Causes of Diabetes in the Pima Indian Adult Female Population

By: Avery Gump, Kezia Regis, and Luis Santin

Presentation Outline

1. Problem Description
2. Problem Statement and Significance
3. Existing and/or Possible Analysis Methods and Metrics
4. Developed Analysis Methods and Metrics
5. Code and simulation results
6. Conclusion and Future Research
7. References

Problem Description

According to the CDC's National Diabetes Statistics Report, over 28.7 million people have been diagnosed with diabetes. Diabetes is a national epidemic with signs of becoming an international problem. To promote awareness, The World Health Organization (WHO), estimated that by 2030, over 336 million people will be diagnosed with diabetes.

Problem Statement and Significance

Problem Statement: What variable(s) can we determine as the main factor for diabetes in the female population of the Pima tribe?

Significance: The Pima (or Akimel O'odham), "River People," are a group of Native Americans living in an area consisting of what is now central and southern Arizona, as well as northwestern Mexico.

They have various environmentally based health issues related to the decline of their traditional economy and farming.The Pima tribe have been the subject of intensive study of diabetes. They have the highest prevalence of type 2 diabetes in the world, much more than is observed in other U.S. populations.

Possible Analysis Methods and Metrics

1. Probability Package in R
    a. Discrete Random Variables and Conditional Probability functionality
2. Binomial
    a. Cumulative Probability
3. LASSO
    a. Least Absolute Shrinkage and Selection Operator
    b. Linear regression model
4. Ridge Regression
    a. Shrinkage estimators that use bias for approximations

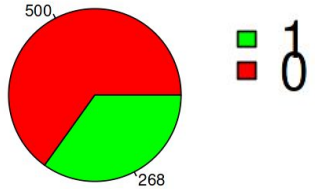
Developed Analysis Methods and Metrics

1. Statistical Overview
    a. Mean
    b. Median
    c. Std
    d. Min/Max
    e. Variance
    f. 1st, 2nd, 3rd Quartiles
    g. F-test (ANOVA/Chi-square)
2. Probability Density Plots
3. Box Plots with outliers
4. Correlation Matrices
5. Fligner-Killeen test
6. Random Forest Classifier

Statistical Overview

	Variable	std.	variance	mean	median	min	max	1st Quartile	2nd Quartile	3rd Quartile	std.	variance	mean	median	min	max	1st Quartile	2nd Quartile	3rd Quartile	Test
1	Outcome	0									1									
2	Pregnancies	3	9.103	3.298	2	0	13	1	2	5	3.7	13.997	4.866	4	0	17	1.75	4	8	F=39.67***
3	Glucose	26.1	683.362	109.98	107	0	197	93	107	125	31.9	1020.139	141.257	140	0	199	119	140	167	F=213.162***
4	BloodPressure	18.1	326.275	68.184	70	0	122	62	70	78	21.5	461.898	70.825	74	0	114	66	74	82	F=3.257*
5	SkinThickness	14.9	221.711	19.664	21	0	60	0	21	31	17.7	312.572	22.164	27	0	99	0	27	36	F=4.304**
6	Insulin	98.9	9774.345	68.792	39	0	744	0	39	105	138.7	19234.673	100.336	0	0	846	0	0	167.25	F=13.281***
7	BMI	7.7	59.134	30.304	30.05	0	57.3	25.4	30.05	35.3	7.3	52.751	35.143	34.25	0	67.1	30.8	34.25	38.775	F=71.772***
8	DiabetesPedigreeFunction	0.3	0.089	0.43	0.336	0.078	2.329	0.23	0.336	0.562	0.4	0.139	0.55	0.449	0.088	2.42	0.262	0.449	0.728	F=23.871***
9	Age	11.7	136.134	31.19	27	21	81	23	27	37	11	120.303	37.067	36	21	70	28	36	44	F=46.141***

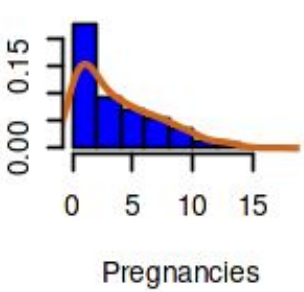
n = 768
Outcome(0) = No Diabetes
Outcome(1) = Has Diabetes
Diabetes Pedigree Function: A measure of family history of diabetes

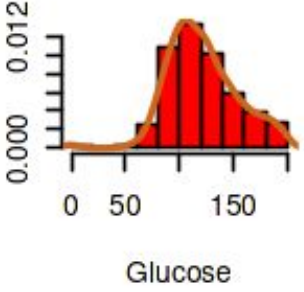


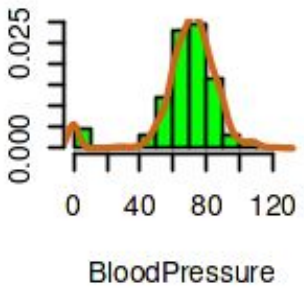
Statistical Overview Code

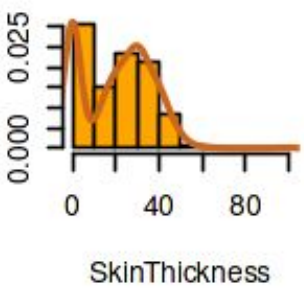
```
library(tidyverse)
## Read Data And Give Summary:
diabetes_df <- read_csv("../input/diabetes-dataset/diabetes.csv")
library(vtable)

st_tab <- st(diabetes_df, out='return',
   summ = list(
     c('sd(x)','min(x)','max(x)','median(x)','var(x)','quantile(x,.25)','quantile(x,.50)','quantile(x,.75)')
   ),
   summ.names = list(
     c('std.','min','max','median','variance','1st Quartile','2nd Quartile','3rd Quartile')
   ),
    group = "Outcome",
    group.test=TRUE,
    )
library(gridExtra)
png("test.png", height = 50*nrow(st_tab), width = 200*ncol(st_tab))
grid.table(st_tab)
dev.off()
st_tab
```

