Universidade Federal da Bahia
Curso: Tópicos B - Introdução a Machine Learning
Docente: Ricardo Rocha
Discentes: Lucas Rabelo, Kim Leone, Mariana Almeida e Sandro Junior

Introdução:

Os dados foram coletados e disponibilizados pelo "Instituto Nacional de Diabetes e Doenças Digestivas e Renais" como parte do banco de dados Pima Indians Diabetes. Em particular, todos os pacientes são mulheres da herança indígena Pima.
O banco de dados é composto por 2000 observações e 9 variáveis sendo as mesmas:

Gravidez: Número de gestações da paciente;
Glicose: concentração plasmática de glicose por mais de 2 horas em um teste oral de tolerância à glicose (mg/dl);
Pressão Arterial: Pressão Arterial Diastólica (mm/Hg);
Espessura da pele : espessura da dobra da pele do tríceps (mm);
Insulina: insulina sérica de 2 horas (mu U / ml);
IMC: Índice de massa corporal (peso em kg / (altura em m²));
DiabetesPedigreeFunction: função pedigree Diabetes (uma função que pontua probabilidade de diabetes com base no histórico familiar);
Idade: Idade (anos);
Resultado ou Diagnóstico: variável de classe (0 se não diabético, 1 se diabético);

Objetivo:

Usar técnicas de Machine Learning para prever se um paciente tem ou não diabetes, com base em certas medidas de diagnóstico incluídas no conjunto de dados.

Técnicas:

Usaremos o R e alguns de seus populares pacotes relacionados à ciência de dados, bem como o python para a criação do gráfico de correlação. Primeiro, foi importado o *CARET* para ler os dados do referido dataset. Foram usados os pacotes *TIDYVERSE* e *GGTHEMES* para visualizações e manipulação dos dados. Em seguida, aplicaremos os modelos de machine learning propostos em sala de aula (Knn, Regressão Logística, Árvore de Decisão, Análise Discriminante Linear e Quadrática e a Floresta Aleatória) e construiremos o melhor modelo de classificação.

Breve análise descritiva dos dados:

Para a realização da análise, o banco foi separado entre os indivíduos que apresentaram diabetes e os saudáveis. No grupo diabético, a média de insulina produzida pelo corpo foi de 100.3 muU/ml, já no grupo sem diabetes a mesma foi de 68.79 muU/ml. Além disso, vale

ressaltar que a média de glicose foi maior para o grupo dos diabéticos sendo este valor de 141.3 mg/dl, o qual entra em contraste com o valor médio de 110 mg/dl presente no grupo de mulheres não diabéticas. Sobre o IMC, o maior foi o dos diabéticos com valor de 35.14 (indicando uma obesidade média no grupo) e o valor do grupo saudável foi de 30.30 (sobrepeso).

Abaixo segue o gráfico da proporção de mulheres diabéticas e não diabéticas do conjunto de dados:

