PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

CRISTIANE A MASSENA - RA 1143503
JUCIANA RODRIGUES - RA 1146483
RICARDO GALIARDI - RA 1143795

GOVERNANÇA DE DADOS NOS DIAGNÓSTICOS DE DOENÇAS

Belo Horizonte
2019

CRISTIANE A MASSENA - RA 1143503
JUCIANA RODRIGUES - RA 1146483
RICARDO GALIARDI - RA 1143795

GOVERNANÇA DE DADOS NOS DIAGNÓSTICOS DE DOENÇAS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2019

SUMÁRIO

1. INTRODUÇÃO

1.1. Contextualização

Vazamentos de dados na área de saúde é um incidente grave, e já houveram casos de grande repercussão devido ao número de pessoas prejudicadas.

Em 2014, a Community Health Systems (CHS), grupo que administra 198 hospitais nos Estados Unidos, teve 4,5 milhões de dados de seus pacientes roubados por conta de falha em um software de criptografia.

A Prefeitura de São Paulo em 2016, por pura negligência, deixou aberto ao público dados registrados entre 2001 e 2007 de mais de 350 mil pacientes.

O IBGE divulgou que entre 2010-2015, houve um aumento no consumo final de bens e serviços de saúde que chegou aos R$ 546 bilhões, quase 10% do PIB nacional. [1]

Neste cenário promissor surgem os empreendimentos na saúde como *healthtechs (Health* de "saúde" em inglês e *tech*, abreviação de *technology*) – startups que aliam tecnologia para fornecer serviços inovadores, sendo um dos setores com maior potencial de crescimento no Brasil**.** Atualmente, são 8% do total de startups no país.

Nestas empresas o uso de dados é intenso e há a necessidade de gerenciar eficazmente estes recursos. Por meio de uma parceria da liderança de negócios e expertise técnica, a função de gestão de dados (Data Management Project) e proteção (Data Protection Officer) pode efetivamente fornecer e controlar ativos de dados e informações (DAMA-DMOBK, 2012).

Para criar as condições que promovam os resultados alinhados com a estratégia, as organizações intensificam esforços no sentido de promover a governança, gestão estratégica e proteção de dados com maior transparência e conformidade com regulamentações e boas práticas.

1 https://blog.nexxera.com/healthtech/

Face ao crescente volume de dados disponíveis para tomada de decisão pelas organizações exige o reconhecimento dos dados como ativos para criação da vantagem competitiva apoiado pela eficiência operacional.

1.2. O problema proposto

Big Data e o Machine Learning podem ajudar a revolucionar a descoberta de medicamentos, o tratamento de doenças como diabetes, câncer entre outros. Dados abertos e recursos computacionais são usados pelos pesquisadores para descobrir novos usos para drogas. Existem entidades que usam tecnologias de Big Data e algoritmos de Aprendizado de Máquina (Machine Learning) para descobrir que um medicamento usado para tratar vermes poderia reduzir um carcinoma que era um tipo de câncer de fígado em camundongos. Este carcinoma em particular, foi o segundo maior contribuinte de mortes por câncer no mundo.

Diante desse contexto, o uso de dados pessoais são vitais para que haja a acurácia nos resultados que podem garantir o sucesso do tratamento do paciente. Dados de saúde são considerados “dados pessoais sensíveis” à luz do artigo 5º, II da LGPD. Assim sendo, as hipóteses para o tratamento desses dados sem o fornecimento do consentimento pelo titular deve obedecer a requisitos com maior nível de rigor (artigo 11º da LGPD) comparando-se ao tratamento de outros dados pessoais (artigo 7º da LGPD).

O inciso II do artigo 11º da LGPD estabelece as situações em que se permite o tratamento de dados pessoais sensíveis sem o consentimento do titular nas hipóteses específicas em que essa dispensa for realmente indispensável. Garantir a segurança destes dados é o desafio no qual a governança de dados contribui de forma efetiva e para sua adequação às leis.

2. Coleta de Dados

Os dados foram obtidos do site oficial dos dados livres para uso em machine learning - UCI.

- https://archive.ics.uci.edu/ml/datasets/diabetes

Visão geral dos dados importados

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age	Class
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1

