AA1

Apresentação do trabalho prático

14 de Dezembro de 2018

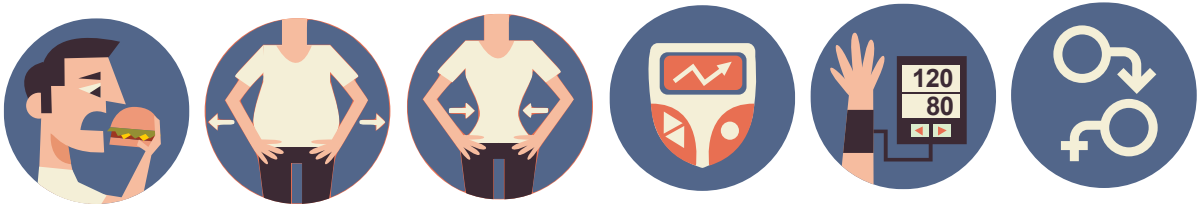
Dataset

Informações do Dataset:

- ✓ Nome: Diabetes
- ✓ Origem: EUA
- ✓ Data: 1997
- ✓ Quantidade: 403 obs. e 19 var

Variáveis

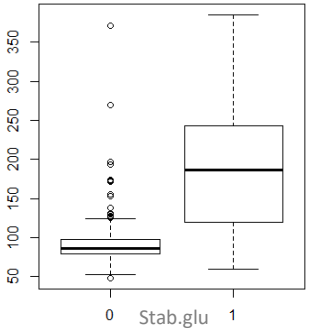
- ✓ Stab.glu
- ✓ Chol
- ✓ Time.ppn
- ✓ Hip
- ✓ Weight
- ✓ Height
- ✓ Bp.1s
- ✓ Bp.1d
- ✓ Bp.2s
- ✓ Bp.2d
- ✓ Gly.hb
- ✓ Ratio
- ✓ Frame
- ✓ Location
- ✓ Hdl
- ✓ Gender
- ✓ Waist

