Escola Estadual Gregoriano Canedo

APLICANDO O MODELO PREDITIVO DE APRENDIZADO DE MÁQUINA PARA O DIAGNÓSTICO DO DIABETES TIPO 2 COM LINGUAGEM R

Monte Carmelo, MG

2023

Rafael Moreira de Melo
Samuel Machado

Orientador: Jair Rodrigues de Andrade

APLICANDO O MODELO PREDITIVO DE APRENDIZADO DE MÁQUINA PARA O DIAGNÓSTICO DO DIABETES TIPO 2 COM LINGUAGEM R

Relatório apresentado à 7ª FEMIC - Feira Mineira de Iniciação Científica.
Orientação do Prof. Jair Rodrigues de Andrade.

Monte Carmelo, MG
2023

RESUMO

O diabetes permanece uma doença incurável, porém o pré-diabetes não. O pré-diabetes é uma condição em que os níveis de glicose no sangue estão elevados, mas ainda não atingiram um limite para serem considerados diabetes tipo 2. Com esse intuito estamos desenvolvendo e aprimorando um modelo de previsão com precisão considerada relevante, visando identificar possíveis futuros casos de diabetes, antecipando sua prevenção e normalizando seus níveis de glicose no sangue, com o objetivo de melhorar o atendimento aos pacientes. O conjunto de dados utilizado para a análise e construção do modelo foi originalmente publicado pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais dos Estados Unidos. A Ciência de Dados tem como objetivo analisar e explorar conjunto de informações de banco de dados, selecionando melhores algoritmos para a construção de modelos de Aprendizado de Máquina ou Machine Learning para prever se um paciente tem ou não diabetes usando alguns parâmetros do conjunto de dados, como por exemplo Insulina e Glicose. A ferramenta utilizada para a preparação, análise, treinamento e teste dos dados foi a linguagem R, desenvolvida originalmente com fins estatísticos e análise de dados. O projeto trata-se de um estudo observacional em que foram gerados dois modelos a partir de um conjunto de dados de mulheres norte-americanas com idade entre 21 e 81 anos utilizando a linguagem R com técnicas de tratamento de dados, o melhor resultado foi observado com o algoritmo *Random Forest* ou Floresta Aleatória que atingiu uma acurácia de 88%, o segundo modelo preditivo utilizou o algoritmo *Logistic Regression* ou Regressão Logística com 77% de assertividade. Sendo assim a maior precisão foi do modelo utilizando o algoritmo *Random Forest*. Lembrando, este campo não tem o objetivo de substituir os profissionais da saúde, e sim apenas melhorar o atendimento dos pacientes.

Palavras-chave: Ciências de Dados, Machine Learning, Algoritmo Random Forest, Algoritmo Regressão Logística, Diabetes Mellitus.

SUMÁRIO

1 INTRODUÇÃO

Segundo a Sociedade Brasileira de Diabetes (SBD), o diabetes é uma doença crônica, que afeta homens, mulheres e crianças, na qual o corpo não produz insulina ou não consegue empregar adequadamente a insulina que produz

A insulina é um hormônio secretado pelo pâncreas para processar a glicose e, assim, transformá-la em energia no corpo, sendo um recurso do corpo para controlar a quantidade de glicose presente no corpo. A incapacidade de produzir insulina e a impossibilidade de incorporá-la de forma eficaz leva a altos níveis de glicose no sangue: a hiperglicemia, o que caracteriza o diabetes.

Em geral, o diabetes pode ser dividido em diabetes tipo 1, diabetes tipo 2, diabetes gestacional e pré-diabetes.

O diabetes tipo 1 surge quando o sistema imunológico ataca inadequadamente as células beta do corpo, que são responsáveis por sintetizar e secretar a insulina, resultando em baixa ou nenhuma liberação de insulina, fazendo com que os níveis de açúcar no sangue se elevem. Esta patologia corresponde a 5 a 10% do total de casos de diabetes.

