**RELATÓRIO DO MODELO DE PREDIÇÃO**

**Autoria: Luiz H Alves do Nascimento**

* **DESCRIÇÃO DO PROBLEMA**

A partir de uma população indígena, os “índios Prima”, localizada na cidade de Phoenix no Arizona, Estados Unidos, foram coletados dados dos residentes maiores de 5 anos a cada dois anos entre 1965 e 1988. Os dados extraídos são o resultado de vários testes a que foram submetidos os integrantes da população indígena e que estão relacionados ao diagnóstico de diabetes mellitus, que pode ser confirmado se a glicemia no sangue em jejum estiver em 126 mg/dl (7,0 mmol/l) ou mais.

O principal intuito da coleta de vários indicadores e características é de diagnosticar precocemente o desenvolvimento de diabetes mellitus através de um modelo de predição a partir de indicadores que podem ou não estar relacionados à doença. Assim, para atingir esse objetivo, a ideia é obter os dados, construir o modelo, analisar sua eficiência, efetuar os ajustes necessários e tentar replicá-lo para situações gerais.

* **SELEÇÃO DAS VARIÁVEIS E DO MÉTODO**

As variáveis a serem utilizadas na construção do modelo serão:

- Número de vezes em que a pessoa ficou grávida (Pregnancies);

- Glicose no plasma sanguíneo em jejum (Glucose);

- Pressão do sangue em mmHg (BloodPressure);

- Espessura da dobra cutânea do triceps em mm (Triceps skin fold thickness);

- Insulina sérica em µU/ml (Insulin);

- IMC em kg/m2 (BMI);

- Função de propensão a diabetes (DiabetesPedigreee);

- Idade em anos (Age).

as quais estão relacionadas a diabetes, tendo em vista que essa doença tende a afetar ou estar relacionada principalmente a esses indicadores, denominados de fatores de risco. Essas variáveis serão as independentes, sendo o resultado da predição (Outcome) a dependente, sendo ela 0 para a ausência de diabetes e 1 para a presença de diabetes.

O modelo será treinado para receber as variáveis independentes e retornar se a pessoa tem uma grande disposição para diabetes ou não. Esse modelo será construído a partir tanto do método de análise discriminante e de uma regressão logística, os quais podem ser utilizados para construir um modelo de predição para variáveis dependentes discretas, os dois modelos gerados serão avaliados e comparados para verificar o com melhor desempenho.

* **COLETA DOS DADOS**

Os dados a serem utilizados podem ser obtidos a partir do link

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>. A planilha contém

tanto as variáveis independentes quanto o resultado esperado a partir dos dados para

treinamento e teste do modelo. Esses dados representam a coleta feita da populção indígena,

os “Índios Prima”, que são o objeto de estudo.

* **APLICAÇÃO DO MÉTODO**

Os métodos que serão aplicados possuem os seguintes pressupostos que devem ser atendidos para que se tenha um modelo adequado para os dados apresentados:

- Ausência de outliers;

- Independência das observações;

- Normalidade multivariada (os dados apresentam distribuição normal);

- Linearidade (variação linear dos dados);

- Ausência de multicolinearidade (correlação entre as variáveis independentes);

- Homogeneidade das matrizes de variância e covariância.

Para construir os modelos, utiliza-se as seguintes bibliotecas e linhas de código no R:

if (!require("readxl")) install.packages("readxl")

library(readxl)

if (!require("dplyr")) install.packages("dplyr")

library(dplyr)

if (!require("MASS")) install.packages("MASS")

library(MASS)

if (!require("e1071")) install.packages("e1071")

library(e1071)

if (!require("pROC")) install.packages("pROC")

library(pROC)

if (!require("caret")) install.packages("caret")

library(caret)

if (!require("DescTools")) install.packages("DescTools")

library(DescTools)

# Importando dados do Excel

Link da base de dados: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

diabetes <- read.csv("diabetes.csv")