Visão do tipo de dados e campos

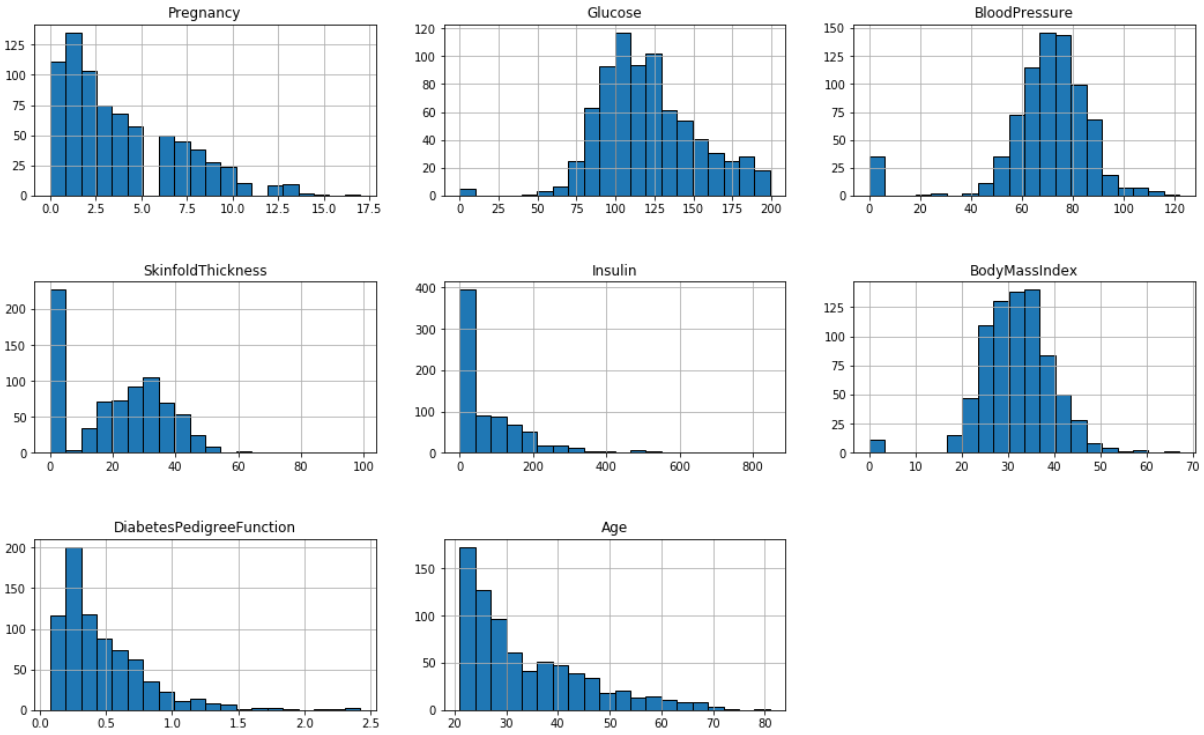
```
Data columns (total 9 columns):
Pregnancy                 768 non-null int64
Glucose                   768 non-null int64
BloodPressure             768 non-null int64
SkinfoldThickness         768 non-null int64
Insulin                   768 non-null int64
BodyMassIndex             768 non-null float64
DiabetesPedigreeFunction  768 non-null float64
Age                       768 non-null int64
Class                     768 non-null int64
dtypes: float64(2), int64(7)
```

