PROBABILITY OF BEING DIABETES POSITIVE USING LOGISTIC REGRESSION MODEL

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves.

Diabetes, thesedays, is a major public health and global problem that is almost approaching epidemic proportions globally. Worldwide, the prevalence of chronic, noncommunicable diseases is increasing at an worrying level. About 20 million people die every year due cardiovascular disease, where diabetes and hypertension are major leading factors. Today, more than 1.7 billion people around world are overweight, and 312 million of them are obese. In addition, at least 155 million children globaly are overweight or obese. Almost 75% of the total adult diabetics are in developing countries. Two major concerns are that much of this increase in diabetes will occur in developing countries and that there is a growing incidence of Diabetes at a younger age including some obese children even before puberty.

Each year 9 million people develop any type of Diabetes and the most dramatic increases in Diabetes have occurred in populations where there have been rapid and major changes in lifestyle, demonstrating the important role played by lifestyle factors and the potential for reversing the global epidemic. A person with diabetes is 2 – 4 times more likely to get cardiovascular disease, and 80% of people with Diabetes will die from it. Premature mortality caused by diabetes results in an estimated 12 to 14 years of life lost. A person with Diabetes incurs medical costs that are two to five times higher than those of a person without diabetes, and the World Health Organization estimates that up to 15% of annual health budgets are spent on diabetes-related illnesses. The annual direct healthcare costs of diabetes worldwide, for people in the 20–79 age groups, are estimated to be as much as 286 billion USD.

Diabetic neuropathy is probably the most common complication. Diabetic retinopathy is a leading cause of blindness and visual disability. Research findings suggest that, after 15 years of diabetes, approximately 2% of people become blind, while about 10% develop severe visual handicap. Diabetes is among the leading causes of kidney failure, but its frequency varies between populations and is also related to the severity and duration of the disease. Diabetic foot disease, due to changes in blood vessels and nerves, often leads to ulceration and subsequent limb amputation. Diabetes is the most common cause of non-traumatic amputation of the lower limb.

The data set is comprised of 768 observations and 9 variables. It is available in the package mlbench. We will be using diabetes as our response/target variable.

Data Description for the 9 variables are as follows.

- pregnant - Number of times pregnant
- glucose - Plasma glucose concentration (glucose tolerance test)
- pressure - Diastolic blood pressure (mm Hg)
- triceps - Triceps skin fold thickness (mm)
- insulin - 2-Hour serum insulin (mu U/ml)
- mass - Body mass index (weight in kg/(height in m)^2)
- pedigree - Diabetes pedigree function
- age - Age (years)
- diabetes - Class variable (test for diabetes)

The following libraries were used to create this project: tidyverse, caret, dplyr, naniar.

CODE: data("PimaIndiansDiabetes2", package = "mlbench")

data <- PimaIndiansDiabetes2

We first loaded our database which is located within the mlbench package. Then we named the database with „data“.

Using the head (data) code, we will see what our database looks like:

pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
1	89	66	23	94	28.1	0.167	21	neg
0	137	40	35	168	43.1	2.288	33	pos
3	78	50	32	88	31.0	0.248	26	pos
2	197	70	45	543	30.5	0.158	53	pos
1	189	60	23	846	30.1	0.398	59	pos
5	166	72	19	175	25.8	0.587	51	pos

A tibble: 9 × 3

	variable	n_miss	pct_miss
	<chr>	<int>	<dbl>
1	insulin	374	48.7
2	triceps	227	29.6
3	pressure	35	4.56
4	mass	11	1.43
5	glucose	5	0.651
6	pregnant	0	0
7	pedigree	0	0
8	age	0	0
9	diabetes	0	0

A tibble: 9 × 3

	variable	n_miss	pct_miss
	<chr>	<int>	<dbl>
1	pregnant	0	0
2	glucose	0	0
3	pressure	0	0
4	triceps	0	0
5	insulin	0	0
6	mass	0	0
7	pedigree	0	0
8	age	0	0
9	diabetes	0	0

Using the code miss_var_summary (data) we check if there are missing values in our database. As we can see in the database we had missing values which we solved with the code data <- na.omit (PimaIndiansDiabetes2) so there are no more missing values in the database. Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.