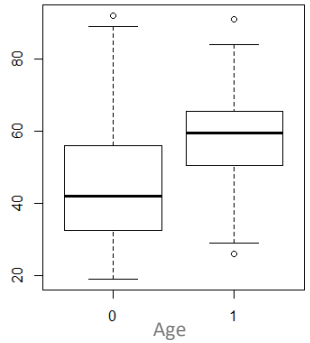


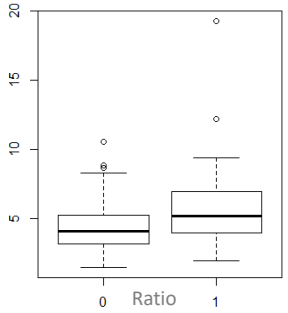
Análise exploratória

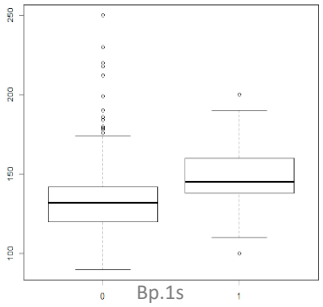
Análise de influencia

Var Binária

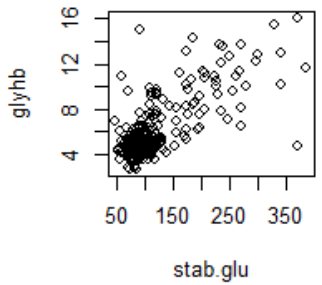


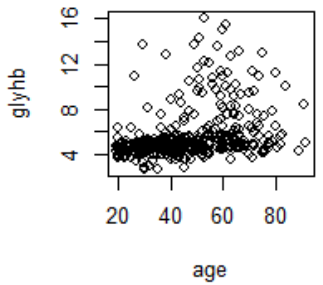


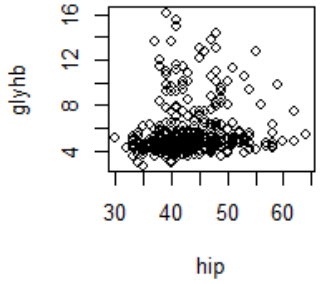


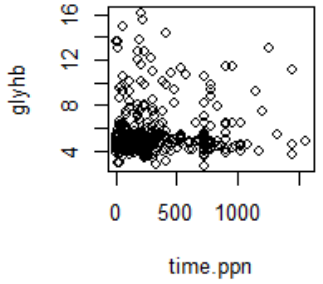


Gly.hb









Ranking variáveis

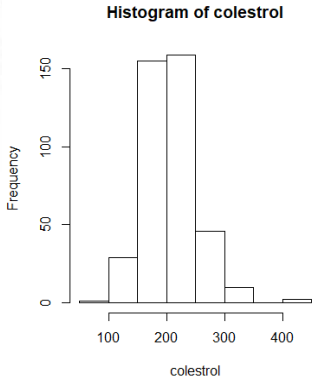
✓ Var Binária

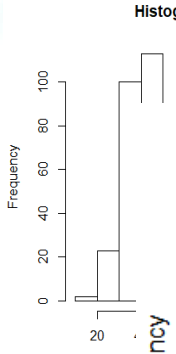
1. Stab.glu
2. Age
3. Bp.1s
4. Racio
5. Waist
6. Weight
7. Hip
8. Hdl
9. Chol
10. Bp.1d

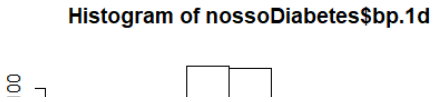
✓ Gly.hb

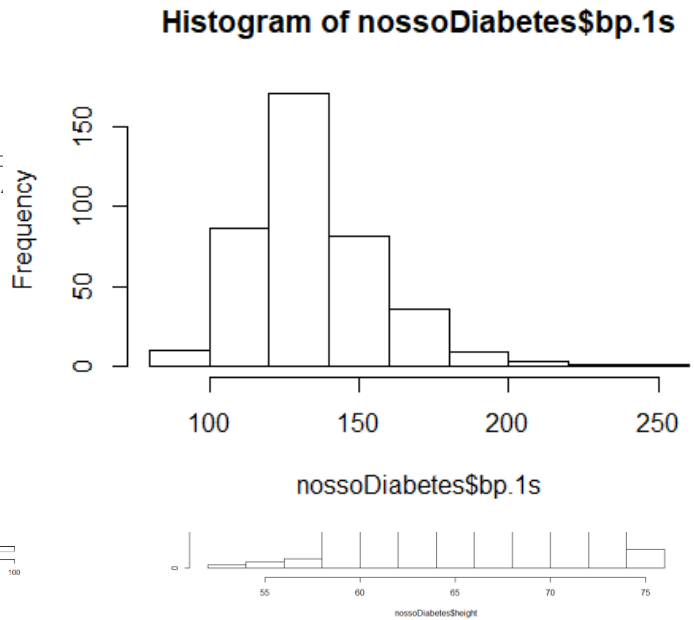
1. Stab.glu
2. Age
3. Time.ppn
4. Hip
5. Bp.1s
6. Waist
7. Bp.1d
8. Chol
9. Height
10. Weigth

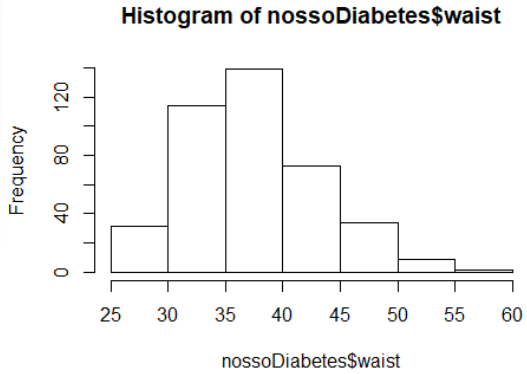
Análise de normalidade

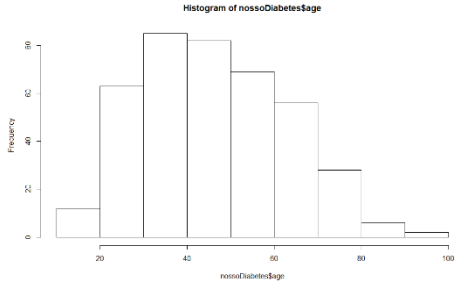


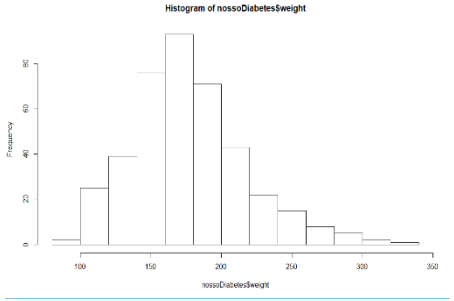












Seleção do modelo

LM vs GLM vs QDA vs KNN vs LDA

Regressão linear (LM)