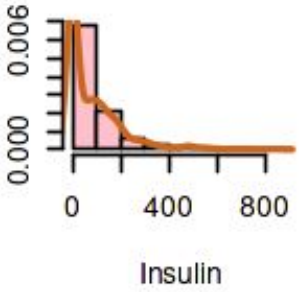
Probability Density Plots

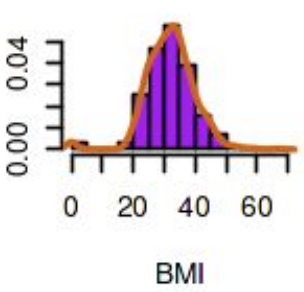


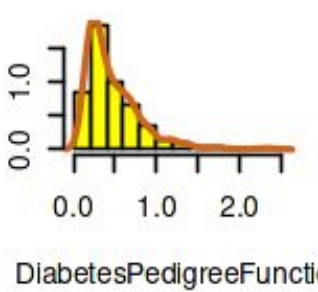


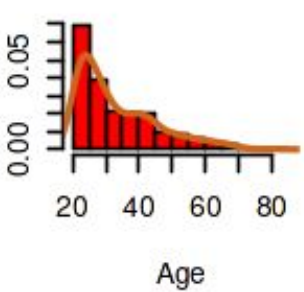








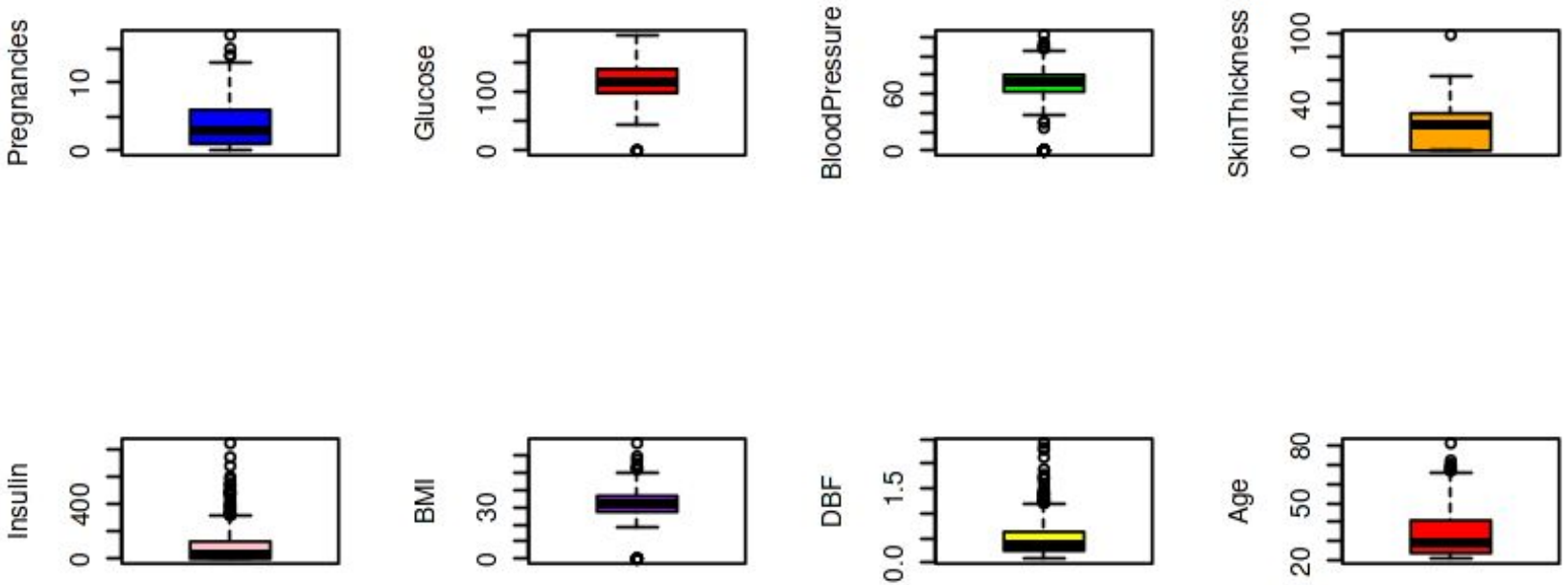




Probability Density Plots Code

```
## Histograms
par(mfrow=c(4,4))
colors_col = c("blue","red","green","orange","pink","purple","yellow","red","green")
counter = 1
for(i in colnames(diabetes_df)){
if(i == "Outcome"){next}

hist(diabetes_df[[i]],
    col=colors_col[counter],
    border="black",
    prob = TRUE,
    xlab = i,
     ylab = "",
    main = "")

lines(density(diabetes_df[[i]]),
    lwd = 2,
    col = "chocolate3")
counter = counter +1
    }
```