Non-numerical variable(s) ignored: diabetes
Descriptive Statistics
data
N: 392

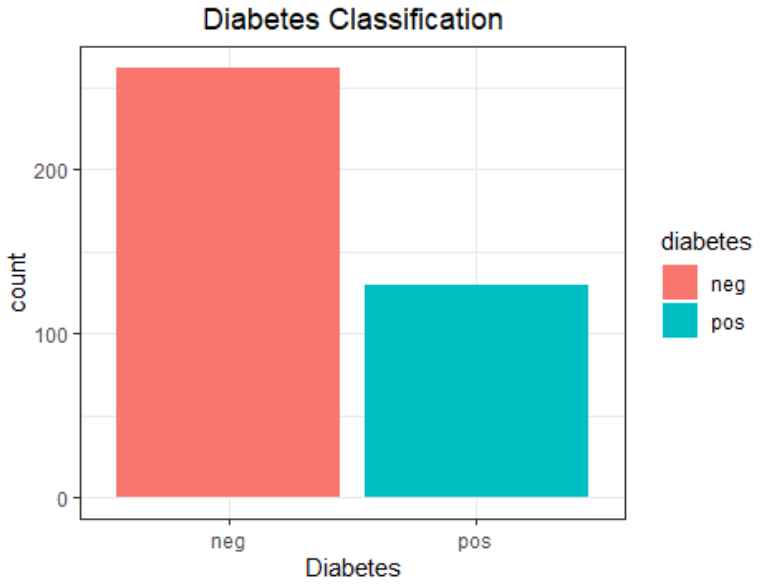
	age	glucose	insulin	mass	pedigree	pregnant	pressure	triceps
Mean	30.86	122.63	156.06	33.09	0.52	3.30	70.66	29.15
Std.Dev	10.20	30.86	118.84	7.03	0.35	3.21	12.50	10.52
Min	21.00	56.00	14.00	18.20	0.09	0.00	24.00	7.00
Q1	23.00	99.00	76.50	28.40	0.27	1.00	62.00	21.00
Median	27.00	119.00	125.50	33.20	0.45	2.00	70.00	29.00
Q3	36.00	143.00	190.00	37.10	0.69	5.00	78.00	37.00
Max	81.00	198.00	846.00	67.10	2.42	17.00	110.00	63.00
MAD	7.41	31.13	80.80	6.67	0.28	1.48	11.86	11.86
IQR	13.00	44.00	113.25	8.70	0.42	4.00	16.00	16.00
CV	0.33	0.25	0.76	0.21	0.66	0.97	0.18	0.36
Skewness	1.39	0.51	2.15	0.66	1.94	1.33	-0.09	0.21
SE.Skewness	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
Kurtosis	1.68	-0.51	6.21	1.50	6.22	1.43	0.75	-0.48
N.Valid	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00
Pct.Valid	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Using the summarytools :: descr (data) code, we got descriptive statistics. Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness. For example, for variable Age we can conclude that the average age of our sample is 30,86 with a standard deviation of 10,20. The youngest person is a 21 year old while the oldest person is a 81 year old.

Using the following code, we obtained the distribution of our variable of interest i.e. the variable diabetes

CODE: ggplot(data, aes(data$diabetes, fill = diabetes)) +

```
geom_bar() +
theme_bw() +
labs(title = "Diabetes Classification", x = "Diabetes") +
theme(plot.title = element_text(hjust = 0.5))
```



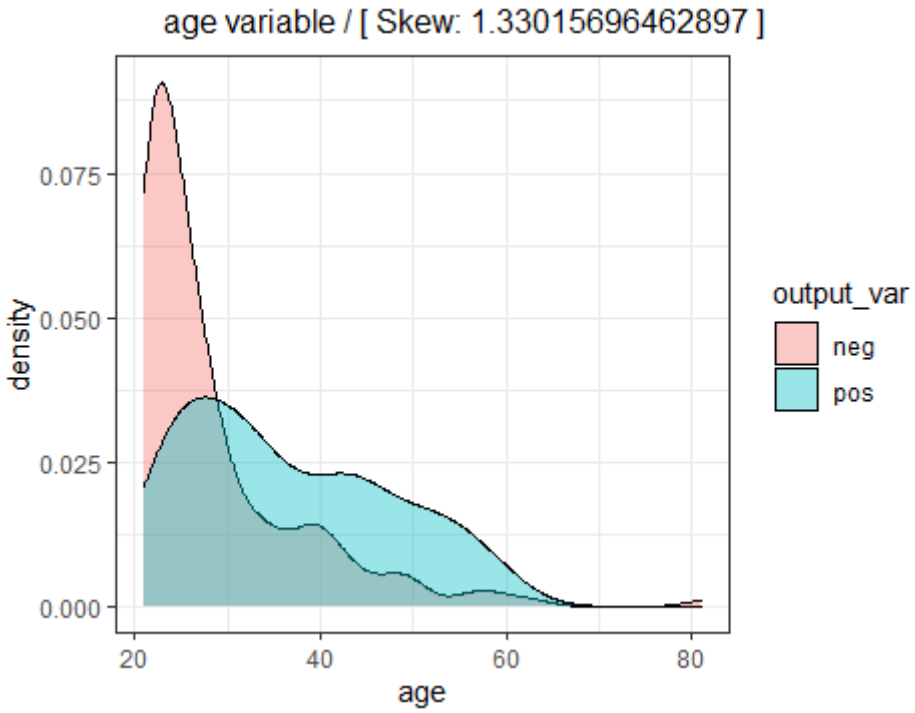
As we can see in the figure above which shows the distribution of the variable diabetes we see that we have over 250 cases that are not recorded as positive for diabetes while this is not the case with the rest, i.e. over 150 people are diagnosed as positive for diabetes.

Using the following code, we obtained distributions of all variables with a degree of skewness.

CODE: univar_graph <- function(univar_name, univar, data, output_var)

```
g_2 <- ggplot(data, aes(x=univar, fill=output_var)) +
    geom_density(alpha=0.4)+
```