3. Processamento/Tratamento de Dados

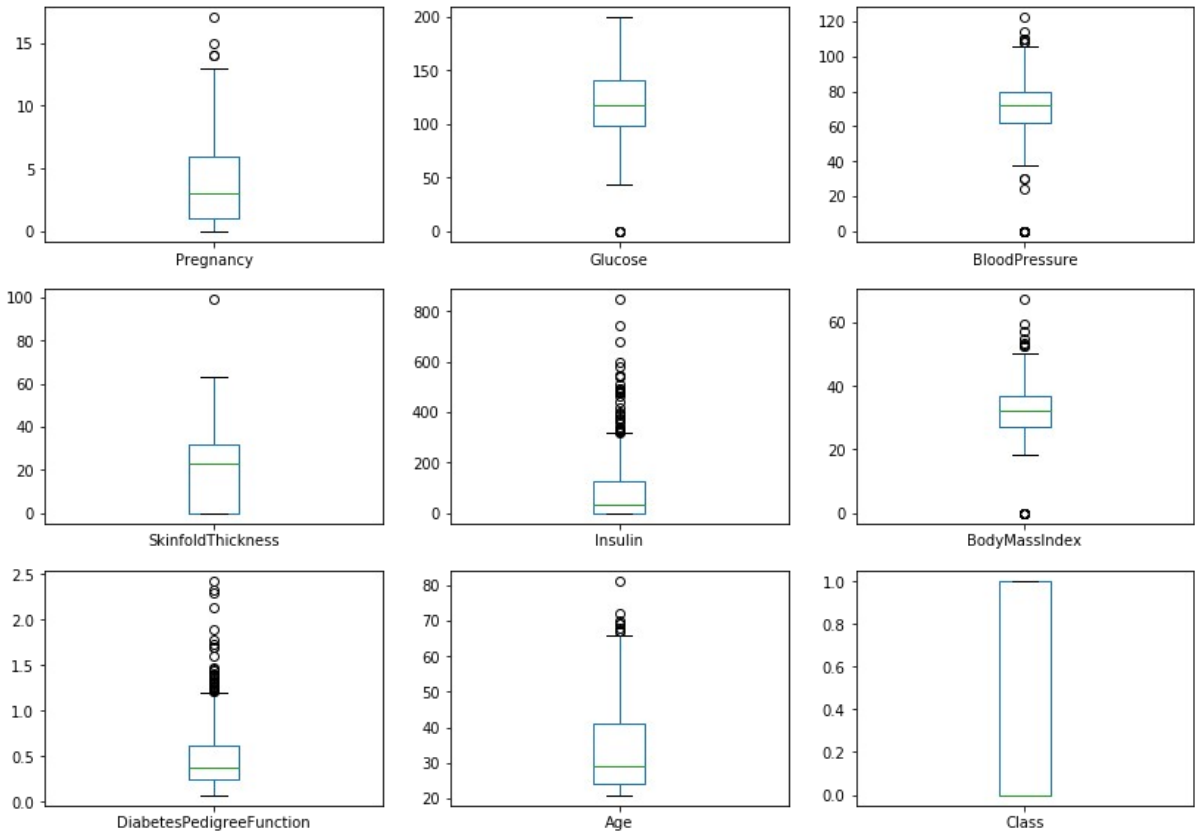
1. Resumo estatístico dos dados

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age	Class
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