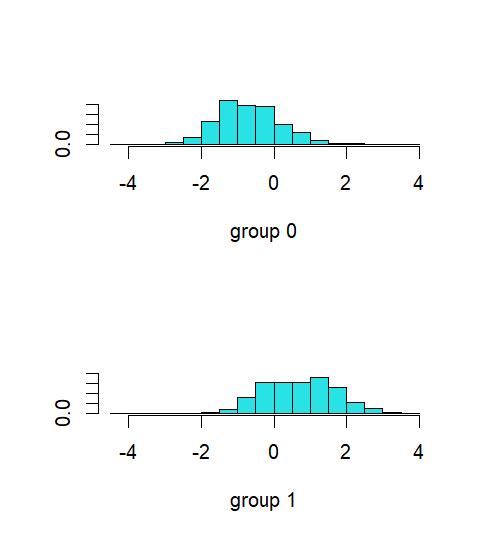
**- ANÁLISE DISCRIMINANTE**

# Modelo de Análise Discriminante

modelo\_lda <- lda(Outcome~., data = diabetes)

# Gráfico dos Grupos

plot(modelo\_lda)



# Previsões

predictions\_lda <- modelo\_lda %>% predict(diabetes)

# Atribuindo os Valores de Cada Classificação

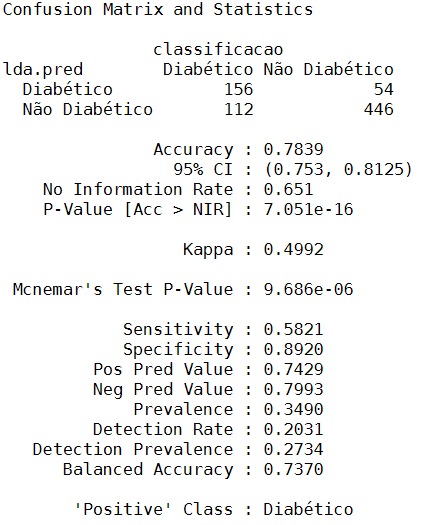
lda\_pred <- ifelse(predictions\_lda$class == "1", "Diabético", "Não Diabético")

classificacao\_lda <- ifelse(diabetes$Outcome == "1", "Diabético", "Não Diabético")

confusion\_matrix\_lda <- table(lda\_pred, classificacao\_lda)

# Matriz de Confusão

confusionMatrix(confusion\_matrix\_lda)



**- REGRESSÃO LOGÍSTICA**

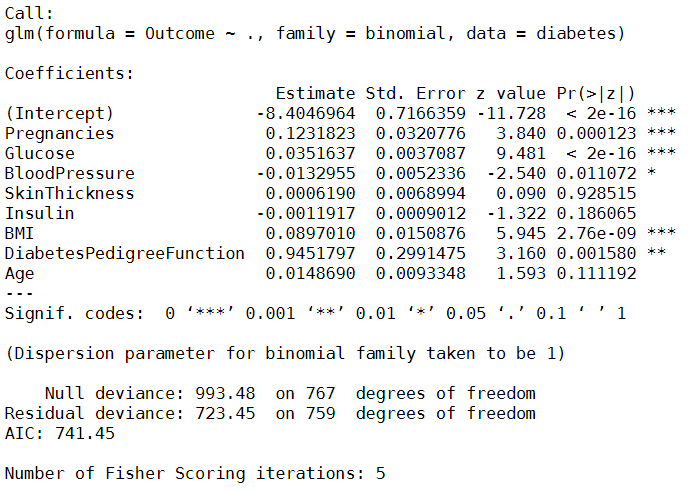
# Modelo de Regressão Logística

modelo\_regLog <- glm(Outcome~., data = diabetes, family = binomial)

# Visualização do Modelo

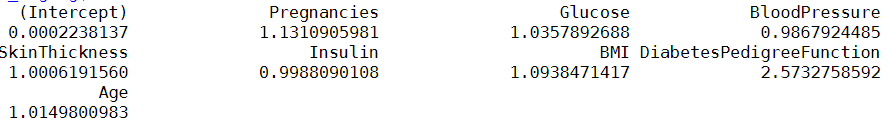
summary(modelo\_regLog)