O diabetes tipo 2 acontece quando o organismo não consegue usar adequadamente a insulina produzida, ou que não tenha uma produção de insulina suficiente para que a taxa de glicemia seja controlada. Esta patologia corresponde, aproximadamente, a 90% dos casos.

> O Tipo 2 aparece quando o organismo não consegue usar adequadamente a insulina que produz; ou não produz insulina suficiente para controlar a taxa de glicemia. Cerca de 90% das pessoas com diabetes têm o Tipo 2. Ele se manifesta mais frequentemente em adultos, mas crianças também podem apresentar. Dependendo da gravidade, ele pode ser controlado com atividade física e planejamento alimentar. Em outros casos, exige o uso de insulina e/ou outros medicamentos para controlar a glicose.(SOCIEDADE BRASILEIRA DE DIABETES, s.d., online).

O conjunto de dados utilizado no projeto pode ser encontrado na plataforma Kaggle, uma comunidade on-line de cientistas de dados e profissionais do aprendizado de máquina subsidiária da Google, este conjunto de dados foi originalmente publicado pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais dos Estados Unidos. A Ciência de Dados tem como objetivo analisar e explorar informações de banco de dados, selecionando melhores algoritmos para a construção de modelos de Aprendizado

de Máquina ou Machine Learning, tendo com objetivo prever o acontecimento de algo com base nas informações contidas no banco de dados, como neste projeto, utilizando o modelo para prever se um paciente tem ou não diabetes usando alguns parâmetros do conjunto de dados, como por exemplo Insulina e Glicose. Outras variáveis preditoras, ou dados de entrada, incluem o número de gestações que a paciente teve, pressão diastólica, espessura da pele, seu IMC e idade.

O estudo incluiu a busca da melhor precisão (pelo menos 78%) de dois modelos de aprendizado treinados, utilizando os algoritmos Regressão Logística e Floresta Aleatória ou *Random Forest*, e também o uso da matriz de confusão e a curva ROC que permite visualizar o desempenho de cada modelo preditivo.

A ferramenta usada para preparação da base, treinamento e teste dos modelos foi a linguagem de programação R, utilizando o ambiente de desenvolvimento integrado (IDE) na plataforma Visual Studio Code ou VS Code.

De acordo com as Diretrizes da Sociedade Brasileira de Diabetes, publicado em 2021 (https://diretriz.diabetes.org.br/diagnostico-e-rastreamento-do-diabetes-tipo-2/) sobre o diagnóstico e rastreamento da população com diabetes tipo 2 ou DM2. O rastreamento é recomendado para todos os indivíduos com 45 anos ou mais, mesmo sem fatores de risco, e para indivíduos com sobrepeso/obesidade que tenham pelo menos um fator de risco adicional para DM2, sendo alguns destes, histórico familiar em parentes de primeiro grau, hipertensão arterial, sedentarismo e até algumas etnias oferecem um risco maior como afro descendentes, hispânicos ou indígenas.

Um modelo preditivo é uma representação matemática de um objeto ou processo projetado para simular fenômenos do mundo real como um passo adiante, na esperança de entender mais claramente o que realmente está acontecendo (BARI et al., 2017, online). O modelo preditivo é obtido através de um processo de desenvolvimento de uma ferramenta ou modelo matemático em combinação com dados para resolver um problema, gerando uma previsão assertiva (KUHN e JOHNSON, 2013, online; BARI et al., 2017, online), onde estes são gerados pela análise preditiva.

O projeto de pesquisa consiste em explorar a inteligência artificial (IA), em particular os algoritmos Regressão Logística e Floresta Aleatória para criar dois modelos de previsão para o diagnóstico do diabetes tipo 2. A Regressão Logística é um dos algoritmos usados para modelagem preditiva, cuja principal característica é o fato

de que sua variável dependente, isto é, a variável que representa o positivo ou negativo do teste, ser categórica e geralmente binária, representando, por exemplo, 0 ou 1, sendo representados negativo e positivo, respectivamente.