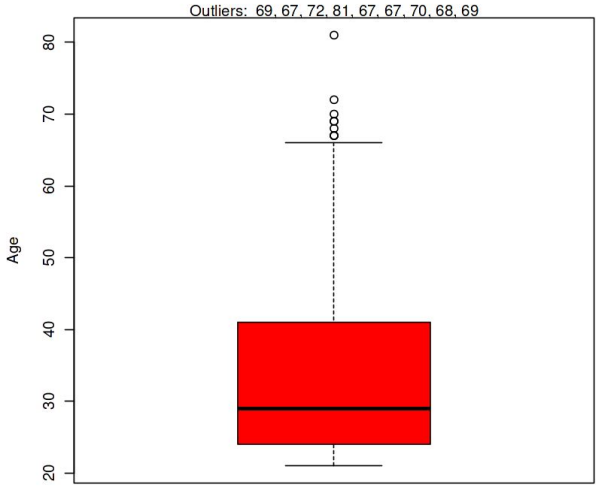
Box Plots With Outliers

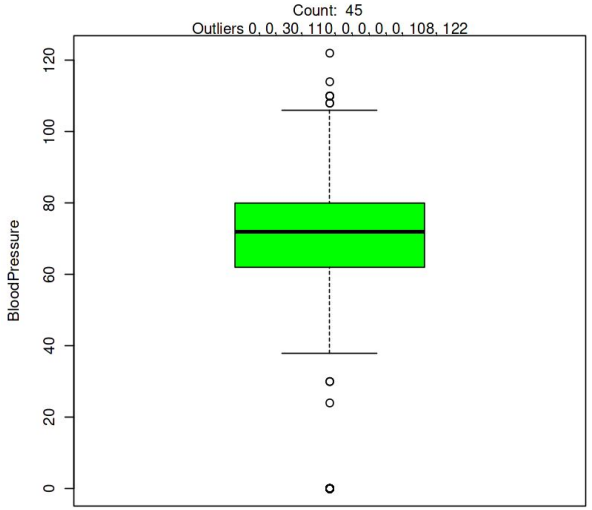


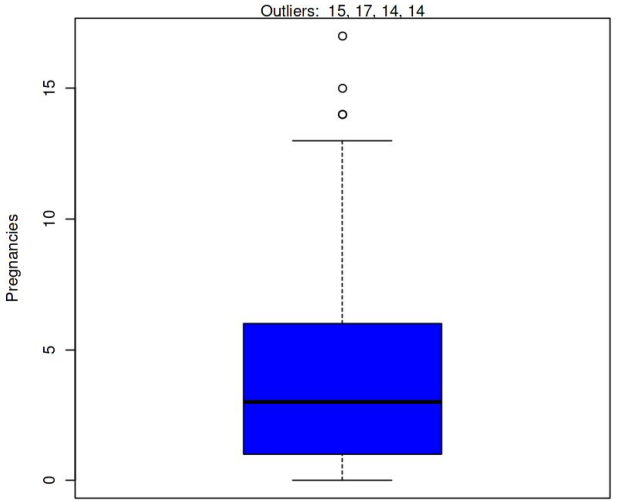
Box Plots With Outliers Code

```
## Boxplots WITHOUT OUTLIERS/ COUNT
par(mfrow=c(4,4))
colors_col = c("blue","red","green","orange","pink","purple","yellow","red","green")
counter = 1
for(i in colnames(diabetes_df)){
if(i == "Outcome"){next}

out <- boxplot.stats(diabetes_df[[i]])$out
if(i == "DiabetesPedigreeFunction"){
    boxplot(diabetes_df[[i]],
  ylab = "DBF",
  main = "",
  col = colors_col[counter]

)
}
else{
boxplot(diabetes_df[[i]],
  ylab = i,
  main = "",
  col = colors_col[counter]

)}
counter = counter +1
    }
```