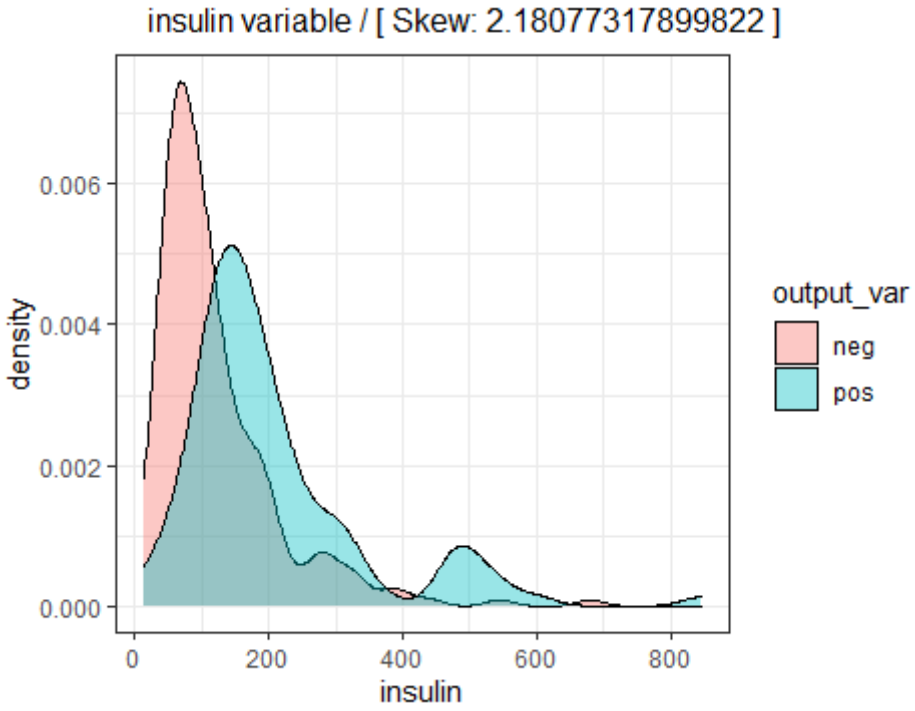
```
            xlab(univar_name) +

            theme_bw()

            gridExtra::grid.arrange( g_2, ncol=1, top = paste(univar_name,"variable", "/ [
Skew:",timeDate::skewness(univar),"]"))

        }

    for (x in 1:(ncol(train.data)-1)) {

        univar_graph(univar_name = names(train.data)[x], univar = train.data[,x], data = train.data,
output_var = train.data[,'diabetes'])

    }
```



People can develop diabetes at any age, from early childhood to adulthood, but the average age at diagnosis is 13 years. An estimated 85% of all type 1 diagnoses take place in people aged under 20 years.[1] As we can see in the picture above as people get older and the chances of getting diabetes are significantly higher. Most people who are negative for diabetes are aged between 20 and 25 years, after which, the older people are and the chances of getting diabetes are increasing. Most people diagnosed with diabetes are in the age group of 25 to 60 years, while most people with diabetes are recorded in the 30s.

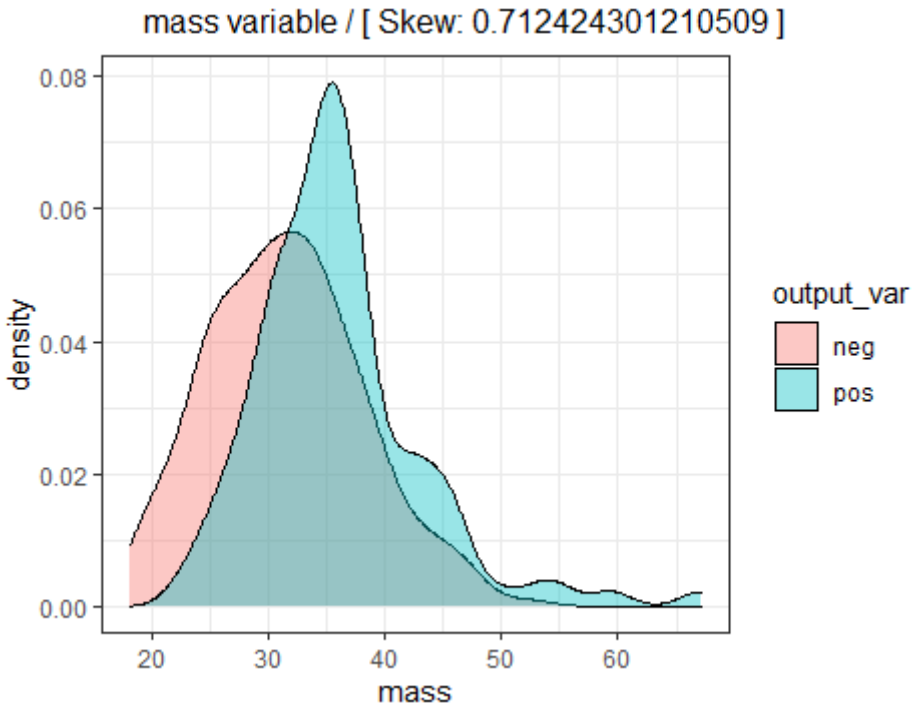
[1] https://www.medicalnewstoday.com/articles/284974

It is important that the concentration of glucose in the blood is maintained at a constant level and controlled carefully. Insulin is a hormone - produced by the pancreas - that regulates glucose concentrations in the blood.[2] As the picture above shows, the higher the glucose concentration, the higher the chances of getting diabetes. The smallest number of positive for diabetes are at glucose levels up to 100 mmol / L, where at higher concentrations than the above, the chances of getting diabetes increase.