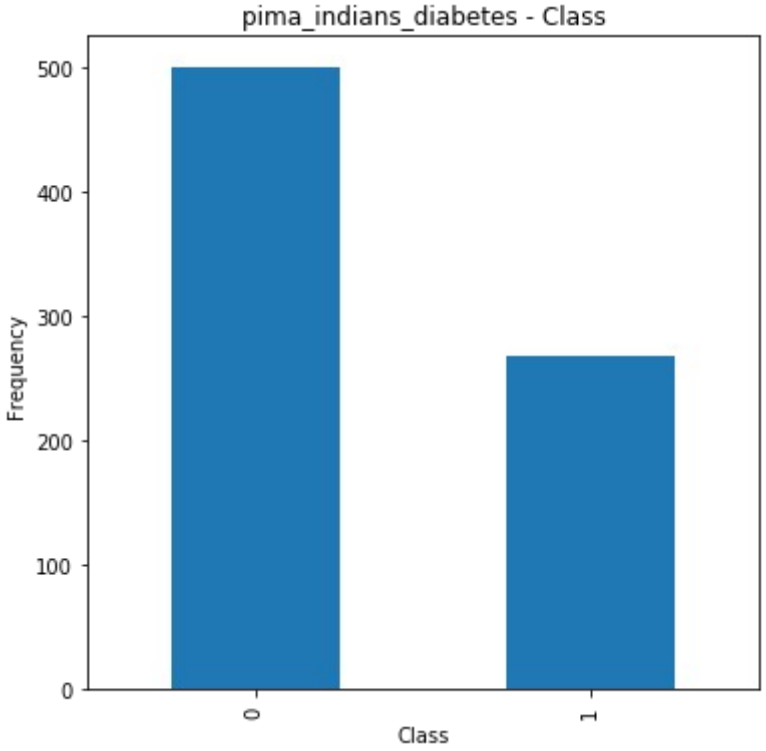
2. Análise dos Dados - Distribuição



3. Análise dos Dados - Outliers

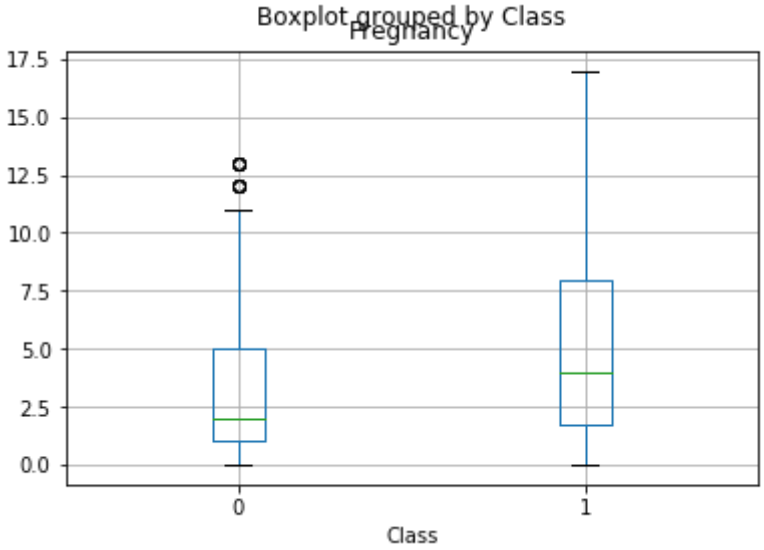


4. Análise dos Dados - Variável alvo e Frequência

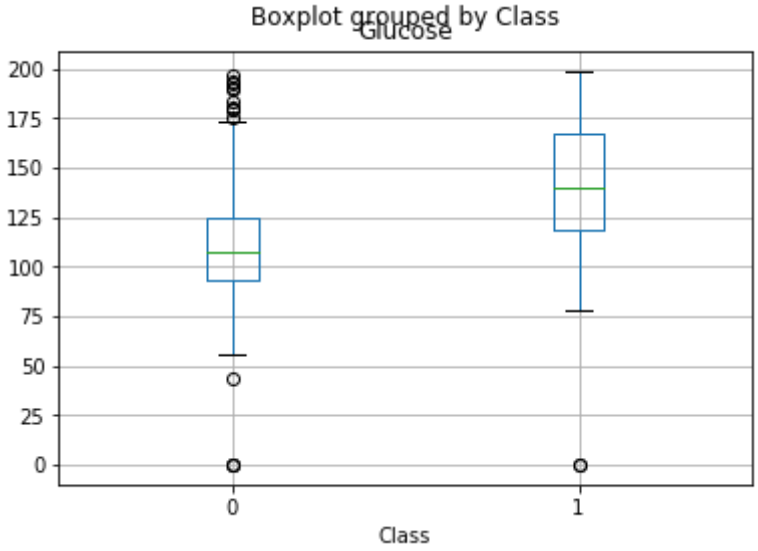


5. Relacionamento entre a variável alvo e seus atributos

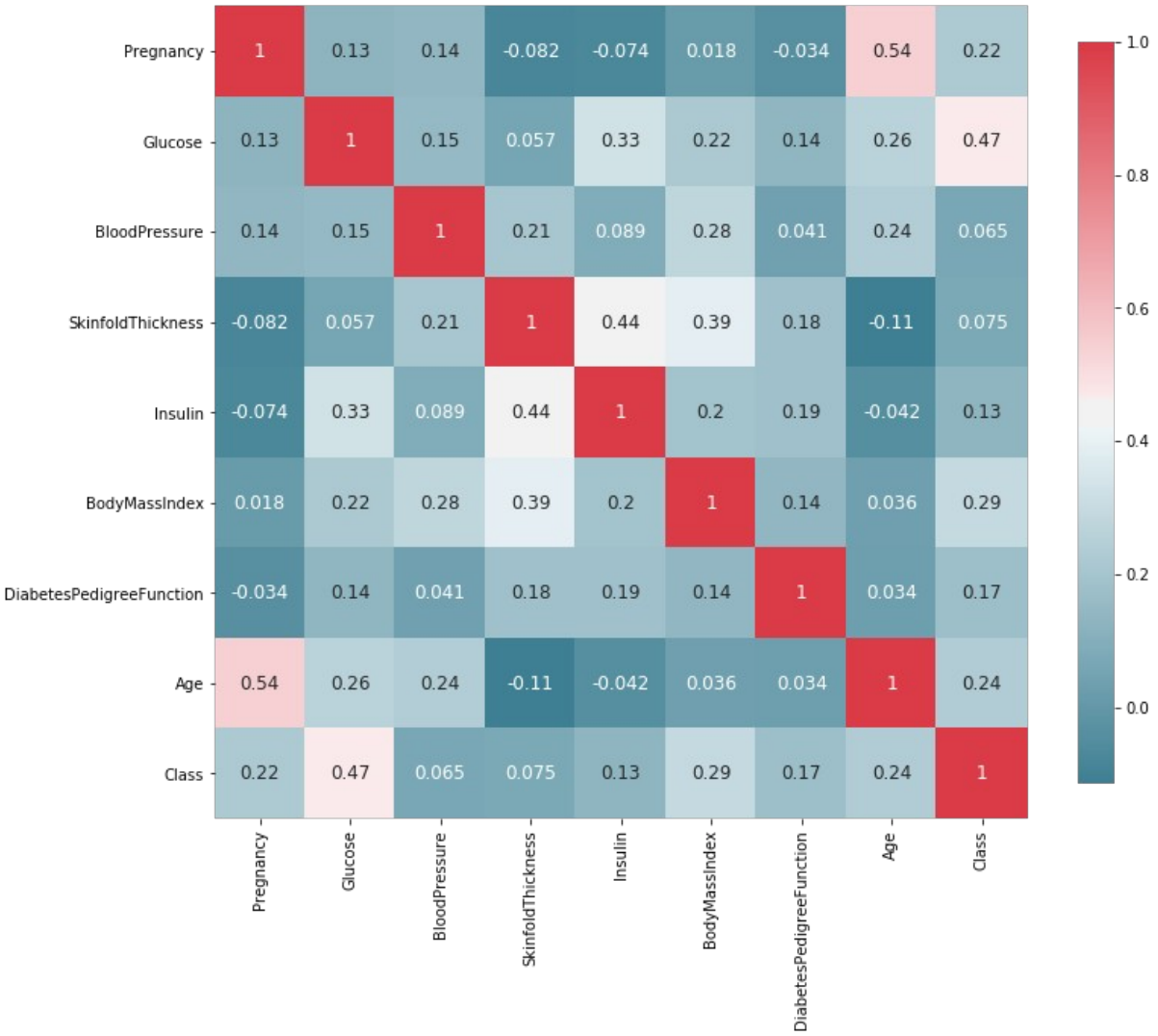
Classe x Gravidez



Classe x Glicose



6. Matriz de Correlação



7. Relacionamento entre Colunas

8. Limpeza dos Dados

Analisando os valores “Zero”

```
Número de valores missing: 5
Class
0    3
1    2
Name: Class, dtype: int64
```

Foram substituídos os valores zero pelo valor da média das classes, seguindo as melhores práticas para redução de variáveis nulas que podem comprometer o modelo.