Single Plot Examples - Age, Blood Pressure, Pregnancies





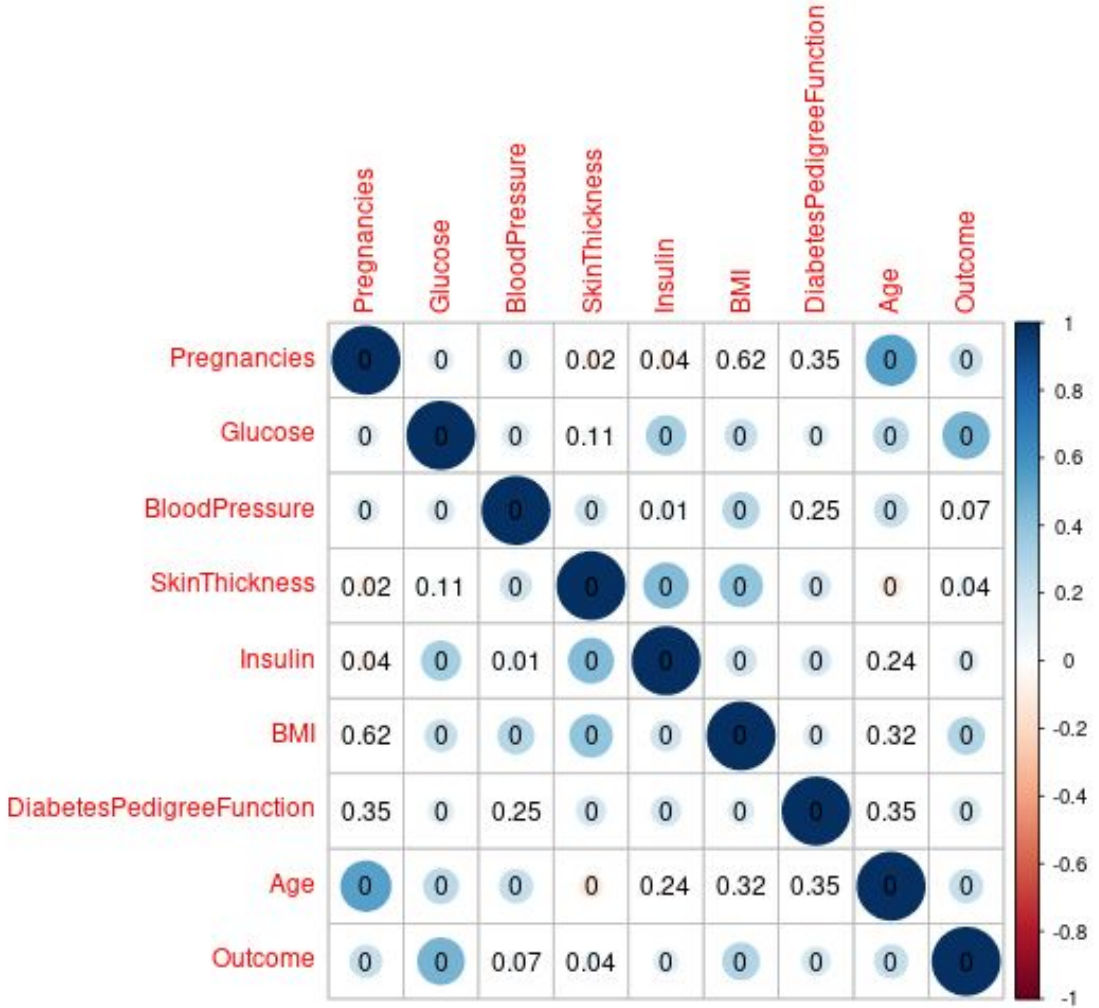


Single Plot Examples - Age, Blood Pressure, Pregnancies

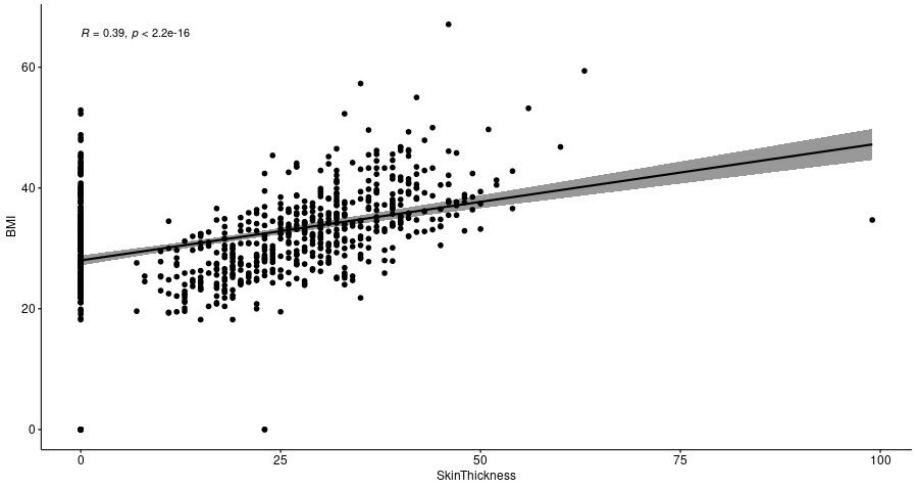
```
## Boxplots WITH OUTLIERS/ COUNT
#par(mfrow=c(4,4))
colors_col = c("blue","red","green","orange","pink","purple","yellow","red","green")
counter = 1
for(i in colnames(diabetes_df)){
if(i == "Outcome"){next}

out <- boxplot.stats(diabetes_df[[i]])$out
boxplot(diabetes_df[[i]],
  ylab = i,
  main = "",
  col = colors_col[counter]

)
counter = counter +1
if(length(out) > 10){
    if(i == "DiabetesPedigreeFunction"){mtext(paste("Count: ", paste(length(out),collapse = ", "), "\nOutliers", paste(out[1:5],collapse=", ")))}
    else{
    mtext(paste("Count: ", paste(length(out),collapse = ", "), "\nOutliers", paste(out[1:10],collapse=", ")))}
}
#print(length(out))

else{
mtext(paste("Outliers: ", paste(out, collapse = ", "))) }
}
```

