Aluno : Lucas Araujo Azevedo
Matrícula: 2017104188

Temas: Diabetes

Motivação:

Verificar a possibilidade de descobrir diabetes a partir de alguns dados sobre a saúde do indivíduo. Possibilitando uma precaução antes mesmo de tê-la, previnindo a doença.

Perguntas:

Trabalho irá utilizar a base de dados do Kagle sobre diabetes (https://www.kaggle.com/uciml/pima-indians-diabetes-database) e focará em tentar resolver as seguintes perguntas:

- A grossura da pele do triceps tem alguma relação com a chance de ter diabetes?
- A gravidez gera uma tendência a ter diabetes?
- Existe relação entre glicose, insulina, pressão sanguinea?

Out[1]:

Show Cell

Out[2]:

Show Cell

In [3]:

```
df.head(2)
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctio
0	6	148	72	35	0	33.6	0.62
1	1	85	66	29	0	26.6	0.35

Ajustando as colunas

Out[4]:

Show Cell

In [5]:

```
df.head(2)
```

Out[5]:

	quant_gravides	glicose	pressao_sangue	pele_triceps	insulina	imc	prob_diabetes_familia
0	6	148	72	35	0	33.6	0.627
1	1	85	66	29	0	26.6	0.351

Tendo ajustado o DataFrame, partiremos para a análise de dados, realizando verificações simples sobre balanceamento e buscando alguma relação entre os dados.
Partindo de uma visão mais simples, dividindo o nosso DataFrame em dois grupos principais: os que são diabéticos e os que não são, para comparar alguns números e buscar algum gráfico normal.

Out[6]:

Show Cell

```
df_diabetico.mean() - df_saudavel.mean()
```

```
quant_gravides            1.567672
glicose                  31.277463
pressao_sangue            2.640627
pele_triceps              2.500179
insulina                 31.543821
imc                       4.838337
prob_diabetes_familia     0.120766
idade                     5.877164
diabetico                 1.000000
dtype: float64
```

Out[7]:

Show Cell

Observamos que a insulina e a glicose possuem uma média maior para pessoas diabéticas. Esse fato já era esperado, para verificação se não existe possíveis outliers, compararemos a mediana:

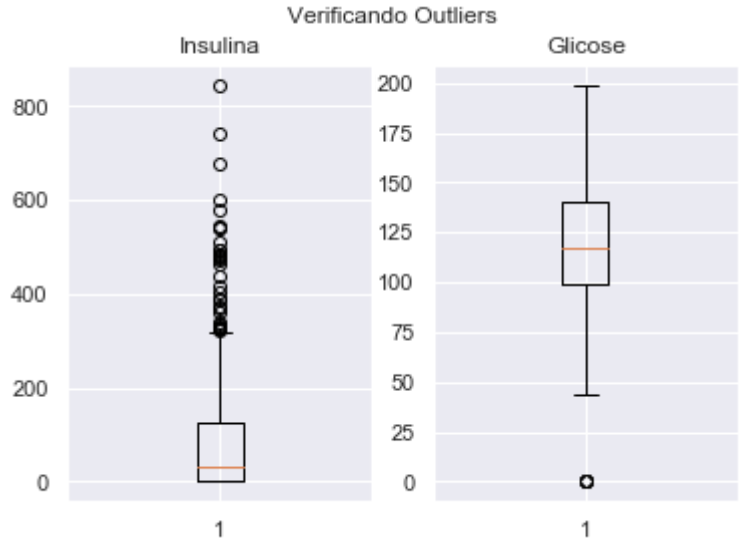
```
Saudáveis
Média glicose Saudaveis   : 109.98
Mediana glicose Saudaveis : 107.0
Média insulina Saudaveis  : 68.792
Mediana insulina Saudaveis: 39.0

Diabéticos
Média glicose Diabéticos   : 141.25746268656715
Mediana glicose Diabéticos : 140.0
Média insulina Diabéticos  : 100.33582089552239
Mediana insulina Diabéticos: 0.0