O *Random Forest* ou Floresta Aleatória é um dos algoritmos de classificação muito robustos, pois em suas classificações é levado em conta a correlação entre as variáveis, não observado em algoritmos como Naive Bayes, mas que o mesmo ainda é um dos mais usados.

A ferramenta digital utilizada, o Visual Studio Code ou VS Code é um ambiente digital de acesso aberto e disponibilizado pela Microsoft, fornece apoio a diversas linguagens de programação e assim uma oportunidade de aplicação do conhecimento de ciência de dados na área, tem práticas facilitadas por meio de extensões produzidas até mesmo pela comunidade.

2 JUSTIFICATIVA

A pandemia de Covid-19 teve um impacto muito negativo no ganho de peso e no controle glicêmico em pessoas com diabetes tipo 2, segundo dados de uma pesquisa internacional realizada por pesquisadores da Fulda University of Applied Sciences, na Alemanha, e publicada pela revista. assistência médica. Curiosamente, o efeito foi revertido em pacientes com diabetes tipo 1 avaliados na pesquisa. Ainda sobre o problema, ele ressalta o momento pandêmico, mas também vale lembrar que a digitalização do mundo e o posicionamento das pessoas quanto a isso vem transformando a sociedade aos poucos, e cada vez mais temos pessoas mais sedentárias.

Com o problema apresentado, iremos desenvolver uma pesquisa com o grupo de iniciação científica para realizar um estudo com banco de dados, em que os diagnosticados com diabetes tipo 2, tenha algumas características em comum, utilizando da Inteligência Artificial para o estudo e construção de algoritmos de *Machine Learning* ou Aprendizado de Máquina, incentivando a interdisciplinaridade proporcionando a exploração dos processos criativos em diferentes áreas, conhecimentos, espaços, ensino e a organização da instituição.

3 OBJETIVOS

3.1 Objetivo geral

A ideia de desenvolver um projeto de ciência de dados com aprendizado de máquina foi acordada e teve como objetivo entender e resolver situações cotidianas a fim de promover o desenvolvimento local e melhorar a qualidade de vida da comunidade, quanto a conscientização sobre o assunto. Assim, tendo em conta os maus hábitos alimentares e a elevada propensão ao diabetes, consideramos utilizar este tópico como base.

3.2 Objetivos específicos

- O projeto tem como premissa explorar por meio da inteligência artificial o campo do Aprendizado de Máquina ou *Machine Learning*.
- Utilizar dos algoritmos Regressão Logística e Floresta Aleatória, visando construir dois modelos para a previsão do diagnóstico do diabetes tipo 2.
- Mostrar a importância do campo da Ciência de Dados, que inclusive tem crescido muito nos últimos tempos, para desenvolver a multidisciplinaridade nos diversos campos do conhecimento.
- Discutir sobre metodologias de pesquisa na área das Ciências da Natureza, Matemática e suas Tecnologias com banco de dados explorados e publicados em artigos científicos.
- Promover o diálogo sobre as metodologias de análise de dados com linguagem de programação R, Python ou alguma utilizada para o mesmo fim, nas pesquisas em Ciências da Natureza, Matemática e suas Tecnologias.