FIGURA 1

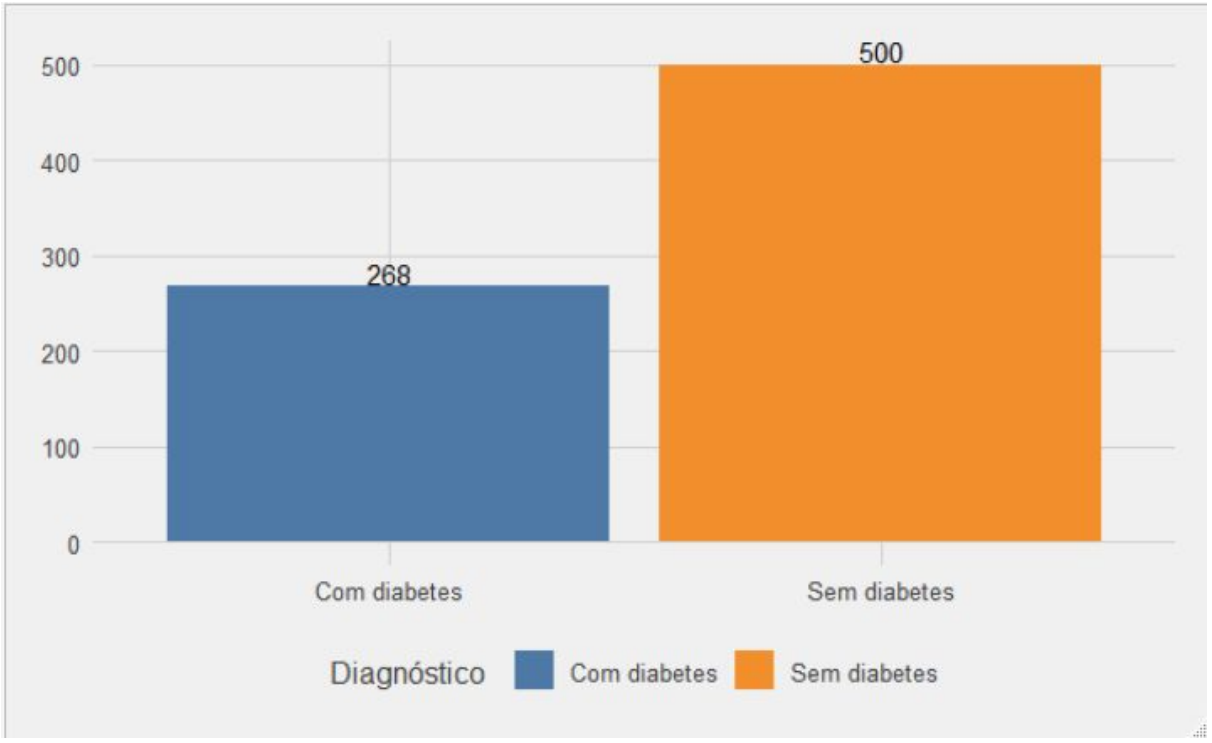
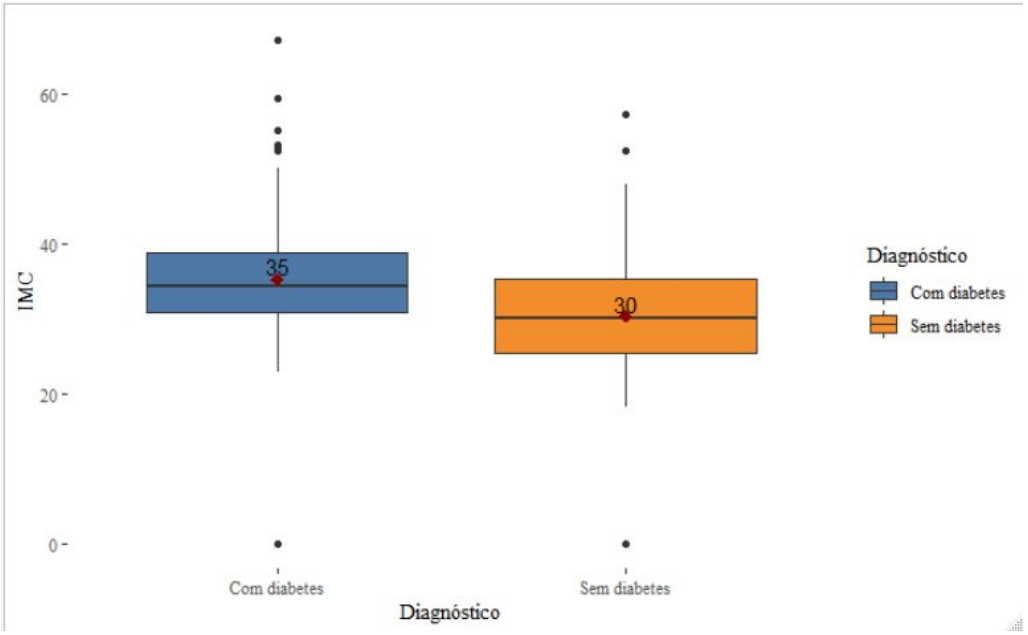


Gráfico referente a proporção de mulheres diabéticas e não diabéticas do banco.