Vemos que Em geral não existe outliers (apenas no caso da insulina em diabét
icos)
```

Out[8]:

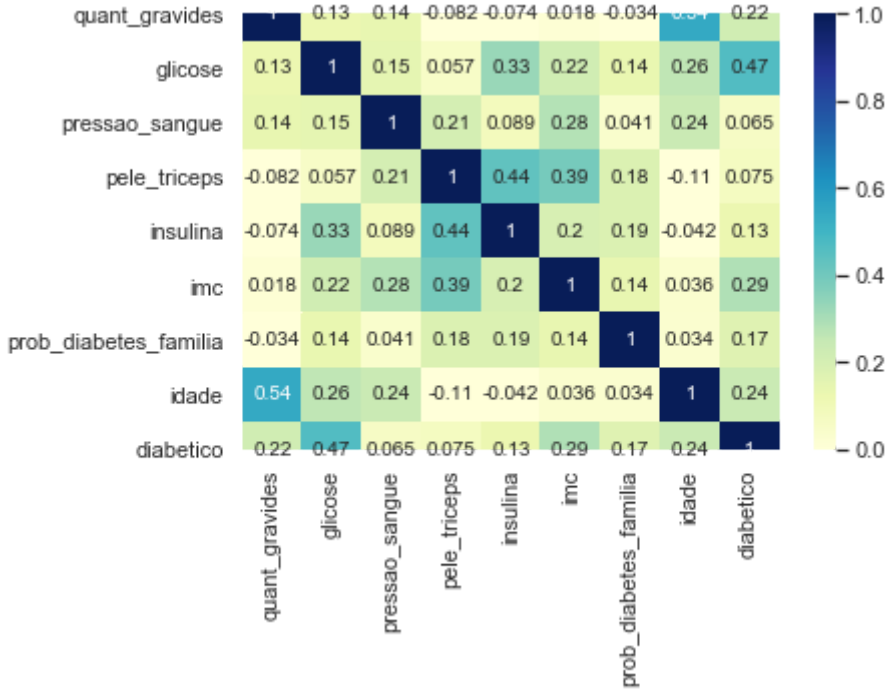
Show Cell



Out[9]:

Show Cell

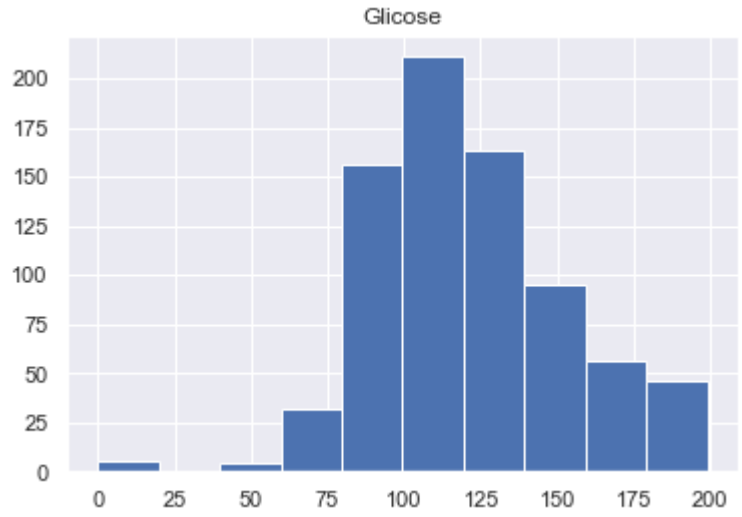
Vamos realizar uma verificação de correlação, para maior simplicidade e melhor visualização, plotaremos um gráfico de calor com as correlações



Out[10]:

Show Cell

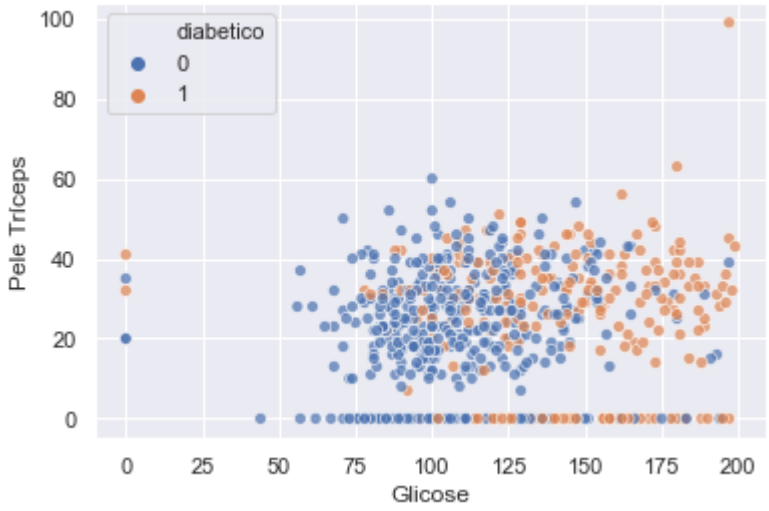
Vemos que não existe uma correlação forte entre nenhum dos dados, mas isso não significa que não podemos utiliza-los
Vamos explorar mais profundamente o data frame seccionando apenas para as partes importantes para nossa pergunta.
Sabemos que glicose é fator importante na diabetes, então utilizaremos sempre essa variável como target para comparação dos outros dados. Ela também segue a reta normal mesmo sem a aplicação do Bootstrap



Out[11]:

Show Cell

Primeira pergunta "A grossura da pele do triceps tem alguma relação com a chance de ter diabetes?"



Out[12]:

Show Cell

Não existe correlação aparente a partir do gráfico de dispersão

Tentaremos aplicar um modelo de regressão para um teste se realmente não são relacionados e não geram

nenhum resultado plausível