,



# Convertendo os coeficientes do modelo de log odds para razão de chance

exp(modelo\_regLog$coefficients)



# Obetendo o Pseudo R2

PseudoR2(modelo\_regLog)



# Previsão das Classes

glm\_probs <- predict(modelo\_regLog, type = "response")

# Classificação dos Indivíduos

glm\_pred <- ifelse(glm\_probs > 0.5, "Diabético", "Não Diabético")

# Matriz de Predição

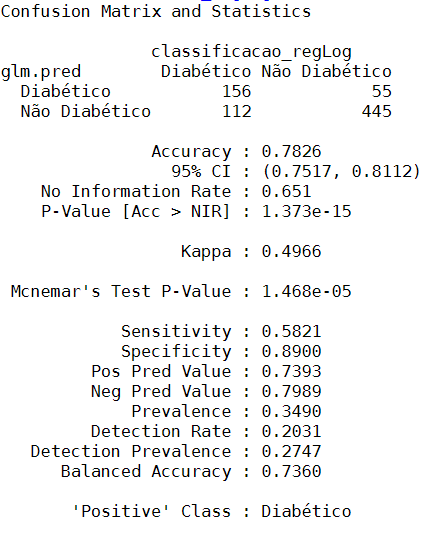
table(glm\_pred, diabetes$Outcome)

# Matriz de Confusão

classificacao\_regLog <- ifelse(diabetes$Outcome > 0.5, "Diabético", "Não Diabético")

confusion\_matrix\_regLog <- table(glm\_pred, classificacao\_regLog)

confusionMatrix(confusion\_matrix\_regLog)



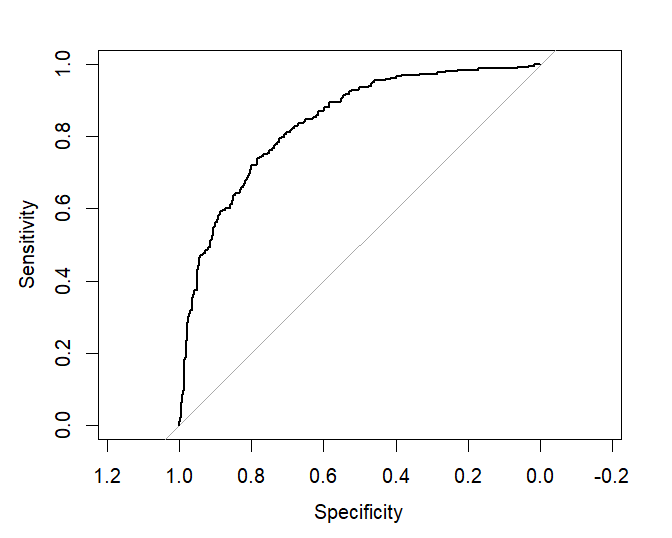
# Curva ROC

roc <- roc(diabetes$Outcome, glm\_probs)

# Área sob a curva ROC

roc

plot.roc(roc)



* **ANÁLISE DO DESEMPENHO**

**- ANÁLISE DISCRIMINANTE**

Analisando o modelo de predição criado através da análise discriminante, observa-se que ele acertou a classificação em 78 dos 100 casos, ou seja, obteve uma acurácia de 78,3% (acurracy = 0.7839). O p-valor foi menor que 0.05 para a acurácia, rejeitando a hipótese nula e indicando uma boa precisão do modelo, e também para o teste de Mcnemar, indicando que há diferenças significativas entre as proporções de discordância entre as variáveis.

A sensitividade (verdadeiros positivos) foi baixa (sensitividade = 0.5821) e a especificidade (verdadeiros negativos) foi alta (especificidade = 0.8920), tendo a classe “Diabético” como positiva, assim, o modelo consegue identificar moderadamente os casos em que a pessoa tem verdadeiramente a doença e muito os casos em que ela realmente não possui.