É possível perceber que a proporção de não diabéticos é maior, porém, o valor da presença de diabéticos é alto para mulheres referente a herança dos índios Pima.

FIGURA 2

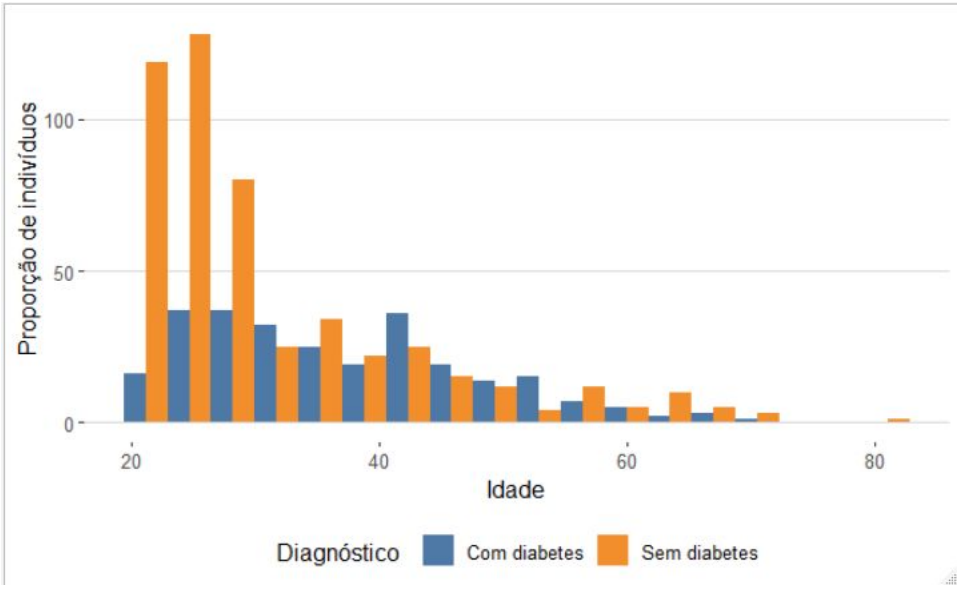


Boxplot referente ao diagnóstico e IMC, o ponto vermelho se refere a média.

Ao fazer uma análise do boxplot, podemos notar que a média do índice de massa corporal é maior em mulheres diabéticas e a presença de valores discrepantes do IMC também.

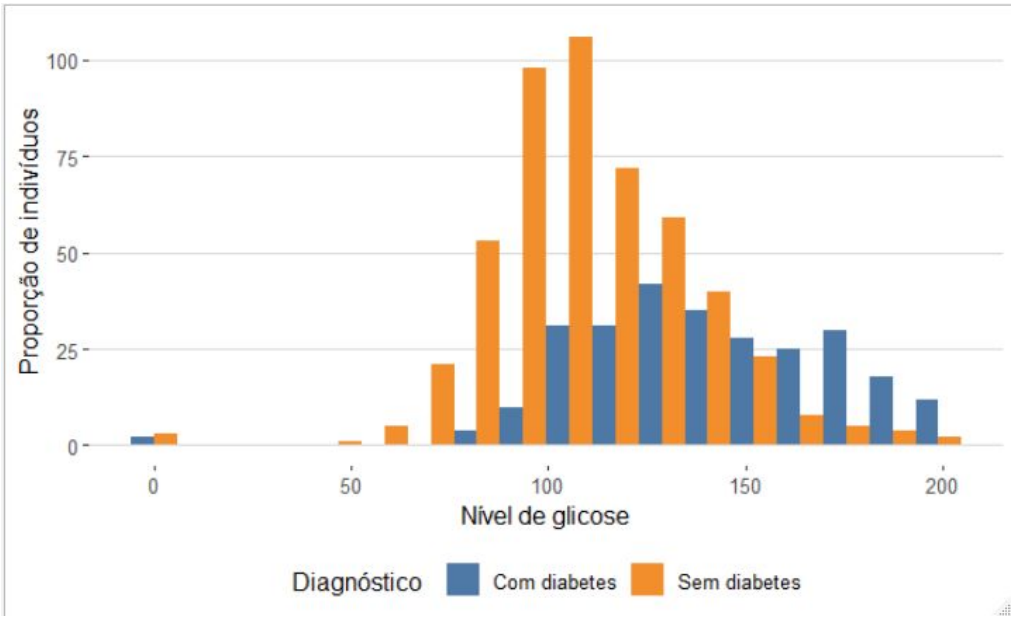
Além disso, é importante explorar a idade dos dois grupos. Para tal, foi construído o histograma:

FIGURA 3



É perceptível que a maioria das mulheres não diabéticas possuem menos de 40 anos de idade, porém, a mais velha também não possui a doença. Já o grupo diabético teve uma variação maior nas idades.

FIGURA 4

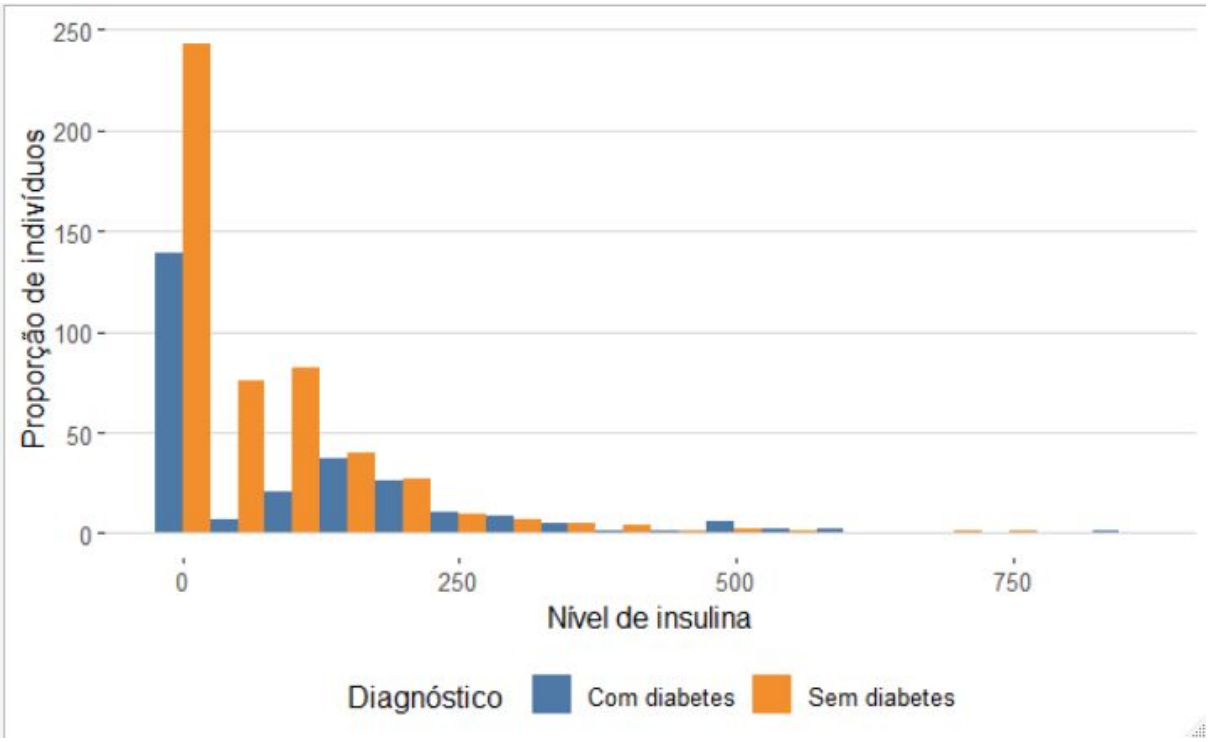


Histograma do nível de glicose das mulheres do banco

No gráfico, é possível perceber que as mulheres diabéticas possuem um nível consideravelmente maior de glicose do que aquelas sem a doença (apenas uma foge dessa observação). Tanto no grupo diabético quanto no não diabético o nível de glicose está acima de 100 mg/dl.

