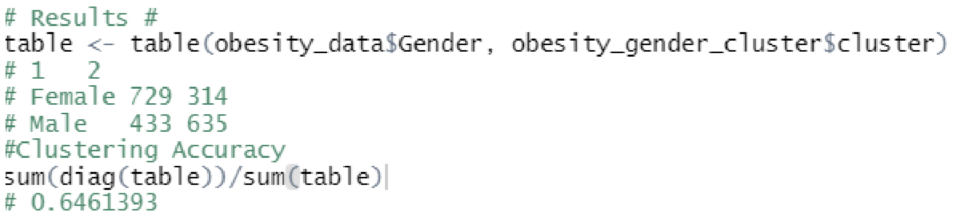


Figure 17 Clustering With Two Clusters

The reason this model was built was due to what was seen during the data discovery phase of the assignment. Looking at Figure 9, we see a scatter plot of height and weight; however, if we an additional element to color code the points we can see that there are two clusters. The only concern with this is the large overlap in the middle of the two groups. From the lower and upper ends, the two groups are separable but, in the middle, it'll be hard to cluster. For this model we set the *nstart* equal to 20 and set the clusters to 2 and only used height and weight to train. In Figure 17 we see the result of the clustering. There is a very large overlap across both clusters and would lead to a low accuracy score but one that is good and predicting the lower and upper ends.



```
# Results #
table <- table(obesity_data$Gender, obesity_gender_cluster$cluster)
# 1   2
# Female 729 314
# Male   433 635
#Clustering Accuracy
sum(diag(table))/sum(table)
# 0.6461393
```

Figure 18 Resutls of Kmeans Cluster =2

Looking at the Figure above we see that we have an accuracy score of 64%.

Random Forest: Predicting *NObeyesdad* Category

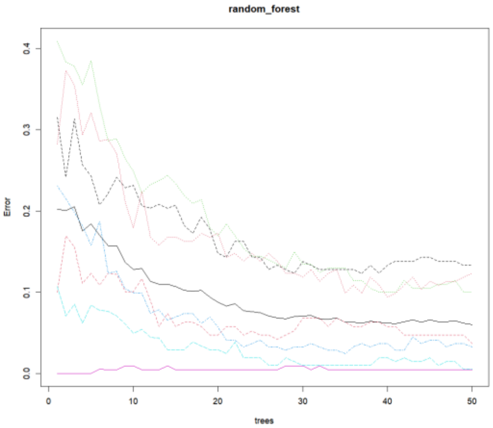
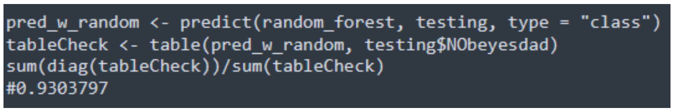


Figure 20 Randomforest Results

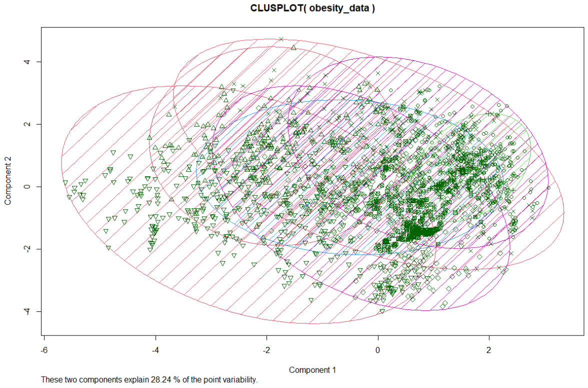


```
pred_w_random <- predict(random_forest, testing, type = "class")
tableCheck <- table(pred_w_random, testing$NObeyesdad)
sum(diag(tableCheck))/sum(tableCheck)
#0.9303797
```

Figure 19 Randomforest Prediction Results

For the Obesity data by selecting *NObeyesdad* as the dependent variable was a good choice. All the data was included in the model; but, the non-numerical data was mapped to numeric values so it could be processed easier. Looking at the plot of the random forest we see the error rate trending down and eventually flatting out after about 20 trees. Out of all 6 models that have been made this is the best model for predicting the dependent variable. Looking at Figure 19 we see an accuracy of 93% which is good. Another way to improve the data is to take it a step further and introduce principle component analysis. The dataset contain strong enough data to predict what category a person may fall into for their weight and health.