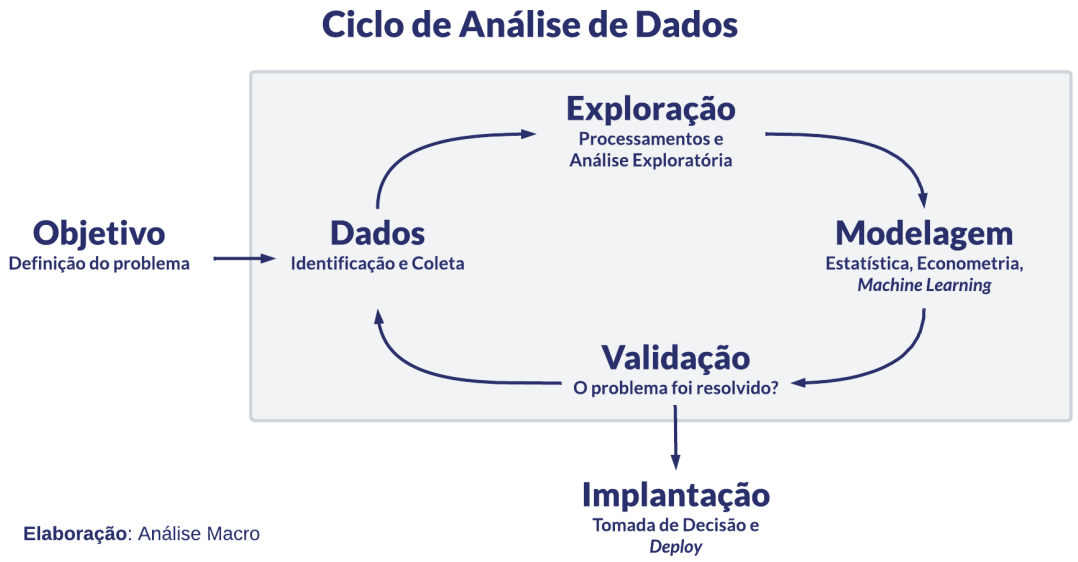
4 METODOLOGIA

Um dos desafios enfrentado por muitos pesquisadores na área de ciência de dados é trabalhar com um banco de dados editados sobre as normas internacionais, este problema pode ser resolvido com o acesso à plataforma Kaggle. Dessa maneira, este

estudo mostra, de maneira introdutória e muito básica, o conceito de mineração de dados. Os conceitos de conjunto de dados, banco de dados, *dataset* ou *data frame* (df) tem basicamente o mesmo significado, para um melhor entendimento, sendo assim, estes são um conjunto de informações, geralmente colocados sobre linhas e colunas, para uma melhor compreensão durante a análise exploratória. Estudos anteriores exploraram as relações entre tratamento metódico dos dados e escolha correta do algoritmo para o aprendizado do modelo em virtude de obter uma previsão correta da variável de saída ou resultado que pretende obter uma melhor acurácia.

Como fazer uma análise exploratória dos dados?

Figura 1 - Ciclo da Análise de Dados



Fonte: https://analisemacro.com.br/

Assim podemos apresentar as principais fases para obter um bom estudo, e uma análise e construção de modelo minuciosa, demonstrando a importância do trabalho do cientista de dados no estudo do problema, utilizando um banco de dados em que conhecemos os dados de entrada (x) e a variável dependente de saída (y).

A base de dados utilizada contém dados de pacientes diagnosticados com diabetes.

Primeiramente utilizamos a função *read.csv*, que é responsável pela leitura de um conjunto de dados, contido em um arquivo com extensão csv, *Comma Separated*

Values (Valores Separados por Vírgula), intitulado diabetes.csv. Nesse processo, o conjunto de dados é lido e armazenado na em formato de linhas e colunas, criando um data frame (Figura 2). O data frame foi criado com o nome de df.

Logo em seguida usamos a função *head* para mostrar de maneira breve, como os dados são representados e suas variáveis.

Figura 2 - Visualização das primeiras 4 observações do data frame

Importando os dados

```
# Carregando o conjunto de dados para o ambiente
df <- read.csv("diabetes.csv")
```

A seguinte tabela mostra os dados que serão analisados, um breve resumo

```
# Vendo as primeiras linhas para visualizar brevemente
head(df, 4)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
```

Fonte: os autores

Com as funções ncol e nrow, ou dim, é possível visualizar as dimensões do conjunto de dados, sendo representados em linhas e colunas, podemos ver quantidade de cada um desses atributos (Figura 3). E também podemos apresentar a função class, onde ela retorna o tipo da variável que está sendo estudada (Figura 3), como por exemplo o número de gestações, onde é representado por um número do tipo inteiro, o que verificamos rapidamente, já que não existe um número quebrado de gestações.