Os campos tratados forma:

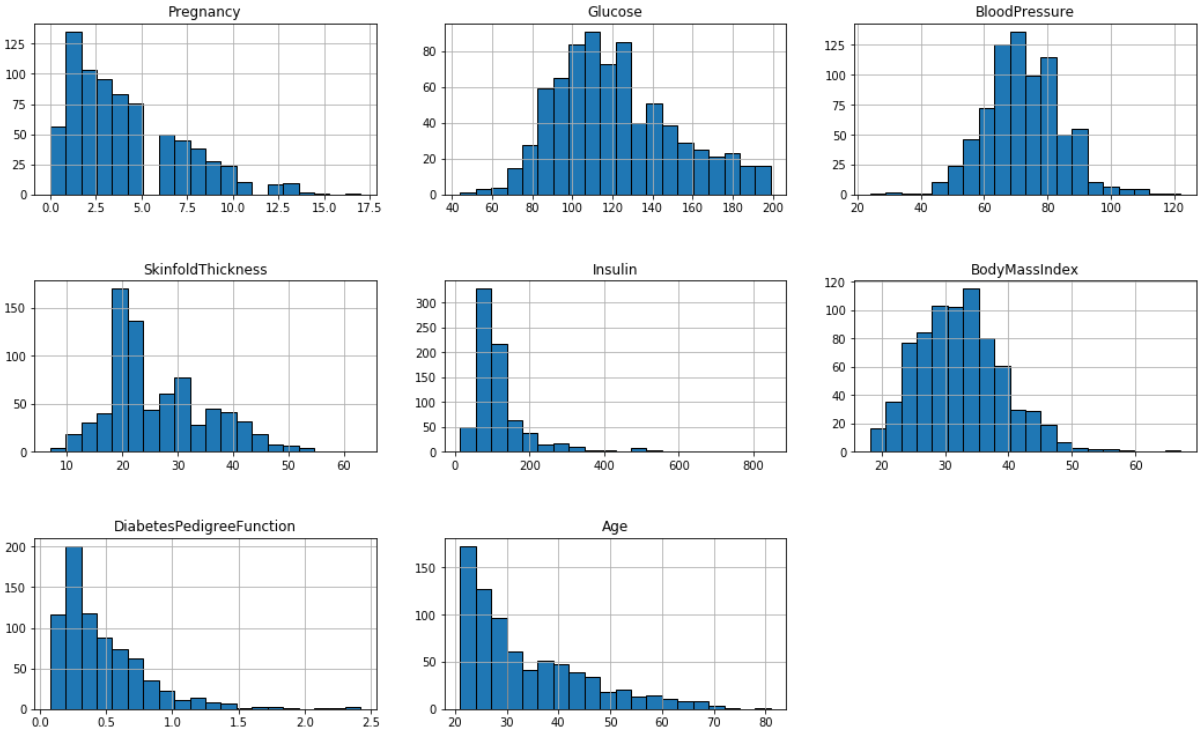
- Class - Classe
- BloodPressure - Pressão do Sangue
- SkinfoldThickness - Espessura da dobra cutânea
- Insulin - Insulina
- BodyMass - Peso do paciente