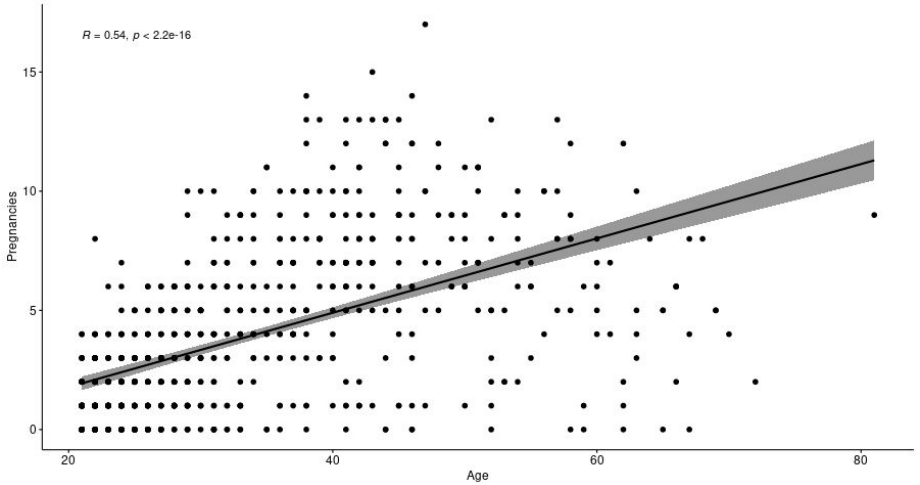
Correlation Matrix

1. Pearson Correlation Coefficient Used
2. H0: true correlation = 0
3. H1: true correlation != 0
4. Numbers in boxes are p-values
5. Circles show which random variables are significantly correlated



Example Correlation





Correlation Code

```r
## Correlation matix
cor_mat <- cor(diabetes_df)
testRes = cor.mtest(diabetes_df, conf.level = 0.95, method=c("pearson","kendall","spearman"))
corrplot(cor_mat, p.mat = testRes$p, insig = 'p-value', sig.level = -1)
```

```r
#Scatter Plots
ggscatter(diabetes_df, x = "Age", y = "Pregnancies",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Age", ylab = "Pregnancies")

ggscatter(diabetes_df, x = "SkinThickness", y = "BMI",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "SkinThickness", ylab = "BMI")
ggscatter(diabetes_df, x = "Age", y = "Outcome",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Age", ylab = "Outcome")
```

Hypothesis Testing(1)

1. First we tried t-test for our hypothesis testing in R. For example given
   a. H0: mean BMI for people 65 or older is the same as the mean BMI for people under this age
   b. H1: mean BMI for people 65 or older is less than the mean BMI for people over this age.
2. Because our p-value is .04212 < .05 we reject the null hypothesis. This makes sense as people lose body mass as they age [3]

```
data:  diabetes_df$BMI[diabetes_df$Age >= 65]
t = -1.8492, df = 15, p-value = 0.04212
alternative hypothesis: true mean is less than 32.08418
95 percent confidence interval:
     -Inf 31.85562
sample estimates:
mean of x
  27.6875
```

Hypothesis Testing(2)

1. What about variance? Our hypothesis was older people will also have less variance in their BMI.
2. F-test
   a. Problem - Need normal distributions
3. Shapiro-Wilk Test
   a. For the Shapiro-Wilk test because our p-value for the two distributions we are trying to compare is < .05, neither will fit the the normal distribution with 95% confidence.

```
        Shapiro-Wilk normality test

data:  diabetes_df$BMI[diabetes_df$Age >= 65]
W = 0.86553, p-value = 0.02325

        Shapiro-Wilk normality test

data:  diabetes_df$BMI[diabetes_df$Age < 65]
W = 0.95207, p-value = 6.415e-15
```

Hypothesis Testing(3)

1. This led us to the Fligner-Killeen test, which tests homogeneity of variances in distributions. For this test, the null hypothesis is that variances in the two groups are the same.
2. Two tests:
   a. Diabetes vs No Diabetes
      i. BP, Insulin, BMI - Homogeneous
   b. Age >= 50 vs Age < 50
      i. Pregnancies, Glucose, BMI – Homogenous

	Outcome	Over50
Pregnancies	0.00000	0.42428
Glucose	0.00000	0.78711
BloodPressure	0.33763	0.01530
SkinThickness	0.00000	0.00114
Insulin	0.36537	0.01221
BMI	0.05022	0.60592
DiabetesPedigreeFunction	0.00000	0.04201
Age	0.00035	0.03461

Hypothesis Testing

```r
## Hypothesis testing
library(stats)
## Tests for Groups Over 50
hypothesis_df = diabetes_df
hypothesis_df$Over50 <- as.factor(ifelse(hypothesis_df$Age>=50,1,0))
hypothesis_df

## Check to make sure we have enough in our >age category
slices <- c(sum(hypothesis_df$Over50 == 1),sum(hypothesis_df$Over50 == 0))
print(slices)

#Hypothesis - people who are older have less variation in their BMI

## T-test for means, we want variance
t.test(diabetes_df$BMI[diabetes_df$Age >= 65],mu = mean(diabetes_df$BMI[diabetes_df$Age < 65]),alternative="less")

## first test if the distributions are normal
shapiro.test(diabetes_df$BMI[diabetes_df$Age >= 65])
shapiro.test(diabetes_df$BMI[diabetes_df$Age < 65])

## F-test -- Need Normal Distributions
#var.test(diabetes_df$BMI[diabetes_df$Age >= 65],diabetes_df$BMI[diabetes_df$Age < 65],"less")

## Fligner Test -- Over 50
fligner.test(BMI ~ Over50, hypothesis_df)
fligner.test(Pregnancies ~ Over50, hypothesis_df)
fligner.test(Glucose ~ Over50, hypothesis_df)
fligner.test(Age ~ Over50, hypothesis_df)
fligner.test(DiabetesPedigreeFunction ~ Over50, hypothesis_df)


## Fligner Test -- Diabetes Outcome
fligner.test(BMI ~ Outcome, hypothesis_df)
fligner.test(Pregnancies ~ Outcome, hypothesis_df)
fligner.test(Glucose ~ Outcome, hypothesis_df)
fligner.test(Age ~ Outcome, hypothesis_df)
fligner.test(DiabetesPedigreeFunction ~ Outcome, hypothesis_df)
```

```r
flignerTestDf <- data.frame(Outcome=numeric(),Over50=numeric())

for(i in colnames(hypothesis_df)){
    if(i == "Outcome" || i == "Over50"){next}
    ## Calculate Fligner Test for Outcome and Over50 For the column
    over50F <- fligner.test(hypothesis_df[[i]],hypothesis_df$Over50)$p.value
    outcomeF <- fligner.test(hypothesis_df[[i]],hypothesis_df$Outcome)$p.value
    ## add this to our dataFrame
    flignerTestDf[nrow(flignerTestDf) +1,] <- c(format(round(outcomeF,5), nsmall=5),format(round(over50F,5), nsmall=5))
    ## rename the row
    row.names(flignerTestDf)[nrow(flignerTestDf)] <- i
}

flignerTestDf
library(gridExtra)
png("flignerTest.png", height = 50*nrow(flignerTestDf), width = 200*ncol(flignerTestDf))
grid.table(flignerTestDf)
dev.off()
```