Desse modo, o modelo atende aos pressupostos estabelecidos para uma análise discriminante, podendo ser considerado um bom modelo.

**- REGRESSÃO LOGÍSTICA**

Analisando o modelo de regressão, verificamos que as variáveis “Pregnancies”, “Glucose”, “BloodPressure”, “IMC” e “DiabetesPedigreeFunction” tiveram um p-valor < 0.05, ou seja, a hipótese nula (insignificância estatística) nesses casos foi rejeitada, sendo as variáveis em questão estatisticamente significativas para o modelo. O contrário ocorreu com as outras variáveis independentes, “SkinThickness”, “Insulin” e “Age”, sendo estatisticamente insignificantes para o modelo. O valor do pseudo R2 ajustado foi de 0.2718, distante do ideal que seria 1, assim, o modelo não explica muito bem a variação dos resultados, ou seja, o modelo de forma linear é pouco preciso no geral.

A partir da matriz de confusão, podemos verificar que a acurácia de acerto do modelo preditivo foi de 0.7826 (78 acertos em 100), um valor bom que indica uma boa capacidade de predição. A sensitividade foi moderada (sensitividade = 0.5821), e a especificidade foi alta (especificidade = 0.8900) para o modelo, um resultado praticamente igual ao do modelo de análise discriminante.

Na curva roc, temos uma área abaixo do gráfico de 0.8394, como é possível verificar também na figura, indicando que o modelo possui uma alta capacidade de discriminação entre os eventos de interesse e não interesse.

Assim, o modelo atende aos pressupostos estabelecidos para uma regressão logística, podendo ser considerado um bom modelo.

* **COMPARAÇÃO DOS MÉTODOS**

Na comparação entre os métodos aplicados, tem-se que acurácia foi praticamente igual e alta nos dois casos, aproximadamente 0.78. A sensitividade nos dois métodos foi a mesma e a especificidade foi ligeiramente maior na análise discriminante, mas uma diferença pouco significativa, sendo baixa a sensitividade e alta a especificidade, um resultado bom, mas que poderia ter sido melhor para a primeira característica.

Ademais, o modelo da regressão logística mostrou que nem todas as variáveis contribuem para explicar o resultado da predição e que esse método tem uma boa capacidade de discriminação entre os eventos de interesse e não interesse.

* **CONCLUSÃO**

Portanto, em razão das poucas diferenças práticas entre o método de análise discriminante e da regressão logística, considera-se aquele que mais se encaixa teoricamente nos dados e variáveis utilizados, no caso a regressão logística, em razão da variável dependente ser binária, possui apenas dois resultados.

No modelo de regressão logística, verifica-se que as variáveis “SkinThickness”, “Insulin” e “Age” poderiam ser cortadas de coletas futuras devido ao seu pouco impacto no resultado da variável final, ou seja, na discriminação do diagnóstico de diabético ou não.

Ademais, a partir dos resultados obtidos percebe-se um modelo com resultados bons para o diagnóstico precoce de diabetes, e que pode seguir adiante para a aplicação em mais amostras populacionais para se avaliar a qualidade do modelo em outras pessoas de diferentes regiões, o que contribuiria também para aumentar a amostra de dados necessária para atender ao pressuposto da grande amostra na regressão linear, onde há grandes chances do modelo apresentar resultados bem diferentes. Assim, o modelo é útil em um pré-diagnóstico, mas está longe de substituir os diagnósticos oficiais e garantir a não ocorrência do diabetes mellitus, sendo os cuidados básicos de saúde e o consumo de alimentos saudáveis ainda indispensável.

* **REFERÊNCIAS BILIOGRÁFICAS**

SMITH, J.W. et al. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. **Proc Annu Symp Comput Appl Med Care**. Hopkins, p. 5, nov. 1988.