9. Avaliação preliminar após a limpeza dos dados

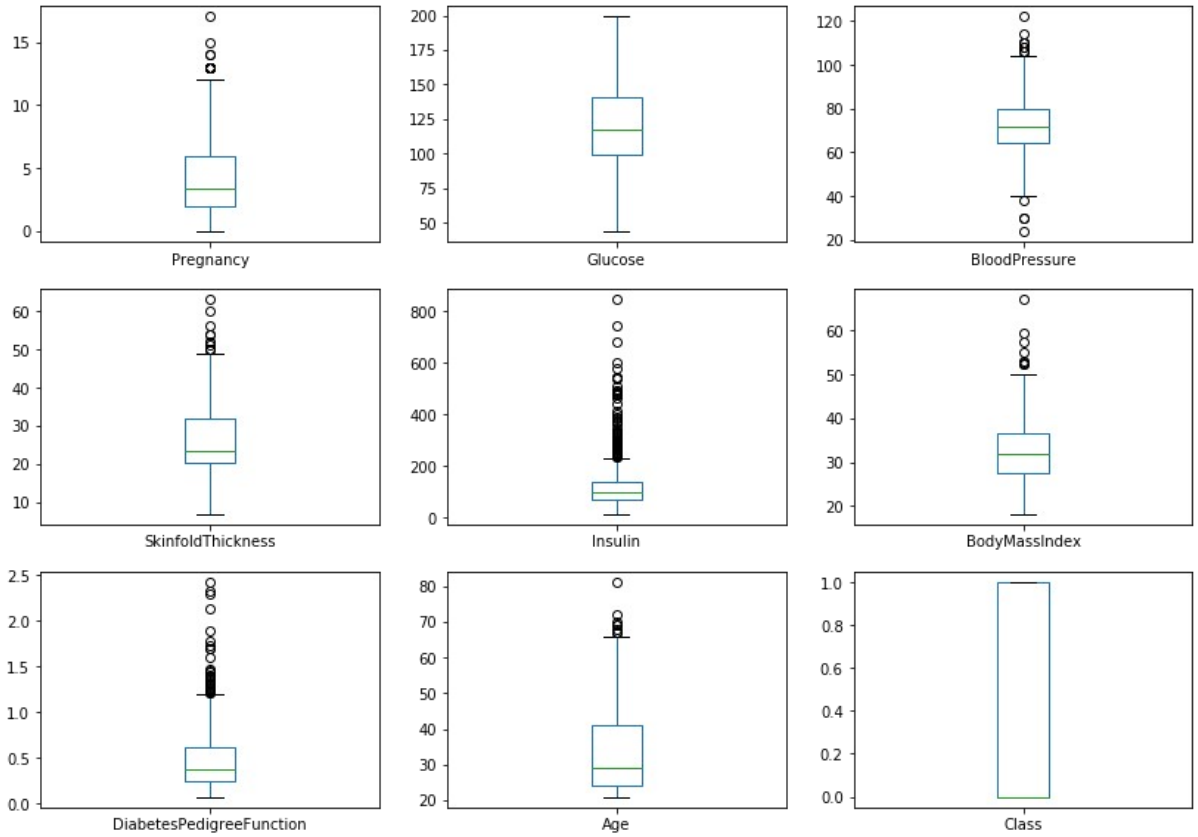
Análise dos Dados - Resumo Estatístico

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age	Class
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	4.126263	121.691999	72.267826	26.770604	124.771038	32.441053	0.471876	33.240885	0.348958
std	3.202732	30.461151	12.115948	9.144460	91.935806	6.880054	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	2.000000	99.750000	64.000000	20.371904	71.691720	27.500000	0.243750	24.000000	0.000000
50%	3.324384	117.000000	72.000000	23.404712	100.000000	32.050000	0.372500	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	136.297569	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	63.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Análise dos Dados - Distribuição



Análise dos Dados - Outliers



4. Análise e Exploração dos Dados

Contagem dos valores da variável target

```
0    500
1    268
Name: Class, dtype: int64
```

Os passos abaixo foram seguidos para ajuste das variáveis para que não houvessem alterações no modelo proposto.

- Criando a variável para manter a distribuição sempre padrão
  - Classe 0 - Total 500
  - Classe 1 - Total 268
- Valida a quantidade de linhas e colunas das variáveis preditoras
  - Shape: 1000 - 9
- Valida a quantidade de linhas e colunas da variável alvo
  - Shape: 1000 - 0
- Contagem de registros da variável alvo
  - Classe 0 - 500

- Classe 1 - 500

Análise dos Dados - Resumo Estatístico

	Pregnancy	Glucose	BloodPressure	SkinfoldThickness	Insulin	BodyMassIndex	DiabetesPedigreeFunction	Age
1	0.352941	0.670968	0.489796	0.500000	0.146992	0.314928	0.234415	0.483333
2	0.058824	0.264516	0.428571	0.392857	0.090194	0.171779	0.116567	0.166667
3	0.470588	0.896774	0.408163	0.292941	0.111922	0.104294	0.253629	0.183333
4	0.058824	0.290323	0.428571	0.285714	0.096154	0.202454	0.038002	0.000000
5	0.000000	0.600000	0.163265	0.500000	0.185096	0.509202	0.943638	0.200000

5. Criação de Modelos de Machine Learning

Utilizando o pacote estatístico para programação em Python - Sklearn, forma mapeados alguns dos melhores algoritmos para classificação e solução do problema para predição da doença do tipo "***Diabetes***".

Abaixo os modelos apurados e avaliados para solução:

- **K-Nearest Neighbours Classifier - SMOTE**
  - Classificação padrão:
    - Matriz de Confusão:

```
[[80 23]
 [11 86]]
```

    - Resultado:

```
              precision    recall  f1-score   support

           0       0.88      0.78      0.82       103
           1       0.79      0.89      0.83        97

 avg / total       0.84      0.83      0.83       200
```

  - Classificação após normalização:
    - Matriz de Confusão:

```
[[103   0]
 [  0  97]]
```

    - Resultado:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	103
1	1.00	1.00	1.00	97
avg / total	1.00	1.00	1.00	200

- Resumo:

  Após a normalização dos dados tivemos uma melhor significativa no resultado apurado.