Random Forest Classifier

1. We trained a Random forest classifier, here are the results with no feature engineering, or pre-processing.
2. Accuracy - .7516
   a. 95% CI : (0.6754, 0.8179)
3. Precision - .4906
4. Recall - 0.7027
5. F1 - 0.5778

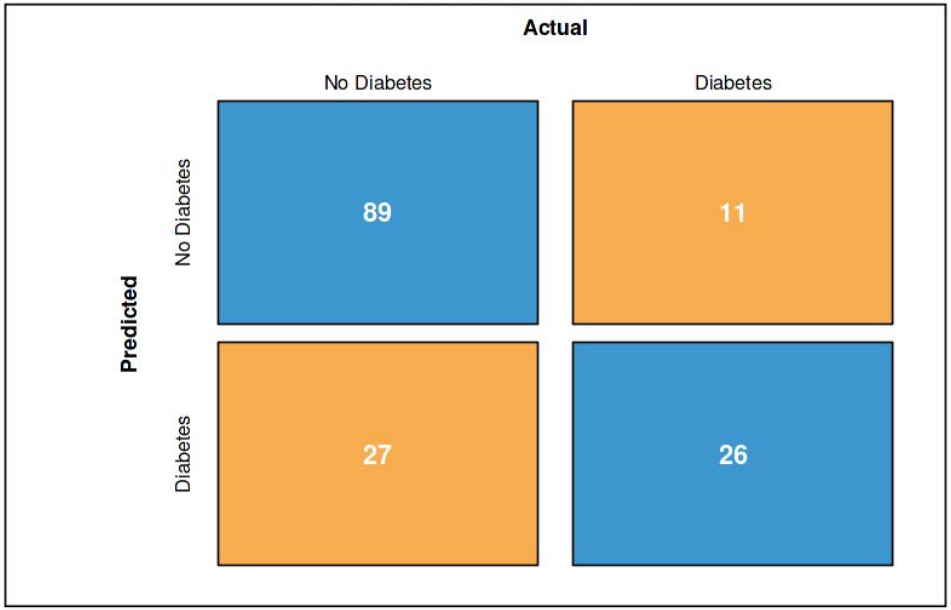
Confusion Matrix - Train

```
Confusion matrix:
    1   2 class.error
1 367  46   0.1113801
2  84 118   0.4158416
```

Confusion Matrix - Test

```
          Reference
Prediction  1  2
         1 74 13
         2 32 34
```