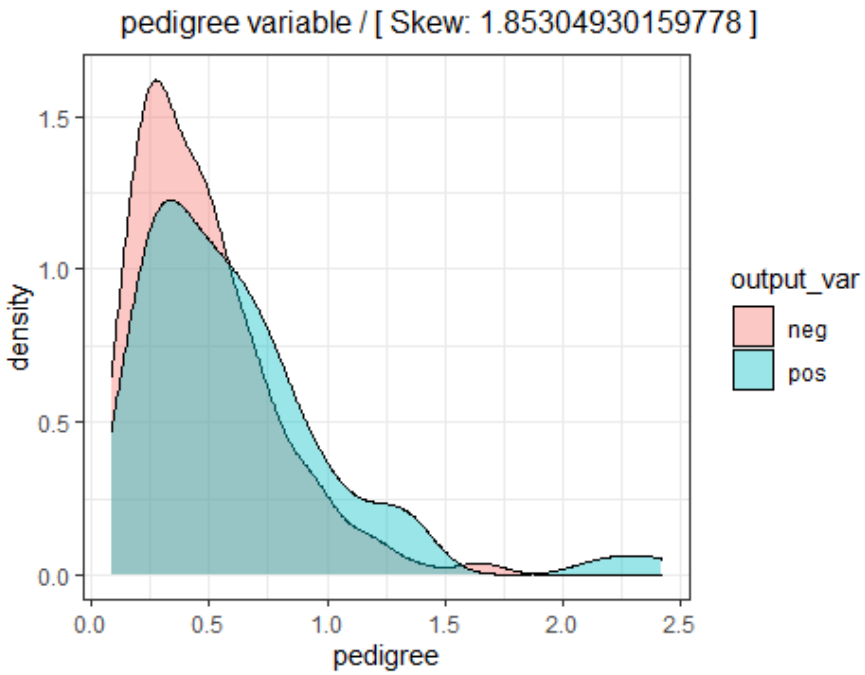
[2] https://www.bbc.co.uk/bitesize/guides/zgqcmsg/revision/3

If you have diabetes: Your glucose levels will continue to rise after you eat because there's not enough insulin to move the glucose into your body's cells. People with type 2 diabetes don't use insulin efficiently (insulin resistance) and don't produce enough insulin (insulin deficiency).[3] Similar to glucose, the higher the level of insulin, the higher the chances of getting diabetes. Most people with diabetes have insulin levels of 200 mg / dl

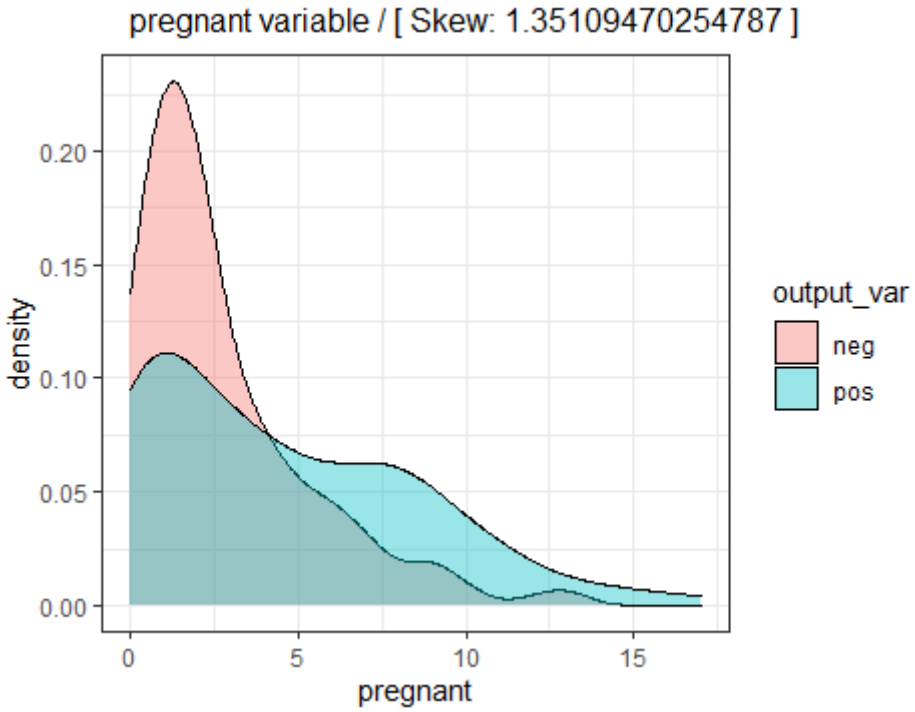


The higher the BMI the higher the chances of getting it. According to the WHO, all people with a BMI over 30 are considered hanged. As we can see in the graph above, most people with diabetes belong to the group of obese people. In general, the higher the BMI, the higher the chances of being diagnosed with diabetes.

[3] https://www.mayoclinic.org/diseases-conditions/diabetes/in-depth/diabetes-treatment/art-20044084#:~:text=If%20you%20have%20diabetes%3A,enough%20insulin%20(insulin%20deficiency).



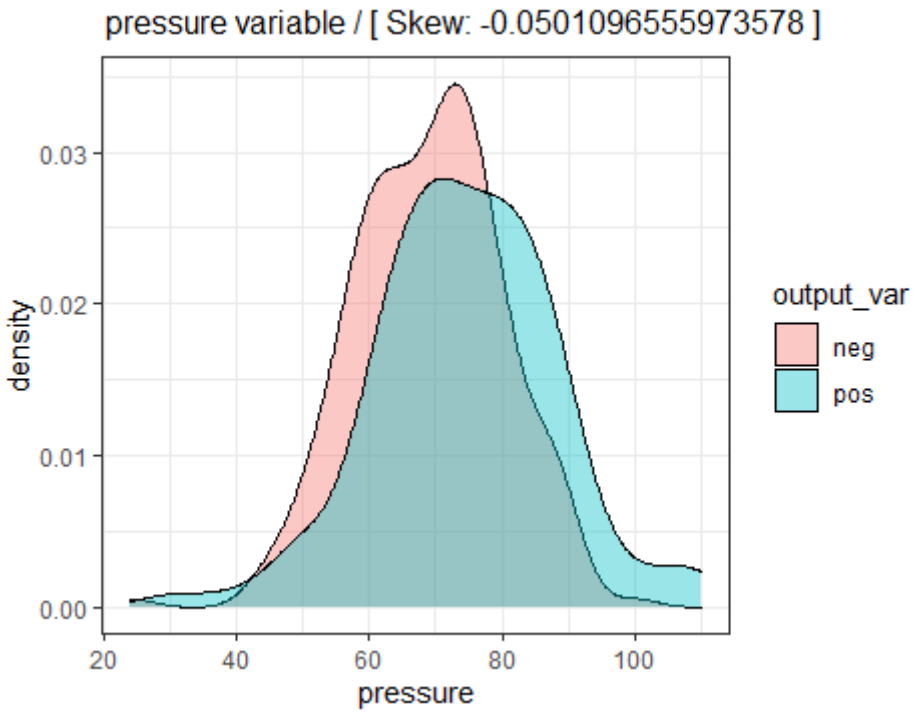
Diabetes pedigree indicates the function which scores likelihood of diabetes based on family history. As we can see in the picture above, the chances of getting diabetes are higher due to the increase in cases of diabetes in the family.