- **K-Nearest Neighbours Classifier - SEM SMOTE**
  - Classificação padrão:
    - Matriz de Confusão:

```
[[83 17]
 [10 44]]
```

    - Resultado:

	precision	recall	f1-score	support
0	0.89	0.83	0.86	100
1	0.72	0.81	0.77	54
avg / total	0.83	0.82	0.83	154

- **Decision Tree Classifier**
  - Classificação padrão:
    - Matriz de Confusão:

```
[[79 21]
 [16 38]]
```

    - Resultado:

	precision	recall	f1-score	support
0	0.83	0.79	0.81	100
1	0.64	0.70	0.67	54
avg / total	0.77	0.76	0.76	154

- **Random Forest Classifier**
  - Classificação padrão:
    - Matriz de Confusão:

```
[[84 16]
 [10 44]]
```

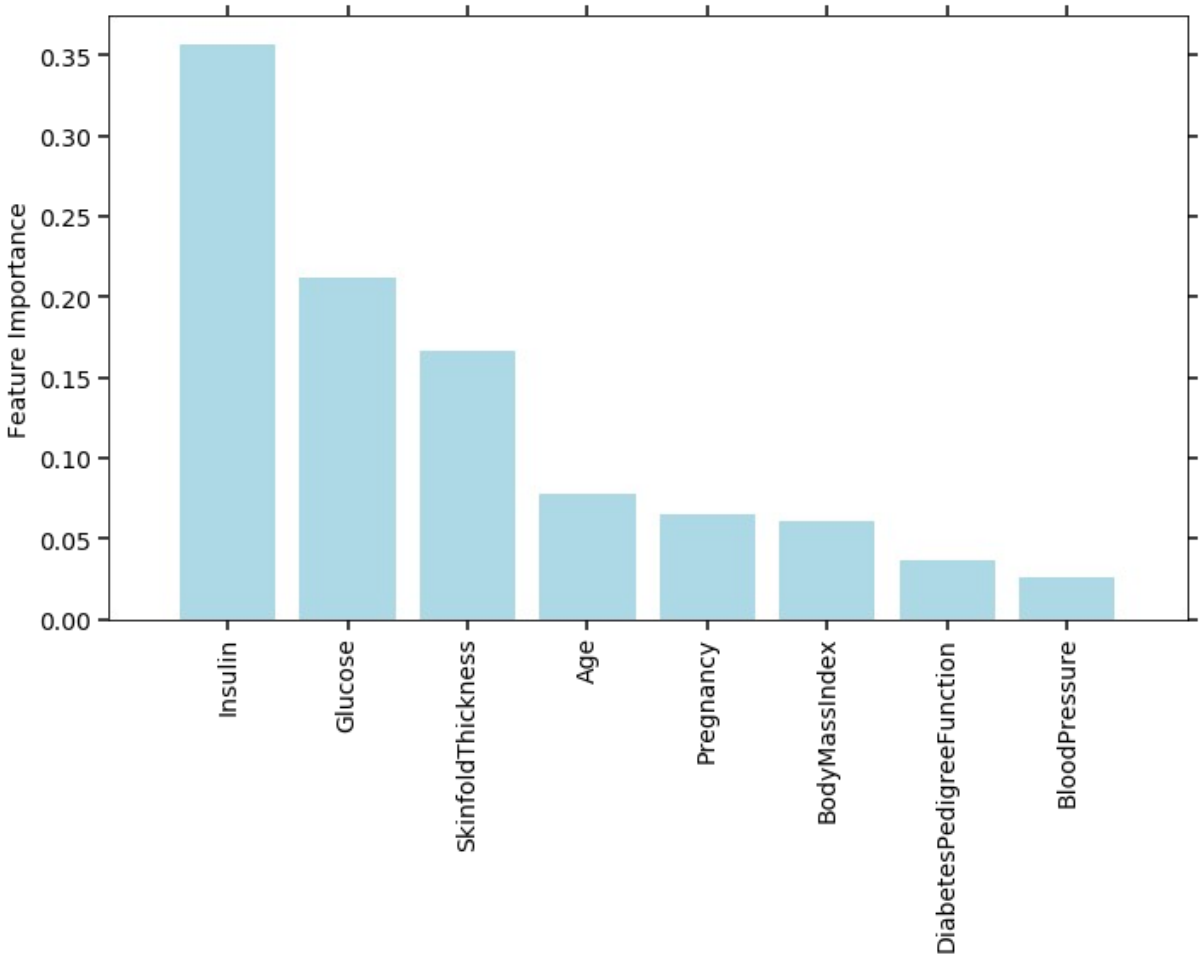
- Resultado:

	precision	recall	f1-score	support
0	0.89	0.84	0.87	100
1	0.73	0.81	0.77	54
avg / total	0.84	0.83	0.83	154

6. Apresentação dos Resultados

Avaliando as variáveis mais relevantes como característica para a predição do Diabetes, tendo como base as variáveis disponíveis na base temos:

- **Variáveis analisadas:**
  - Pregnancy - Gravidez
  - Glucose - Glicose
  - BloodPressure - Pressão Sanguinea
  - SkinfoldThickness - Espessura da dobra cutânea
  - Insulin - Insulina
  - BodyMassIndex - Indice de Massa Corporal
  - DiabetesPredigreeFuncion - Função para Previsão do Diabetes
  - Age - Idade
- **Variáveis representativas:**
  - Insulin - Insulina foi a única variável com maior representatividade e característica mais importante para a evolução da doença.