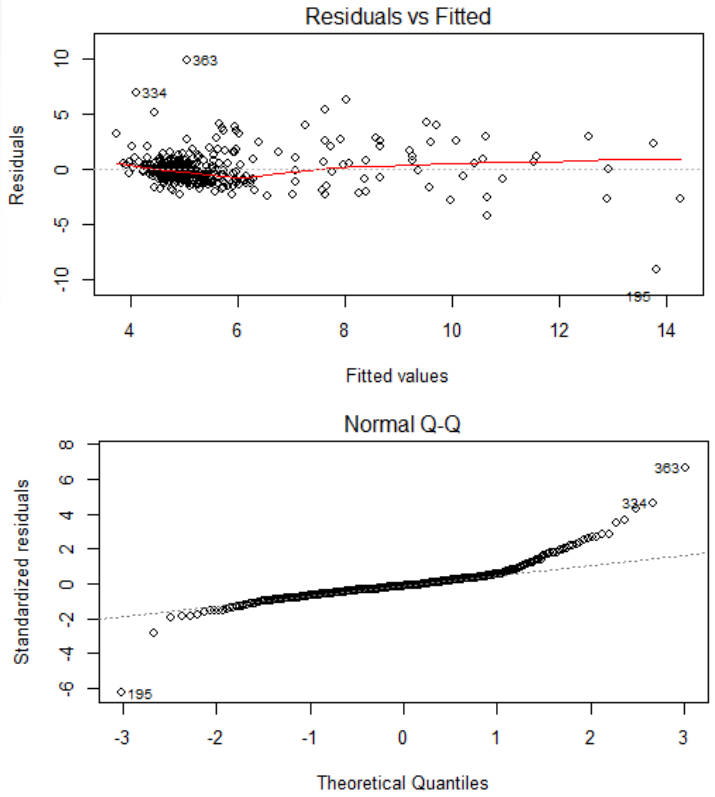
Modelos:

Dataset:

- stab.glu
  Adj. R-squared: 0.5602
- stab.glu^2
  Adj. R-squared: 0.5678
- stab.glu^3
  Adj. R-squared: 0.5877

CV(K=10):

- stab.glu
  Adj. R-squared: 0.5678
- stab.glu^2
  Adj. R-squared: 0.5621
- stab.glu^3
  Adj. R-squared: 0.5581

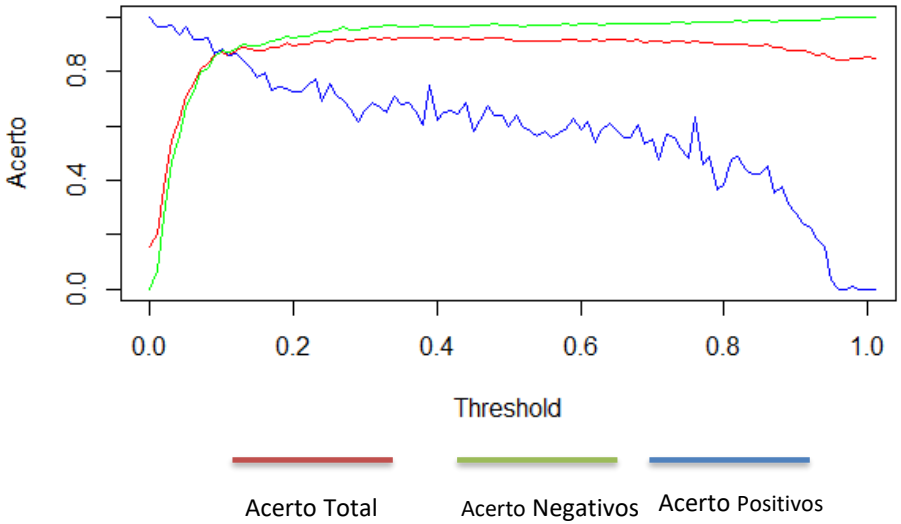


Regressão logística (GLM)

Modelos:

Fórmula	Thr	Acerto
Stab.glu^2	0.39	3.5714
Stab.glu^2+ratio	0.39	3.5714
Stab.glu+ratio + stab.glu x ratio	0.39	3.5520

Gráfico Threshold:



KNN

Nossa abordagem

Variáveis	K	Acerto ponderado
Stab.glu	14	3.4825
Stab.glu+age	9	3.4707
Stab.glu+age+bp.1s	10	3.5320
Stab.glu+age+bp.1s+ratio	12	3.5316
Stab.glu+age+bp.1s+ratio+waist	10	3.5307

Regsubsets

Variáveis	K	Acerto ponderado
Stab.glu+chol	9	3.4902
Stab.glu+ratio+age	16	3.4766
Stab.glu+ratio+age+time.ppn	3	3.4286
(F)stab.glu+chol+age	8	3.4684
(F)stab.glu+chol+age+time.ppn	4	3.4014
(B)stab.glu+ratio	15	3.4820

QDA E LDA

QDA

Variáveis	Threshold	Acerto Ponderado
Age	0.22	2.8265
Age+bp.1s	0.24	2.8260
Age+bp.1s+ waist	0.18	2.9115

LDA

Variáveis	Threshold	Acerto Ponderado
Age	0.18	2.8578
Age+bp.1s	0.26	2.8039
Age+bp.1s+ waist	0.16	2.9364

Seleção melhor modelo (CV)

Modelo	Regressão linear	Regressão logística	KNN
Fórmula	Stab.glu + age	stab.glu^2	stab.glu + age + bp.1s
Threshold	>7	0.39	K=10
Acerto ponderado	$adj.r^2 = 56.23\%$	**3.5714**	3.5320

Seleção melhor modelo
(Teste e Treino)

Modelo	Regressão linear	Regressão logística	KNN
Fórmula	Stab.glu + age	stab.glu^2	stab.glu + age + bp.1s
Threshold	>7	0.39	K=10
Acerto ponderado	**3.6318**	3.5991	3.5665

Modelo escolhido

Modelo: Regressão linear

- ✓ **Fórmula: glyhb = 1.693 + 0.027*stab.glu + 0.020*age**
- ✓ **Adjusted R^2 : 56.23%**
- ✓ **Acerto ponderado: 3.63**
- ✓ **Acerto total: 95.88%**
- ✓ **Acerto positivos: 71.43%**
- ✓ **Acerto negativos: 100%**

Discussão de resultados

- **Será que o colesterol/pressão arterial/tempo após refeição/fatores corporais afetam os diabetes?**
- **Qual fator corporal explica melhor o valor da diabete?**
- **Quais fatores influenciam mais o resultado final?**
- **De que forma os fatores selecionados para a explicação dos resultados o influenciam? (crescentemente, decrescentemente, linearmente)**

Discussão de resultados

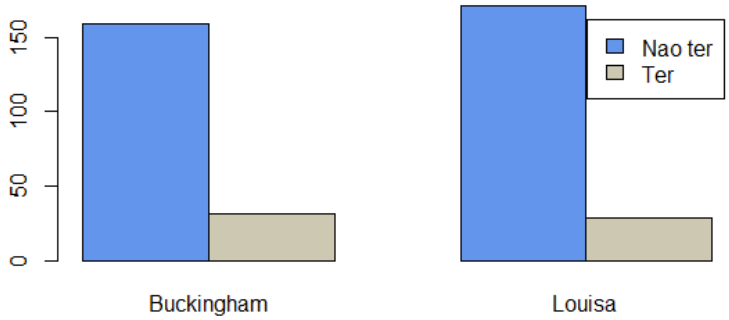
? **Qual a probabilidade (ou confiança no resultado) de uma pessoa com as caraterísticas X (por exemplo colesterol=180, altura=175, peso=67, etc.) ter diabetes?**

Stab.glu	Age	Confiança
151	60	0.5%
171	60	29.9%
191	60	54.7%
211	60	73.8%
231	60	86.3%

Stab.glu	Age	Confiança
191	20	16.7%
191	40	37%
191	60	54.7%
191	75	66%

Discussão de resultados

- **Qual a taxa de incidência em pessoas com menos e com mais de 50 anos?**
- **Qual a cidade apresenta maior incidência? (visto serem só dois podemos comparar)**
- **O resultado é mais exato utilizando um modelo de classificação ou de regressão (e de seguida classificando)?**



Incidência em pessoas com mais/menos do que 50 anos

- ✓ 94.22% das pessoas que têm menos de 50 anos não têm diabetes
- ✓ 70,4% das pessoas que têm mais de 50 anos não têm diabetes
- ✓ 42.55% dos que não têm diabetes têm mais de 50 anos
- ✓ 83.54% dos que têm diabetes têm mais de 50 anos

Q & A