Figura 3 -Visualizando as propriedades do data frame

```
# Verificando o formato dos dados
cat(sprintf("Linhas: %s\nColunas: %s\n\n", nrow(df), ncol(df)))
```

```
## Linhas: 768
## Colunas: 9
```

```
for (i in colnames(df)) {
    cat(sprintf("%s - %s\n",i, class(df[,i])))
}
```

```
## Pregnancies - integer
## Glucose - integer
## BloodPressure - integer
## SkinThickness - integer
## Insulin - integer
## BMI - numeric
## DiabetesPedigreeFunction - numeric
## Age - integer
## Outcome - integer
```

Fonte: os autores

A análise exploratória de dados será apresentada com mais detalhes no capítulo do apêndice em que é descrito todas as fases da preparação dos dados com a análise estatística descritiva, construção da tabela de correlação entre as variáveis de entrada do dataset, constatar a presença de *outliers* ou valores discrepantes com a construção de boxplot, identificando valores ausentes ou ocultos e finalmente substituindo os valores ausentes pela mediana (*median*).

4.1 CONSTRUÇÃO DO MODELO PREDITIVO

Com a função *sample*, é gerado uma amostra aleatória que iremos usar como índices, ou seja, um número que representa o local onde a informação do conjunto de dados é guardada, para a divisão dos dados em treino e teste, tendo aproximadamente, 70% dos dados para treino e 30% para testes (Figura 4).

Figura 4 - Criando os índices dos dados de treino e teste

```
# Usando 70% dos dados para treino
indices <- sample(1:nrow(df), nrow(df)*0.7)

# Verificando o tamanho da amostra
cat(sprintf("O tamanho do conjunto de treino é de %s%s\n", round(length(indices)/nrow(df)*100, 2), "%"))
```

```
## O tamanho do conjunto de treino é de 69.92%
```

Fonte: os autores

Agora iremos dividir o conjunto de dados em 4 partes, treino_x, treino_y, teste_x, teste_y (Figura 5), sendo assim, o modelo treina com os dados x para prever a variável y, representada no data frame como *Outcome.*

Figura 5 - Dividindo em partes o conjunto de dados

```
# Definindo o conjunto de dados para treino
treino_x <- df[indices, -9]
treino_y <- df[indices, 9]
# Definindo o conjunto de dados para teste
teste_x <- df[-indices, -9]
teste_y <- df[-indices, 9]
```

Fonte: os autores

5 RESULTADOS OBTIDOS

Dentre os dois modelos criados, o que apresentou melhor acurácia no teste depois do treinamento foi a Floresta Aleatória com 88% de precisão, e em segundo a Regressão Logística com 77% de precisão. De acordo com nosso objetivo de atingir pelo menos 78%, apenas o algoritmo *Random Forest* ou Floresta Aleatória conseguiu alcançar e superar o objetivo. Conforme o resultado apresentado na matriz de confusão, na linha *Accuracy* (Figura 6), o modelo Regressão Logística apresentou 77% de precisão para os dados de teste.