```
Aplicando o KNN

Acurácia: 0.7086614173228346

Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.77      0.80      0.78       168
           1       0.57      0.53      0.55        86

    accuracy                           0.71       254
   macro avg       0.67      0.67      0.67       254
weighted avg       0.70      0.71      0.71       254
```

Out[13]:

Show Cell

Out[14]:

Show Cell

```
Aplicando uma Regressão Logística
Acurácia: 0.7244094488188977


Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.76      0.86      0.80       168
           1       0.62      0.47      0.53        86

    accuracy                           0.72       254
   macro avg       0.69      0.66      0.67       254
weighted avg       0.71      0.72      0.71       254
```

Out[15]:

Show Cell

Utilizando um modelo de regressão observamos que a pele do triceps não é um fator totalmente determinante para a ocorrência do diabetes.

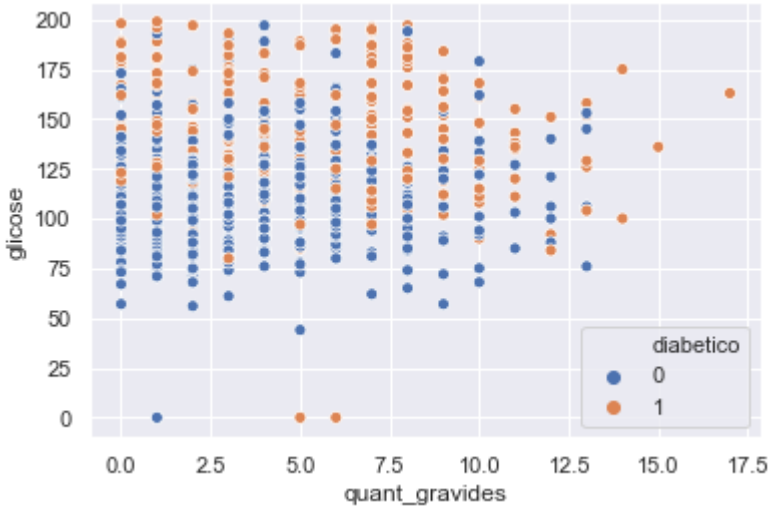
Segunda pergunta "A gravidez gera uma tendência a ter diabetes?"

Out[16]:

Show Cell

Out[17]:

Show Cell



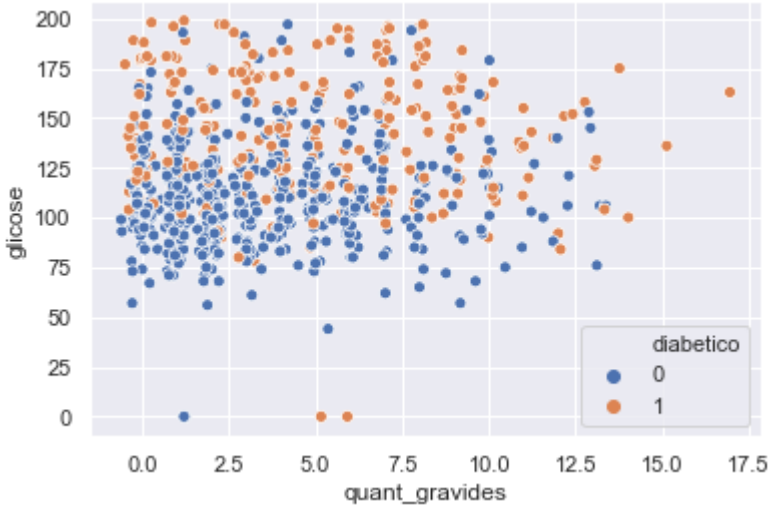
Foi adicionado um ruído, para verificação de possíveis clusteres.

Out[18]:

Show Cell

Out[19]:

Show Cell



Vemos que não existem clusteres e nem uma correção forte entre os eixos

Realizando o mesmo teste aplicado a pele do tríceps