KMeans Clustering: Predicting *NObeyesdad* Using Only Height and Weight



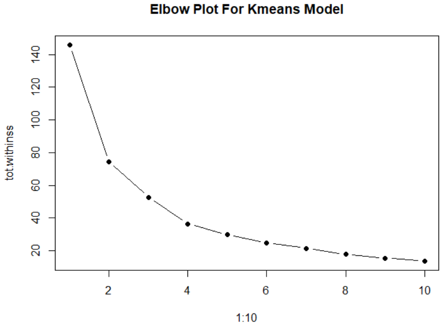
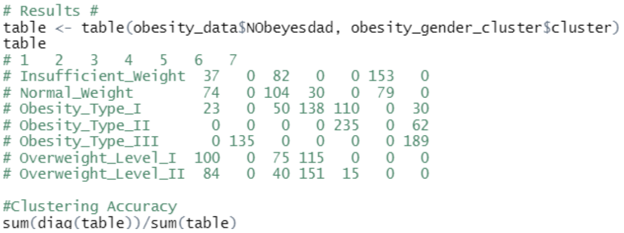


Figure 21 Seven Clusters Using Height and Weight For The Model

Figure 22 Elbow Plot For Kmeans



```
# Results #
table <- table(obesity_data$NObeyesdad, obesity_gender_cluster$cluster)
table
# 1   2   3   4   5   6   7
# Insufficient_Weight  37   0  82   0   0 153   0
# Normal_Weight        74   0 104  30   0  79   0
# Obesity_Type_I       23   0  50 138 110   0  30
# Obesity_Type_II       0   0   0   0 235   0  62
# Obesity_Type_III      0 135   0   0   0   0 189
# Overweight_Level_I  100   0  75 115   0   0   0
# Overweight_Level_II  84   0  40 151  15   0   0

#Clustering Accuracy
sum(diag(table))/sum(table)
```

Figure 23 Results of 7 Clusters Accuracy

For this model the goal was to see if we can predict which category each person belongs to. The reason for this is the scatter plot in Figure 10. I was wondering if we could use Weight and Height to create clusters that can be used to predict the category a person falls into. This model did not perform that well as the clusters that were formed were overlapping. The reason for this is because there isn't enough data to create distinct groups. There are two feature variables in the model, both height and weight. We can see the groups in the scatter plot, but we don't have enough data in the model to create those groups. The result of this is we get a 4% accuracy rate, which is low.

Conclusion

Looking at both of these datasets we can see that there are some strengths and weaknesses when it comes to creating models. For the Diabetes Dataset the method of removing a large of feature and transforming all the non-numeric data to numeric data could have had a negative impact to the models. For the future exploring the rest of these features that were removed will be important in improving the fit of the models. The linear regression and logistic regression models had weak fits but that could very much change by exploring other dependent variables or introducing PCAs. Choosing *readmitted* as the dependent variable was not the right choice as the data didn't strong support it.

For the obesity data set choosing *NObeyesdad* as the dependent variable for the Random Forest model had a good fit. Another technique which had a positive impact on it was including all the feature variables and properly mapping all the non-numeric values to numeric values. We can see the success of it by having a 93% accuracy score. As for the other two other models built using the Obesity data lacked vital information. Each model was trained using two features (height and weight), the difference was trying to predict two different values. In one model we predicted Sex and in the other we wanted the

NObeyesdad. Each model needed more information to successfully create the necessary clusters to identify the dependent variable. The model predicting Sex had an accuracy of 63% which is okay but the other model only had a 4% accuracy; a clear indication that there was too many clusters.

Using a Random Forest model for predicting *NObeyesdad* was very successful and good while the model lacked with the diabetes dataset. This means adjustments to how the data was prepared need to be made and other adjustments should occur to ensure a good fit of the model.

References

Beata Strack, 1. P. (2014). *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.*

UCI. (n.d.). *Diabetes 130-US hospitals for years 1999-2008 Data Set.* Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

UCI Machine Learning Repository. (n.d.). *UCI Machine Learning Repository.* Retrieved from UCI: https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

Libraries Used:

- ggplot2, cluster, caret, randomForest, ggcorrplot**,**