Figura 6 - Matriz de confusão do modelo Regressão Logística

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 133  30
##          1  23  45
##
##                Accuracy : 0.7706
##                  95% CI : (0.7109, 0.8232)
##     No Information Rate : 0.6753
##     P-Value [Acc > NIR] : 0.000957
##
```

Fonte: os autores

Otimização do modelo preditivo com o algoritmo *Random Forest* (Figura 7).

Figura 7 - Matriz de confusão do modelo *Random Forest*

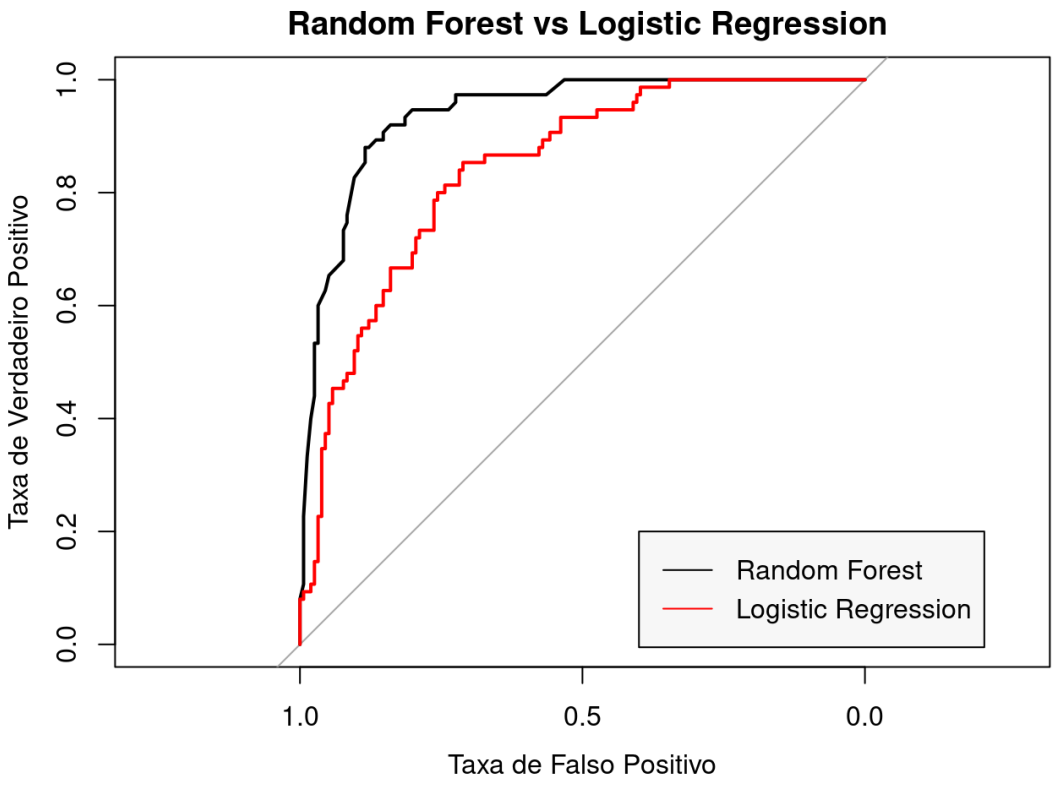
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 138   9
##          1  18  66
##
##                Accuracy : 0.8831
##                  95% CI : (0.8345, 0.9215)
##     No Information Rate : 0.6753
##     P-Value [Acc > NIR] : 1.82e-13
##
```

Fonte: os autores

5.1 COMPARANDO OS MODELOS

Com o objetivo de comparar os dois modelos podemos construir uma curva ROC, que representa o desempenho do modelo, sendo o eixo x, os casos falsos positivos, e o eixo y, os casos verdadeiros positivos (Figura 8). Quanto mais a curva estiver para canto superior esquerdo, melhor o modelo em questão de precisão de suas previsões com base nos dados de teste.

Figura 8 - Curva ROC dos dois modelos



Fonte: os autores

6 CONCLUSÕES OU CONSIDERAÇÕES FINAIS

A identificação do diabetes tipo 2 usando um modelo preditivo com o algoritmo Floresta Aleatória, em dos vários algoritmos do campo *Machine Learning*, que é um subcampo da Inteligência Artificial (IA), permite pré-diagnosticar um paciente com diabetes a partir de uma combinação de dados, informações comuns em pacientes diabéticos. Sendo assim, este projeto não tem o objetivo de substituir o papel de um médico, mas que com essa ideia pode-se melhorar o atendimento da comunidade quanto ao tratamento do diabetes. Com este projeto queremos mostrar a importância dos profissionais e do campo da Ciência de Dados, que atua no estudo dos dados, análise exploratória, escolha do algoritmo preditivo correto, observando correlações e discrepâncias nos dados por meio de estudos estatísticos, em que todas essas decisões afetam na precisão do modelo preditivo.