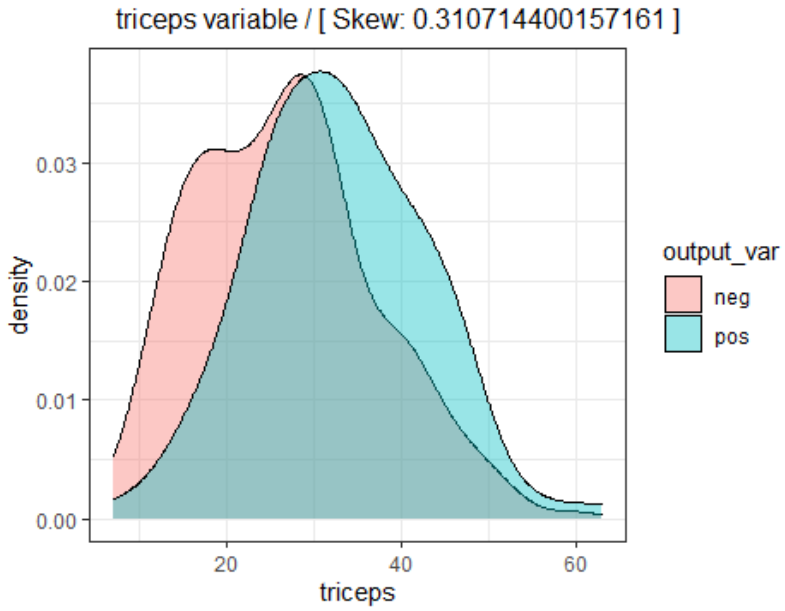
Diabetes during pregnancy—including type 1, type 2, or gestational diabetes—can negatively affect the health of women and their babies. For women with type 1 or type 2 diabetes, high blood sugar around the time of conception increases babies' risk of birth defects, stillbirth, and preterm birth.[4] As

[4] https://www.niddk.nih.gov/health-information/diabetes/diabetes-pregnancy#:~:text=Pregnancy%20can%20worsen%20certain%20long,glucose%20levels%20are%20too%20high.

we can see in the picture above, the higher the number of births, the higher the chances of getting diabetes. Each new birth from the fifth also recorded more people diagnosed with diabetes.



Although within the normal range, early increases of diastolic blood pressure and heart may indicate early cardiovascular changes in response to diabetes and potentially contribute to a greater proclivity for later development of nephropathy. 5 Diabetic pressure levels seems to be normally distributed for negative results and the positive cases range from bp of 60 to 110.



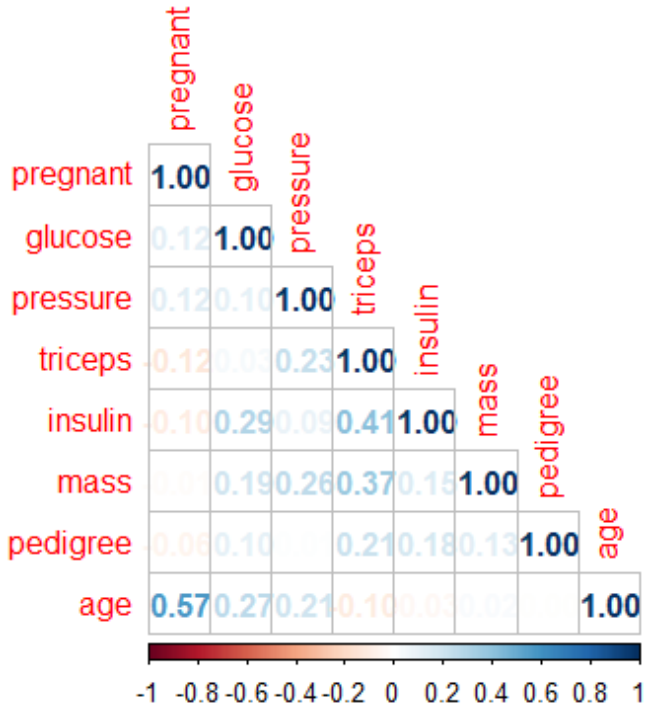
[5] https://pubmed.ncbi.nlm.nih.gov/15207840/

As we can see in the picture above the thickness of the skin has almost no effect on diabetes. We can only notice a small difference that a larger number of patients have thicker skin but the differences are not so significant.

Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y. Correlation measures association, but doesn't show if x causes y or vice versa—or if the association is caused by a third factor. Using the code correlation <- cor (data [, setdiff (names (data (, 'diabetes')]]) we obtained the following correlation values:

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
pregnant	1.00000000	0.2156048	0.21857810	0.0852945	0.08344185	-0.03121094	-0.03476794	0.66703418
glucose	0.21560480	1.0000000	0.22452845	0.1852958	0.58504816	0.21285705	0.08829580	0.37506144
pressure	0.21857810	0.2245285	1.00000000	0.2639493	0.10620819	0.28841518	-0.03058748	0.33609069
triceps	0.08529450	0.1852958	0.26394931	1.0000000	0.11901454	0.67009729	0.15148878	0.18081053
insulin	0.08344185	0.5850482	0.10620819	0.1190145	1.00000000	0.20508466	0.05298820	0.26215913
mass	-0.03121094	0.2128570	0.28841518	0.6700973	0.20508466	1.00000000	0.13905955	0.08593943
pedigree	-0.03476794	0.0882958	-0.03058748	0.1514888	0.05298820	0.13905955	1.00000000	0.04764760
age	0.66703418	0.3750614	0.33609069	0.1808105	0.26215913	0.08593943	0.04764760	1.00000000

The figure above shows the correlation values in a table form, using the code corrplot :: corrplot (correlation, type = "lower", method = "number") the correlation values will also be graphically displayed.



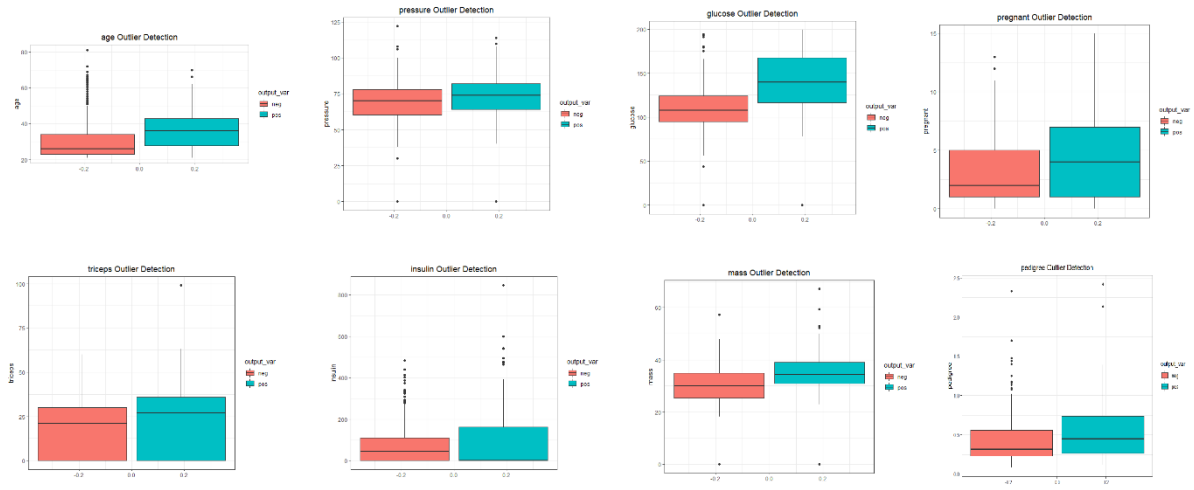
As we note in the figure above, we do not have highly correlated values present in both directions.

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. A

boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

Using the following code using the box plot we will check if we have any outliers present.

```
CODE: box_plot <- function(bivar_name, bivar, data, output_var) {

 g_1 <- ggplot(data = data, aes(y = bivar, fill = output_var)) +

      geom_boxplot() +

      theme_bw() +

      labs( title = paste(bivar_name,"Outlier Detection", sep =" "), y = bivar_name) +

      theme(plot.title = element_text(hjust = 0.5))

 plot(g_1)

}

for (x in 1:(ncol(train_set)-1)) {

 box_plot(bivar_name = names(data)[x], bivar = train_set[,x], data = data, output_var =
data[,'diabetes'])

}
```



We can observe that insulin, pedigree and age have the highest outliers. We will take care of them in the train() function of the CARET package.

As we could notice in the graphical representation of the distribution of variables and their degree of skewness and using the box plot we notice that variables such as Insulin, pedigree and age have high right skewness. Pressure and mass have negative skewness while pregnant, glucose, and triceps have moderate to low right skewness.

In statistics, the (binary) logistic model (or logit model) is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by a indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.

We'll randomly split the data into training set (80% for building a predictive model) and test set (20% for evaluating the model). Also we set seed for reproducibility.

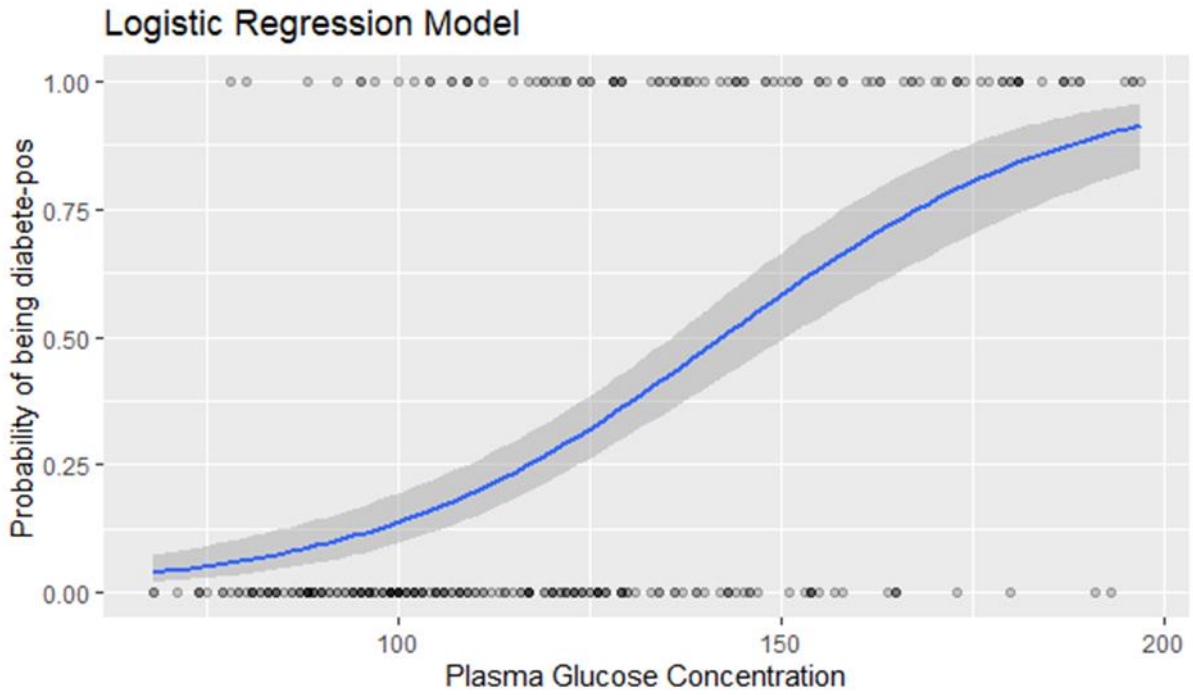
CODE: set.seed(123)

```
training.samples <- data$diabetes %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- data[training.samples, ]
test.data <- data[-training.samples, ]
```