```
Aplicando o KNN

Acurácia: 0.594488188976378

Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.72      0.64      0.68       168
           1       0.42      0.51      0.46        86

    accuracy                           0.59       254
   macro avg       0.57      0.57      0.57       254
weighted avg       0.62      0.59      0.60       254
```

Out[20]:

Show Cell

```
Aplicando uma Regressão Logística

Acurácia: 0.6850393700787402


Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.70      0.92      0.79       168
           1       0.59      0.23      0.33        86

    accuracy                           0.69       254
   macro avg       0.64      0.57      0.56       254
weighted avg       0.66      0.69      0.64       254
```

Out[21]:

Show Cell

Obtemos uma performace melhor do que relacionando a pele do tríceps. Porém não temos uma acurácia grande.

Terceira pergunta "Existe relação entre glicose, insulina, pressão sanguinea?"

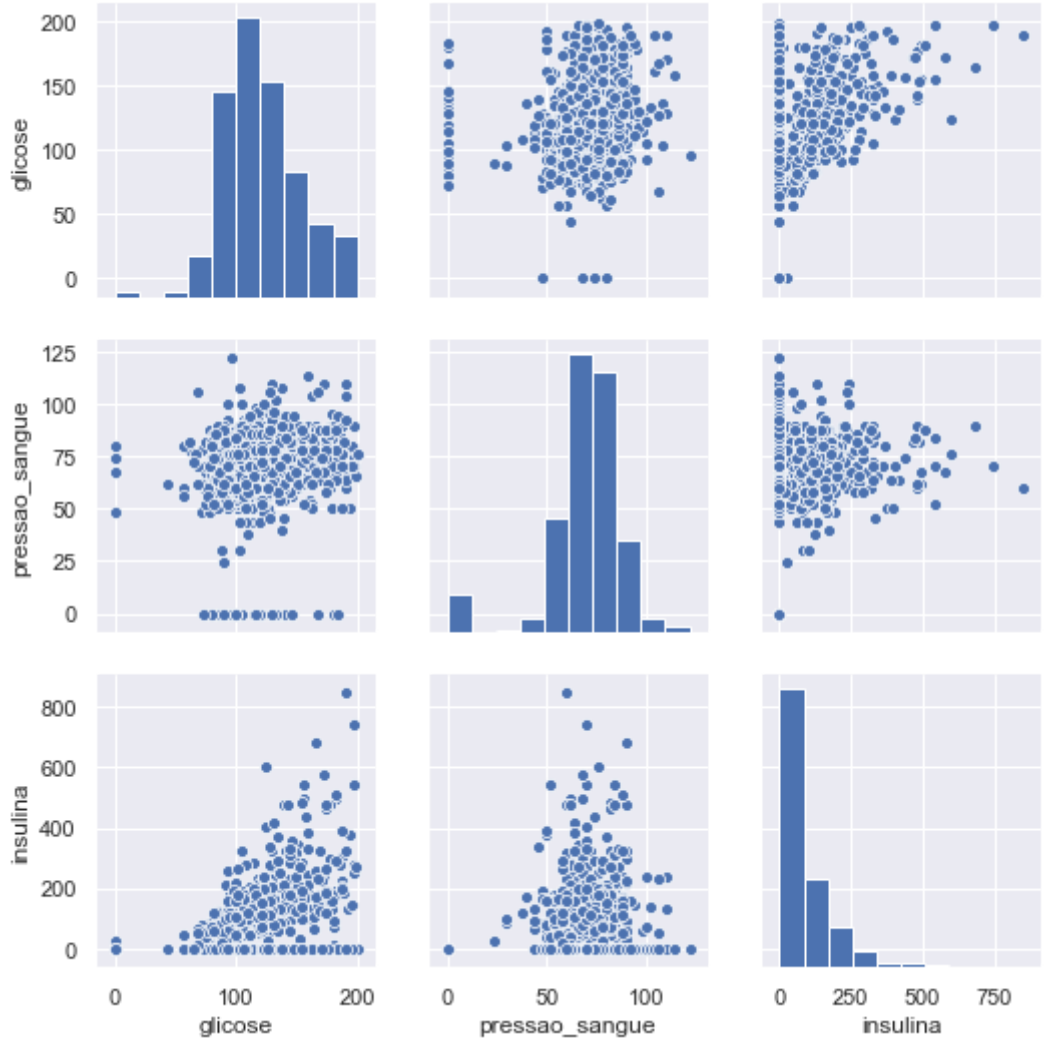
Out[22]:

Show Cell