Os algoritmos Regressão Logística e Floresta Aleatória são aplicações do mundo da inteligência artificial para construir modelos preditivos em diversas áreas do

conhecimento. No entanto, com o conjunto de dados usado nesses modelos, estamos atuando na área da saúde, tornando o rastreamento do diabetes mais eficiente para ajudar médicos, enfermeiros e laboratório de testes clínicos a prestar um melhor atendimento a esses pacientes. É importante lembrar que este campo do aprendizado de máquina não foi desenvolvido para substituir estes profissionais, os algoritmos simplesmente foram desenvolvidos para agilizar tarefas que demoraria muito tempo para que fosse realizada por um ser humano, pois analisar e avaliar 768 observações de um conjunto de dados que tem variáveis que se correlacionam entre si é uma tarefa muito difícil para qualquer profissional

REFERÊNCIAS

Balaraman, S. (2020). Comparison of Classification Models for Breast Cancer Identification using Google Colab. Preprints, pages 1–12.

BARI, Anasse; CHAOUCHI, Mohamed; JUNG, Tommy. Predictive Analytics for Dummies. 2. ed. Hoboken: John Wiley & Sons, Inc, 2017. 9 p.

KUHN, Max; JOHNSON, Kjell. Applied Predictive Modeling. New York: Springer, 2013. 7 p.

MESQUITA, P. S. B. Um Modelo de Regressão Logística para Avaliação dos Programas de Pós–Graduação no Brasil. Dissertação (Mestrado) — Universidade Estadual do Norte Fluminense, Campo dos Goytacazes, 2014. Citado 3 vezes nas páginas 14, 17 e 20.

Sociedade Brasileira de Diabetes.s.d. Disponível em: https://diabetes.org.br/. Acesso em: 08 out. 2023.