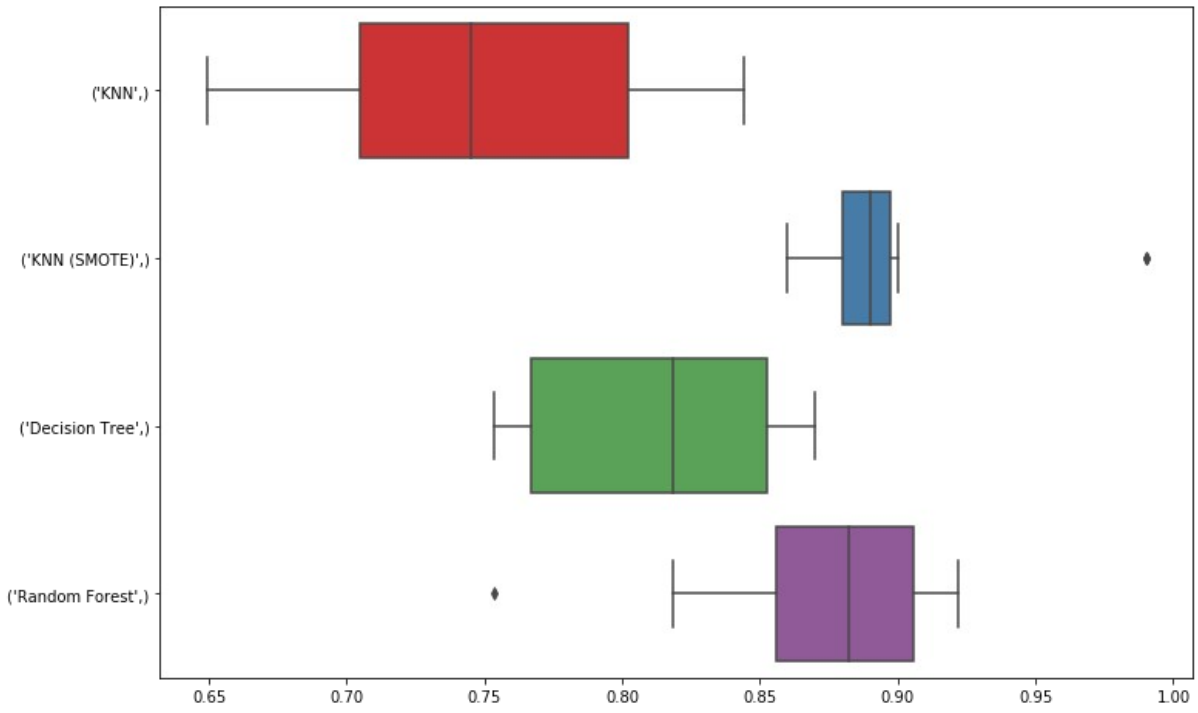


- **Acurácia entre os Modelos utilizados**

	Accuracy
Decision Tree	0.753247
Random Forest	0.831169
KNN	0.818182
KNN (Smote)	0.895000

- **Utilizando o métodos de “Cross Validation”**

	CV Mean
KNN	0.753794
KNN (SMOTE)	0.904000
Decision Tree	0.802085
Random Forest	0.871138



Resumo

Após a exploração dos modelos de classificação mais importantes ou representativos para solução do problema proposto - Previsão de Desenvolvimento da doença do Diabetes por pacientes gerais em seus resultados clínicos pudemos observar que o melhor resultado foi apresentado com o modelo k-nearest neighbors - KNN (SMOTE).

O modelo “KNN” é um modelo supervisionado para classificação simples, usado para classificar objetos com base em exemplos de treinamento que estão mais próximos no espaço de características.

$$\underset{X_{i,j} \subset X}{argmin} \sum_{i=0}^{K} \sqrt{\sum_{j=o}^{N} (X_{i,j} - t_i)^2}$$

7. Conclusão Final

Em nosso trabalho realizamos a aplicação de um modelo de classificação para detecção de doenças do tipo Diabetes, e levamos em consideração a aderência da exploração dos dados em conformidade com a LGPD (Lei Geral de Proteção aos Dados) que entrará em vigor em 20 de agosto de 2020.

Todos os dados apresentados foram revisados e qualificados para que não houvessem exposição de informações sensíveis em suas utilizações. Bem como a não identificação das suas respectivas informações aos seus proprietários - pacientes.

Atualmente os modelos de “machine learning” ou modelos de aprendizado de máquina pode nos auxiliar em uma infinidade de tarefas e soluções de problemas corporativos ou não. Cabe às empresas ou profissionais a utilização de seu uso consciente para que não sejam infringidas as leis vigentes nos países de origem das pesquisas ou trabalhos.

8. Links

- https://drive.google.com/drive/folders/1swx0b0C5TIGgd53yA_MIPWGxTF95Po-9
  - **Documento**: TCC - Machine Learning e Governança em Saude.pdf
  - **Código Fonte**: Classification-Diabetics.ipynb
  - **Código Documentado**: Classification-Diabetics.pdf
  - **Vídeo**: classification-diabetics.mp4

REFERÊNCIAS

BARATA, André Montoia. **Governança de dados em organizações brasileiras: uma avaliação comparativa entre os benefícios previstos na literatura e os obtidos pelas organizações.** 2015. 154 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015