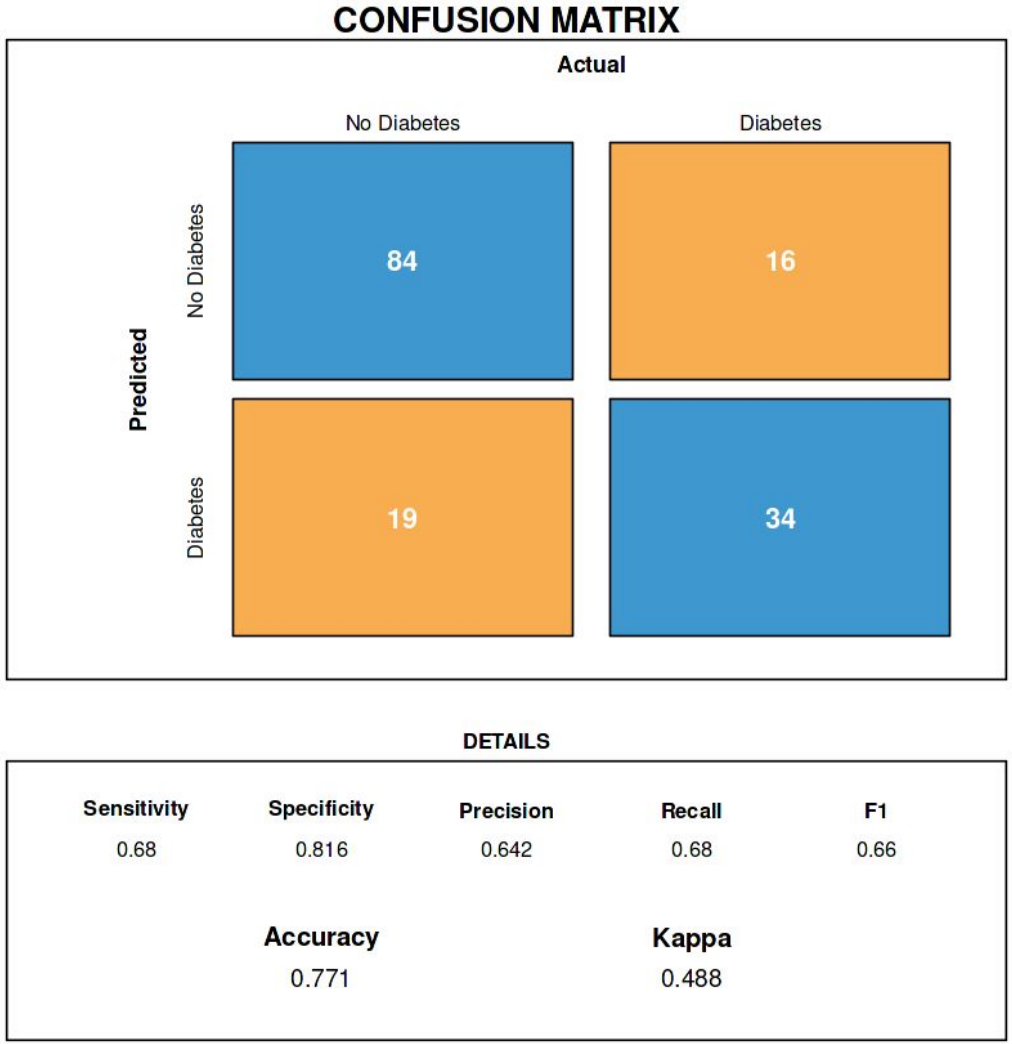
CONFUSION MATRIX



DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.703	0.767	0.491	0.703	0.578

Accuracy	Kappa
0.752	0.41



DETAILS

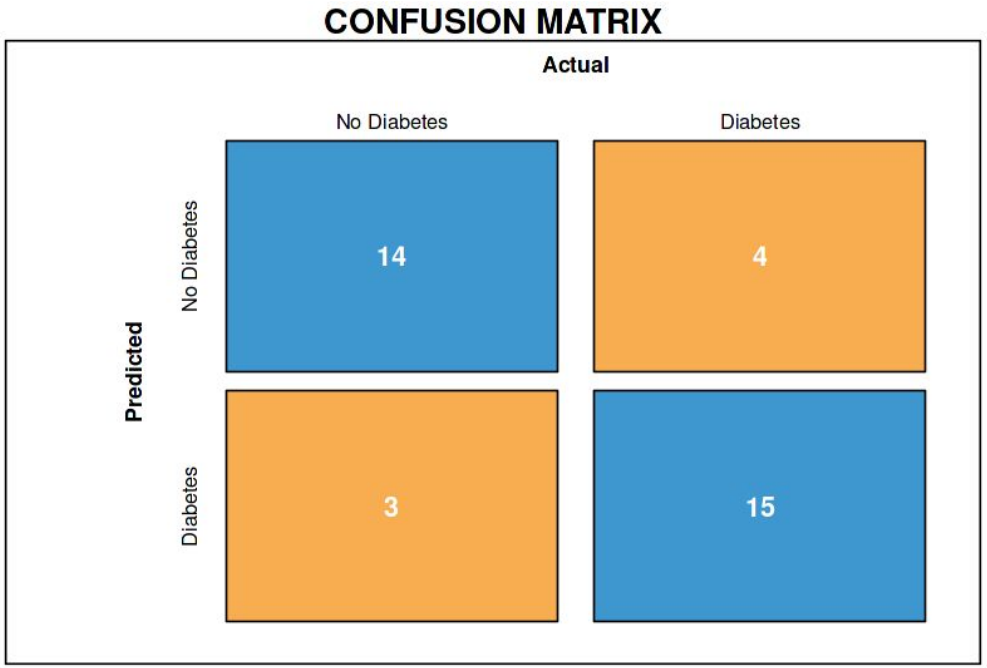
Sensitivity	Specificity	Precision	Recall	F1
0.68	0.816	0.642	0.68	0.66

Accuracy	Kappa
0.771	0.488

Slight improvement when removing Blood Pressure.

Other things tried:

1. Glucose -> X
   a. X^(½)
   b. X^2
   c. X^3
   d. log(X)
2. Glucose -> X, BMI-> Y
   a. X * Y
3. (same with Glucose x BMI x Diabetes Pedigree Function
4. Did not have a large effect



DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.789	0.824	0.833	0.789	0.811

Accuracy	Kappa
0.806	0.611

What did work a bit:

1. Removing outliers
2. Removed rows with 0 for Insulin, BMI, Blood Pressure (assumed to be missing data)
3. Balanced data by undersamplin g