APÊNDICE

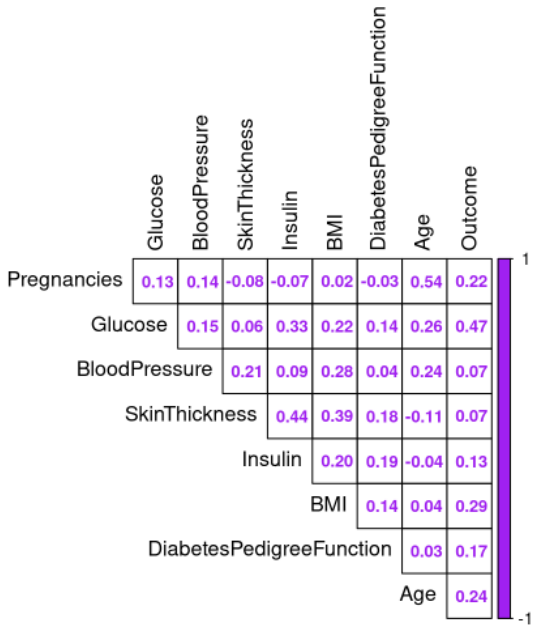
APÊNDICE A - Análise exploratória dos dados com estatística descritiva

```
# Vendo uma breve descrição dos dados
summary(df)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

Fonte: os autores

APÊNDICE B – Tabela de correlação entre as variáveis do conjunto de dados

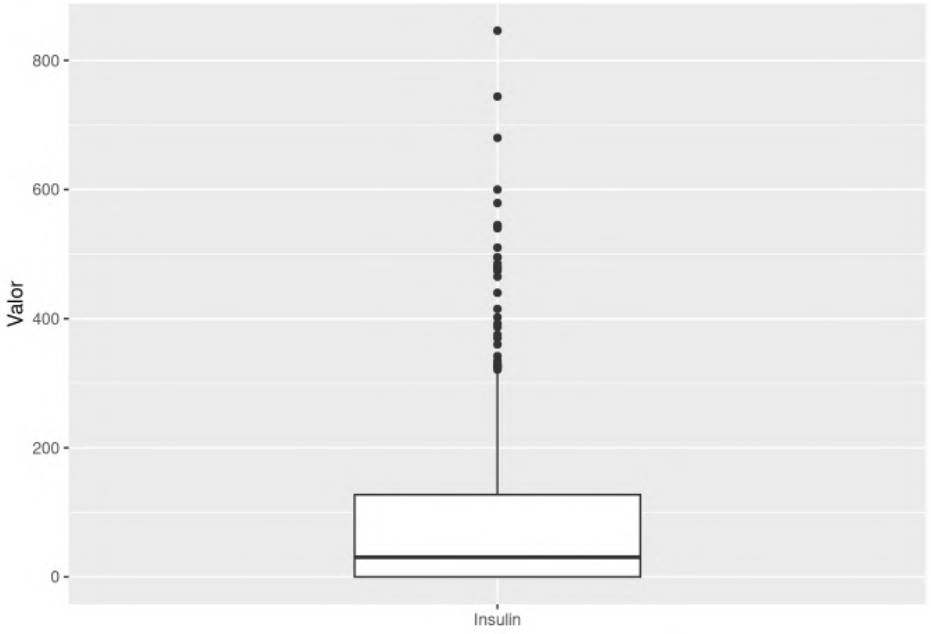


Fonte: os autores

APÊNDICE C – Verificando os *outliers* ou valores discrepantes

Neste gráfico observamos os **Outliers**, um tipo deles, já mecionado antes, é os valores iguais a 0 que são impossíveis.

```
ggplot(df, aes(x="Insulin", y=Insulin)) +
  geom_boxplot(width = 0.4) +
  xlab("") +
  ylab("Valor")
```



Fonte: os autores

APÊNDICE D – Verificando a existência de valores ausentes ou ocultos

```
# Verificando se existem dados ausentes
for (i in colnames(df)) {
    cat(sprintf("%s - %s\n",i, sum(df[,i]==0)))
}
```

```
## Pregnancies - 111
## Glucose - 5
## BloodPressure - 35
## SkinThickness - 227
## Insulin - 374
## BMI - 11
## DiabetesPedigreeFunction - 0
## Age - 0
## Outcome - 500
```

Fonte: os autores

APÊNDICE E – Substituindo os valores ausentes pela mediana

Os dados faltantes que são iguais a 0 atrapalham o modelo em seu aprendizado, então o melhor a se fazer é tratar esses dados, em exemplo de dados que serão tratados são os valores iguais a 0 que são impossíveis como dito anteriormente, a Insulina, por exemplo. Os dados faltantes serão substituídos pela **mediana** dos mesmos dados, menos a variável Gravidez por conta de ela ser um valor possível

```
# Criando uma variável com o nome de todas as colunas que serão tratadas
colunas <- c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI")

# Substituindo os dados da coluna de acordo com o level do Outcome, sendo uma mediana para vada valor do Outcome,
'0' e '1'
for (i in colunas) {
    df[df['Outcome']==0,i] <- replace(df[df['Outcome']==0,i], df[df['Outcome']==0,i]==0, median(df[df[,i]!=0 & df
['Outcome'] ==0,i]))
    df[df['Outcome']==1,i] <- replace(df[df['Outcome']==1,i], df[df['Outcome']==1,i]==0, median(df[df[,i]!=0 & df
['Outcome'] ==1,i]))

}
```

Fonte: os autores