```
                glicose  pressao_sangue  insulina
glicose        1.000000        0.152590  0.331357
pressao_sangue 0.152590        1.000000  0.088933
insulina       0.331357        0.088933  1.000000
```

Out[23]:

Show Cell



Out[24]:

Show Cell

Segundo o Portal da Educação (https://www.portaleducacao.com.br/conteudo/artigos/enfermagem/funcao-da-insulina-no-corpo/34860):

> A insulina promove o transporte de glicose para essas células de modo idêntico ao das outras células do corpo

Vemos que a influência existe e é biológica, porém no nosso dataframe ela não é forte e não é um fator que influência a pressão sanguínea da pessoa

Tentando obter melhores resultados com o DataFrame completo

```
Proporção dentro do Dataframe:
0    500
1    268
Name: diabetico, dtype: int64
```

Out[25]:

Show Cell

Nosso Dataframe está desbalanceado e ao aplicar algum modelo de Machine Learning ele provavelmente irá ficar enviesado. Para evitar isso, balancearemos o DataFrame.

```
0    265
Name: diabetico, dtype: int64
```

Out[26]:

Show Cell

```
1    268
0    265
Name: diabetico, dtype: int64
```

Out[27]:

Show Cell

Separando os dados de treino e teste.
Embaralhamos os dados dentro do Dataframe para que ao treinar o modelo tenha um pouco mais de dificuldade.

Out[28]:

Show Cell

Out[34]:

Show Cell

```
Regressão Logística

Acurácia: 0.7954545454545454


Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.85      0.77      0.81        98
           1       0.74      0.83      0.78        78

    accuracy                           0.80       176
   macro avg       0.80      0.80      0.79       176
weighted avg       0.80      0.80      0.80       176

Verificando a matriz de confusão
[[75 23]
 [13 65]]
```

Out[35]:

Show Cell

Out[36]:

Show Cell

```
Random Forest

Acurácia: 0.7215909090909091


Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.78      0.69      0.74        98
           1       0.66      0.76      0.71        78

    accuracy                           0.72       176
   macro avg       0.72      0.73      0.72       176
weighted avg       0.73      0.72      0.72       176

Verificando a matriz de confusão
[[68 30]
 [19 59]]
```

Out[40]:

Show Cell

Temos o RandomForest com desempenho melhor sobre todo o DataFrame, iremos testar o modelo dentro de todos os dados para ter uma acurácia geral final

```
Random Forest
Acurácia: 0.8255208333333334


Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.94      0.78      0.85       500
           1       0.69      0.91      0.78       268

    accuracy                           0.83       768
   macro avg       0.82      0.84      0.82       768
weighted avg       0.85      0.83      0.83       768

Verificando a matriz de confusão
[[391 109]
 [ 25 243]]
```

Out[41]:

Show Cell

Aplicando o modelo geral, obtemos uma acurácia de 83%!

Explicando o motivo de utilização final do RandomForest e não da Regressão (Mesmo com performace inicial melhor):

```
Regressão Logística sobre todos os dados
Acurácia: 0.74609375


Distribuições dos acertos

              precision    recall  f1-score   support

           0       0.87      0.72      0.79       500
           1       0.60      0.79      0.68       268

    accuracy                           0.75       768
   macro avg       0.73      0.76      0.74       768
weighted avg       0.77      0.75      0.75       768

Verificando a matriz de confusão
[[361 139]
 [ 56 212]]
```

Out[44]:

Show Cell

Acurácia do RandomForest é maior que a da Regressão!