Referente ao nível de insulina foi feito um histograma para analisar a diferença do mesmo dentre os dois grupos:

FIGURA 5



Histograma referente ao nível de insulina nas mulheres diabéticas e não diabéticas.

O gráfico revela que o maior nível de insulina está presente no conjunto de mulheres que apresentam diabetes, porém a maior parte das mulheres diabéticas tem o nível de insulina similar ao das que não conviviam com a doença.

Abaixo o gráfico de correlação entre as variáveis, utilizado para nortear as variáveis consideradas para achar a melhor acurácia.

FIGURA 6

	N_Gravidez	Glicose	Pres_Sang	Esp_pele	Insulina	Imc	DPF	Idade	Diag
N_Gravidez	1.0	0.129	0.141	-0.0817	-0.0735	0.0177	-0.0335	0.544	0.222
Glicose	0.129	1.0	0.153	0.0573	0.331	0.221	0.137	0.264	0.467
Pres_Sang	0.141	0.153	1.0	0.207	0.0889	0.282	0.0413	0.24	0.0651
Esp_pele	-0.0817	0.0573	0.207	1.0	0.437	0.393	0.184	-0.114	0.0748
Insulina	-0.0735	0.331	0.0889	0.437	1.0	0.198	0.185	-0.0422	0.131
Imc	0.0177	0.221	0.282	0.393	0.198	1.0	0.141	0.0362	0.293
DPF	-0.0335	0.137	0.0413	0.184	0.185	0.141	1.0	0.0336	0.174
Idade	0.544	0.264	0.24	-0.114	-0.0422	0.0362	0.0336	1.0	0.238
Diag	0.222	0.467	0.0651	0.0748	0.131	0.293	0.174	0.238	1.0

As cores mais claras indicam mais correlação. Logo, podemos afirmar que os níveis de glicose, idade, IMC e número de gestações apresentaram uma correlação considerável com a variável de resultado (Diag- referente ao diagnóstico).

Separando os dados em treinamento e teste:

Para treinar o modelo de classificação. Foram utilizados os modelos de aprendizado de máquina citados na aula. Primeiro, foi utilizada a função *CreatedataPartition* para separar os dados em teste e treinamento, considerando 30% e 70% respectivamente, e, em seguida, a função train para treinar o modelo. Depois, foi utilizada a função *predict*, a qual permite analisar os resultados do modelo, além da realização da matriz de confusão para assim acessar a acurácia do modelo.

Modelos:

Considerando quatro combinações de variáveis preditoras (Age + BMI, Glucose + BMI, BMI+Age+Pregnancies+Glucose e todas as variáveis),

º KNN

o melhor ajuste foi o que considerou Glicose e IMC, com K = 23 e acurácia de 80,87%;

º Regressão Logística

o melhor ajuste foi o que considerou todas as variáveis, com acurácia 78,70%;.

º Análise discriminante linear

o melhor ajuste foi o que considerou todas as variáveis, com acurácia de 79,57%;

º Análise discriminante quadrática

o melhor ajuste foi o que considerou IMC+Idade+Gravidez+Glicose, com acurácia de 77,83%;

º Árvore de decisão

o melhor ajuste foi o que considerou IMC+Idade+Gravidez+Glicose, com acurácia de 75,65%;

º Floresta Aleatória

o melhor ajuste foi o que considerou todas as variáveis, com acurácia de 76,96%;

Portanto, podemos afirmar que o melhor modelo foi o KNN.

Referências Bibliográficas

https://www.kaggle.com/uciml/pima-indians-diabetes-database

Códigos

https://raw.githubusercontent.com/marreapato/IML-trab/master/Diabetes_analise_exp
https://raw.githubusercontent.com/marreapato/IML-trab/master/Diabetes