The R function glm(), for generalized linear model, can be used to compute logistic regression, so we specify the option family = binomial, which tells to R that we want to fit logistic regression.

CODE

```
train.data %>%
    mutate(prob = ifelse(diabetes == "pos", 1, 0)) %>%
    ggplot(aes(glucose, prob)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "glm", method.args = list(family = "binomial")) +
 labs(
        title = "Logistic Regression Model",
        x = "Plasma Glucose Concentration",
        y = "Probability of being diabete-pos"
      )
exp(coef(model))
```



In Logistic Regression, we don't directly fit a straight line to our data like in linear regression. Instead, we fit a S shaped curve, called Sigmoid, to our observations. Using the code above, we made an S shaped curve distribution of our simple logistic regression model in which the variable glucose concentration was included. As we can see from the picture that the blood glucose level is higher than 150 mmol/L 75% are more likely to be diabetes positive. If the glucose level is higher than 200 mmol L, the chances of having diabetes are approximately 100%.

CODE: model <- glm(diabetes ~., data = train.data, family = binomial

summary(model)

```
Call:
glm(formula = diabetes ~ ., family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5832  -0.6544  -0.3292   0.6248   2.5968

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.053e+01  1.440e+00  -7.317 2.54e-13 ***
pregnant     1.005e-01  6.127e-02   1.640  0.10092
glucose      3.710e-02  6.486e-03   5.719 1.07e-08 ***
pressure    -3.876e-04  1.383e-02  -0.028  0.97764
triceps      1.418e-02  1.998e-02   0.710  0.47800
insulin      5.940e-04  1.508e-03   0.394  0.69371
mass         7.997e-02  3.180e-02   2.515  0.01190 *
pedigree     1.329e+00  4.823e-01   2.756  0.00585 **
age          2.718e-02  2.020e-02   1.346  0.17840
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 398.80  on 313  degrees of freedom
Residual deviance: 267.18  on 305  degrees of freedom
AIC: 285.18

Number of Fisher Scoring iterations: 5
```

summary(model)$coef

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.50372    1.31719  -7.215 5.39e-13
## pregnant     0.04571    0.06218   0.735 4.62e-01
## glucose      0.04230    0.00657   6.439 1.20e-10
## pressure    -0.00700    0.01291  -0.542 5.87e-01
## triceps      0.01858    0.01861   0.998 3.18e-01
## insulin     -0.00159    0.00139  -1.144 2.52e-01
## mass         0.04502    0.02887   1.559 1.19e-01
## pedigree     0.96845    0.46020   2.104 3.53e-02
## age          0.04256    0.02158   1.972 4.86e-02
```

It can be seen, using asterisks indicating the level of significance, that only 4 out of the 8 predictors are significantly associated to the outcome. These include: pregnant, glucose, mass and pedigree.

The obtaines coefficient estimate of the variable glucose is b = 0.04230, which has positive sign. This means that an increase in glucose level is also associated with increase in the probability of being diabetes-positive. However, the coefficient for the variable insulin is b = -0.00159, which has negative sign, means that an increase in insuline will be associated with a decreased probability of being diabetes-positive.

An important concept to understand, for interpreting the logistic beta coefficients, is the odds ratio. An odds ratio measures the association between a predictor variable (x) and the outcome variable (y). It represents the ratio of the odds that an event will occur (event = 1) given the presence of the predictor x (x = 1), compared to the odds of the event occurring in the absence of that predictor (x = 0).

For example, the regression coefficient for glucose is 0.04257. This indicate that one unit increase in the glucose concentration will increase the odds of being diabetes-positive by exp(0.04257) or 1.04 times.

CODE:
```
model_glm <- caret::train(diabetes ~., data = train.data,
                          method = "glm",
                          metric = "ROC",
                          tuneLength = 10,
                          trControl = trainControl(method = "cv", number = 10,
                                                   classProbs = T, summaryFunction = twoClassSummary),
                          preProcess = c("center","scale","pca"))
```

```
Generalized Linear Model

314 samples
  8 predictor
  2 classes: 'neg', 'pos'

Pre-processing: centered (8), scaled (8), principal component signal extraction (8)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 283, 283, 283, 283, 282, 283, ...
Resampling results:

  ROC        Sens       Spec
  0.8526407  0.8666667  0.5790909
```

ROC score in logistic regression are used for determining the best cutoff value for predicting whether a new observation is a "negative" (0) or a "positive" (1). Sensitivity and specificity are two metrics for evaluating the proportion of true positives and true negatives, respectively. In other words, sensitivity is the proportion of 1s that you correctly identified as 1s using that particular cutoff value, or the true positive rate. Conversely, specificity is the proportion of 0s that you correctly identified as 0s, or the true negative rate. A best-case ROC would look like a 90 degree angle (0,9). For our model ROC score is 0,8526407, which is not a bad but we will try to improve them a little bit.

From the model logistic regression results table, it can be noticed that some variables like triceps, insulin, pressure and age are not statistically significant. Keeping them in the model may contribute to overfitting. Therefore, they should be eliminated. So we will exclude this variables from our initial model and create best model with significant variables: pregnant, glucose, mass and pedigree.

CODE: bestmodel <- glm(diabetes ~ pregnant + glucose + mass + pedigree,

data = train.data, family = binomial)

summary(bestmodel)

```
Call:
glm(formula = diabetes ~ pregnant + glucose + mass + pedigree,
    family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6452  -0.6510  -0.3476   0.6296   2.6213

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.357445   1.258048  -8.233  < 2e-16 ***
pregnant      0.160972   0.046852   3.436 0.000591 ***
glucose       0.040099   0.005581   7.184 6.75e-13 ***
mass          0.096389   0.024094   4.001 6.32e-05 ***
pedigree      1.434442   0.477930   3.001 0.002688 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 398.80  on 313  degrees of freedom
Residual deviance: 270.21  on 309  degrees of freedom
AIC: 280.21

Number of Fisher Scoring iterations: 5
```

As we can see all the variables are significant. All variables have a positive sign which means they have a positive impact on increasing the diagnosis of diabetes. What we can conclude from the obtained model is that pregnant will increase the odds of being diabetes-positive by 1.17 times, glucose by 1.04 times, mass by 1.09 times and pedigree by 4.17 times. The greatest influence has the variable pedigree, which increases the probability of getting diabetes, if we have a higher family history of diabetes, by as much as 4 times.

CODE: bestmodel_glm <- caret::train(diabetes ~ pregnant + glucose + mass + pedigree, data = train.data,

method = "glm",

metric = "ROC",

tuneLength = 10,

trControl = trainControl(method = "cv", number = 10,

classProbs = T, summaryFunction = twoClassSummary),

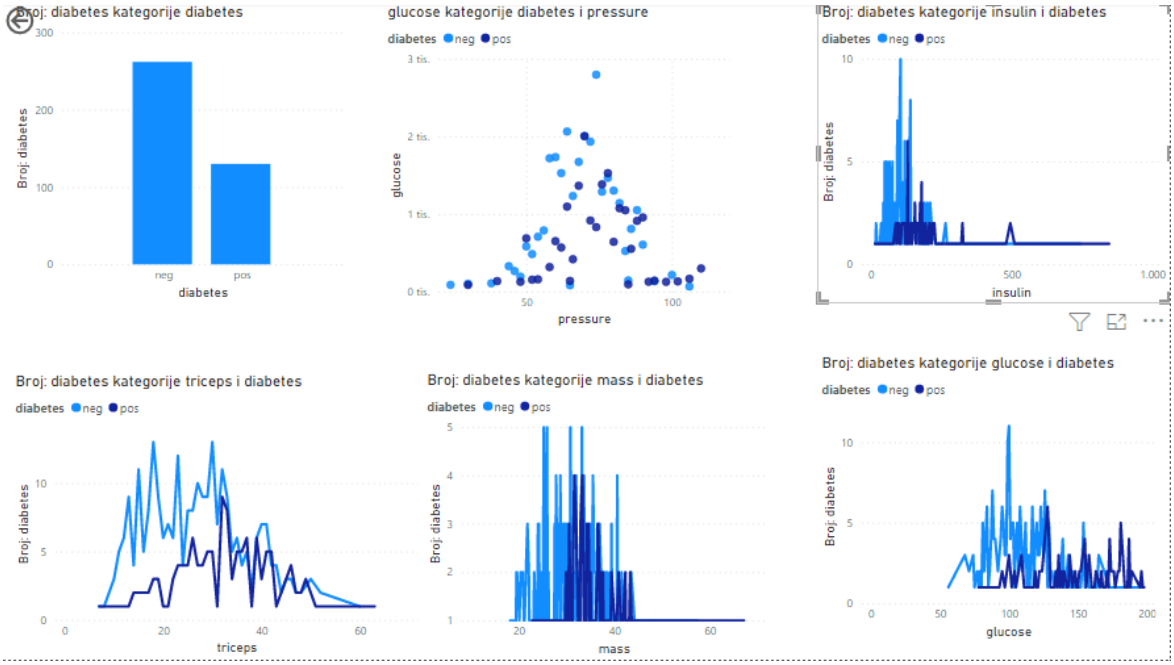
preProcess = c("center","scale","pca"))

```
Generalized Linear Model

314 samples
  4 predictor
  2 classes: 'neg', 'pos'

Pre-processing: centered (4), scaled (4), principal component signal extraction (4)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 282, 283, 282, 283, 282, 283, ...
Resampling results:

  ROC        Sens       Spec
  0.8522078  0.8952381  0.5763636
```

As we can see, there are significant shifts in the performance of the model. The ROC score is not significantly increased, which is reasonable because the previous model also has high predictive capabilities. There are significant improvements in the model in terms of sensitivity and specificity of the model.

Power BI Dashboard

After loading the R script with the base in Power BI we created a few charts that are displayed on the dashboard below:



The dashboard features visuals that have already been shown and explained in this project.