Confusion Matrix Code

```
library('caret')
library('randomForest')
library(creditmodel)
index = createDataPartition(diabetes_df$Outcome,p=.8,list=FALSE)
diabetes_df$Outcome = as.factor(diabetes_df$Outcome)
trainSet <- diabetes_df[index,]
testSet <- diabetes_df[-index,]
fitControl <- trainControl(
    method='cv',
    number = 5,
    savePredictions = 'final',
    classProbs = T
)
predictors<-c("Age","Glucose","Insulin","BMI","Pregnancies","BloodPressure","SkinThickness","DiabetesPedigreeFunction")
outcomeName<-"Outcome"
set.seed(120)
classifier_RF = randomForest(Outcome ~ ., data=trainSet,ntree=500)
classifier_RF

testSet$y_pred<-predict(classifier_RF,newdata=testSet[,predictors])
cm <- confusionMatrix(testSet$Outcome,testSet$y_pred,mode="everything",positive="1")
```

Improvements and Future Questions

1. Future AI approach stuff
2. More testing done on the feature importance, SHAP analysis, deeper models.
3. More statistical features tested (standard deviations for Glucose column)
4. More train/test splits to get more accurate results for model
    a. Develop more higher order statistical features using the features we know are important, e.g. add a feature for standard deviations for important features, and maybe excluded others such as blood pressure from the model.

We wanted to see if we could use machine learning to create an analytical model with better prediction

This is results and analysis dont put in conclusion section of report

1. All three tests agree that blood Pressure is not a good predictor of diabetes outcome but Glucose is.
2. The Fligner-Killeen test suggest neither is BMI (close on this one tho .05022) or Insulin, but the correlation test and F-test (Anova) agree that these two are significant
3. The Fligner-Killeen test suggests skin thickness is a good predictor of diabetes outcome but the F-test suggests this was less significant (but still alpha = .05) and the correlation test says not significantly correlated

References

1. Wild S;Roglic G;Green A;Sicree R;King H; “Global Prevalence of Diabetes: Estimates for the Year 2000 and Projections for 2030.” *Diabetes Care*, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/15111519/.
2. Khare, Akshay Dattatray. “Diabetes Dataset.” *Kaggle*, 6 Oct. 2022, https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset.
3. Volpi E, Nazemi R, Fujita S. Muscle tissue changes with aging. Curr Opin Clin Nutr Metab Care. 2004 Jul;7(4):405-10. doi: 10.1097/01.mco.0000134362.76653.b2. PMID: 15192443; PMCID: PMC2804956.
4. Cybernetic. (2017, March 22). R how to visualize confusion matrix using the caret package. Stack Overflow. Retrieved December 5, 2022, from https://stackoverflow.com/questions/23891140/r-how-to-visualize-confusion-matrix-using-the-caret-package
5. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. mishrapriyank17. “How to Remove Outliers from Multiple Columns in R Dataframe?” GeeksforGeeks, 3 Feb. 2022, https://www.geeksforgeeks.org/how-to-remove-outliers-from-multiple-columns-in-r-dataframe/.
7. Tena, Jezaniah Kira Sarion, "Test Procedures for Equality of Two Variances in Delta Distributions" (2009). Dissertations. 695. https://scholarworks.wmich.edu/dissertations/695
8. “Normality Test in R.” STHDA, http://www.sthda.com/english/wiki/normality-test-in-r.
9. Soetewey, Antoine. “Outliers Detection in R.” Stats and R, 8 Nov. 2020, https://statsandr.com/blog/outliers-detection-in-r/. Accessed 8 Dec. 2022.
10. Stolle, Salome. “Quartiles and Quantiles ~ Differences & Calculation.” BachelorPrint, 11 Nov. 2022, https://www.bachelorprint.com/statistics/quartiles-and-quantiles/#:~:text=Quartiles%20are%20a%20set%20of,in%20research%20as%20statistical%20quantities.

Fligner-Killeen Method

Explanation: This method of hypothesis testing compares the variances of the different groups.

3.3.4 Modified Fligner-Killeen Test (F-K:med χ^2)

This modification of the Fligner-Killeen test (Fligner and Killeen, 1976), suggested by Conover et al. (1981), involves using the ranks of $|X_{ij} - \tilde{X}_i|$, R_{ij}, where $\tilde{X}_i$ is the median of the ith group, and assigning increasing scores of

$$a_{N,R_{ij}} = a_{N,l} = \Phi^{-1}\left(\frac{1}{2} + \frac{l}{2(N+1)}\right)$$

based on those ranks, where $\Phi(x)$ is the cdf of a standard normal distribution. And from these scores, the chi-squared test is formulated based on the statistic

$$X^2 = \sum_{i=1}^{k} n_i(\bar{A}_i - \bar{a})^2/V^2, \tag{3.3}$$

where $\bar{A}_i$ is the mean score in the ith sample, $\bar{a}$ is the overall mean score

$$\bar{a} = \left(\sum_{l=1}^{N} a_{N,l}/N\right),$$

and

$$V^2 = \sum_{l=1}^{N}(a_{N,l} - \bar{a})^2/(N-1).$$

Statistic X^2 has asymptotically χ^2_{k-1} distribution where k is the number of variances compared. This test rejects the hypothesis that k variances are equal (at a significance level of α) if $X^2 > \chi^2_{\alpha,k-1}$.

This modified test was shown to be powerful and robust over various symmetric (uniform, normal and double exponential) and asymmetric (square of the random variables with symmetric distributions considered) distributions in an extensive simulation study conducted by Conover et al. (1981).

Pearson Correlation Formula

Pearson correlation formula

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

m_x and m_y are the means of x and y variables.

The p-value (significance level) of the correlation can be determined :

1. by using the correlation coefficient table for the degrees of freedom : $df = n - 2$, where n is the number of observation in x and y variables.