Verão 2017 - IME

Introdução a Análise Exploratória de Dados e Métodos estatísticos



Thiago Costa

Personal Trainer

Preocupação com saúde

Expertice em diabetes

Estatística Exploratória



Cleibson Almeida

Militar

Preocupação com saúde

Expertice em diabetes

Estatística Exploratória



Reginaldo Florentino

Arquiteto

Preocupação com saúde

Expertice em diabetes

Estatística Exploratória

<



Análise Descritiva de Dados e Aplicação da Regressão Logística

PESQUISADORA: Dra. Denise Beheregaray

INSTITUIÇÃO: ADJ – Associação de Diabetes Juvenil

FINALIDADE DO PROJETO: Publicação

RESPONSÁVEIS PELA ANÁLISE:

Professora Gisela Tunes Professora Viviana Giampaoli Eduardo Carvalho da Silva William Nilson de Amorim

REFERÊNCIA DO TRABALHO:

TUNES, G.; GIAMPAOLI, V; SILVA, E. C.; AMORIM, W. N. Relatório de análise estatística sobre o projeto: "Avaliação do Findrisk como instrumento de detecção de diabetes de tipo 2 em um estudo transversal". São Paulo, IME-USP, 2012. (ERA-CEA -12P34).

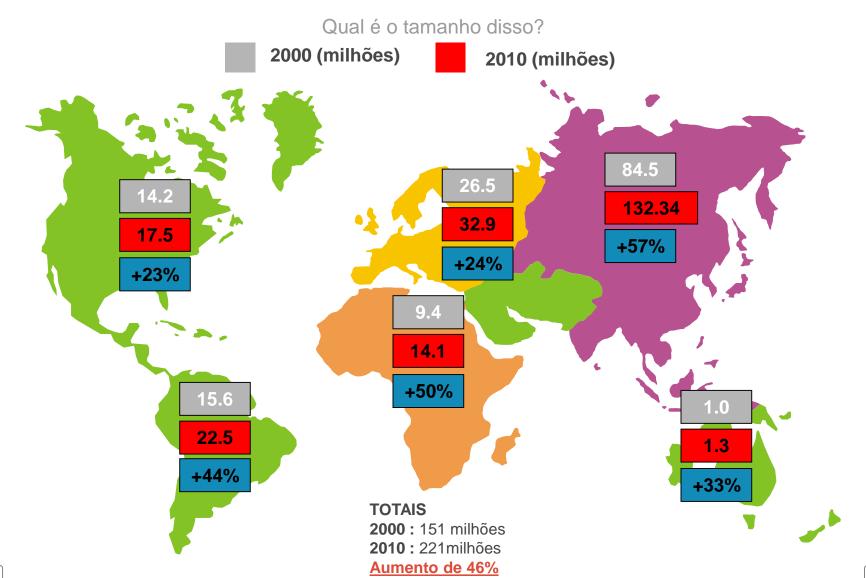


Sobre a Diabetes

Contextualização



Sobre a Diabetes



Sobre a Diabetes

A classificação

Valores de glicose plasmática (em mg/dl) para diagnóstico de diabetes *mellitus* e seus estágios pré-clínicos.

•			
Categoria	Jejum*	2 h após 75 g de glicose	Casual**
Glicemia normal	< 100	< 140	
Tolerância à glicose diminuída	≥ 100 a < 126	≥ 140 a < 200	
Diabetes mellitus	≥ 126	≥ 200	≥ 200 (com sintomas clássicos)***

^{*}O jejum é definido como a falta de ingestão calórica por no mínimo 8 h.

Nota: o diagnóstico do DM deve sempre ser confirmado pela repetição do teste em outro dia, a menos que haja hiperglicemia inequívoca com descompensação metabólica aguda ou sintomas óbvios de DM.

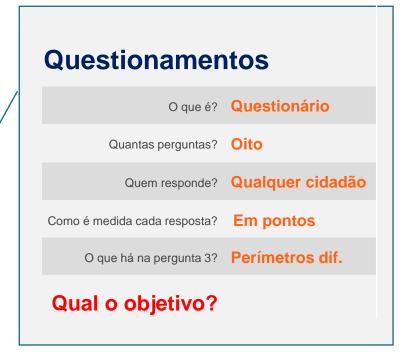
Tipo I	Tipo II
 Doença Auto-Imune Deficiência Insulina Jovens, Adolescentes Auto-Anticorpos Indivíduos Magros 	 90 % S. Metabólica Resistência Insulínica e Deficiência Indivíduos Meia Idade Depósito Amilóide nas Ilhotas Obesidade (85 %) Início Geralmente Lento

^{**}Glicemia plasmática casual é aquela realizada a qualquer hora do dia, sem se observar o intervalo desde a última refeição.

^{***}Os sintomas clássicos do DM incluem poliúria, polidipsia e perda não explicada de peso.

A metodologia adotada para coleta dos dados





detectar o risco de desenvolvimento da Diabetes do tipo II em um período de 10 anos

Características das perguntas [variáveis] [features]



- IDADE > Quantitativa Contínua
- IMC > Quantitativa Contínua
- CIRCUNFERÊNCIA DA CINTURA > Quantitativa Contínua
- 4 ATIVIDADE FÍSICA (> 30min) > Qualitativa Nominal
- 5 CONSUMO DE LEGUMES /FRUTAS > Qualitativa Nominal
- USO DE MEDICAMENTO (ANTI-HIPERTENÇÃO) > Qualitativa Nominal
- GLICEMIA ELEVADA (EXAME SAÚDE)
 > Qualitativa Nominal
- FATOR HEREDITARIEDADE (DIABETES I ou II)
 > Qualitativa Nominal

Discretização das variáveis [categorização] [feature engeneering]

- 1 IDADE > Quantitativa Contínua -> Intervalar
- IMC
 > Quantitativa Contínua -> Intervalar
- CIRCUNFERÊNCIA DA CINTURA
 > Quantitativa Contínua -> Intervalar
- 4 ATIVIDADE FÍSICA (> 30min) > Qualitativa Nominal
- CONSUMO DE LEGUMES /FRUTAS > Qualitativa Nominal
- USO DE MEDICAMENTO (ANTI-HIPERTENÇÃO)
 > Qualitativa Nominal
- 7 GLICEMIA ELEVADA (EXAME SAÚDE) > Qualitativa Nominal
- FATOR HEREDITARIEDADE (DIABETES I ou II)
 > Qualitativa Nominal

Idade, classificada em:

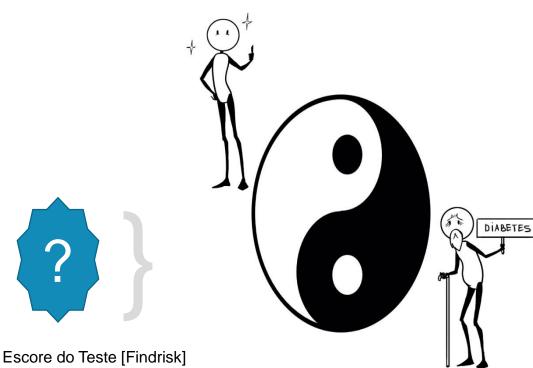
- menos de 45 anos (0 ponto);
- o entre 45 e 54 anos (2 pontos)
- o entre 55 e 64 an<mark>os (3 pontos)</mark>
- o mais de 64 anos (4 pontos).

Categorização em faixas (medida em pontos)

Preparação e entendimento das variáveis para a modelagem







- Risco levemente elevado (1 em 25) pontuação de 7 a 11 (inclusive);
- Risco moderado (1 em 6) pontuação de 12 a 14 (inclusive);
- Risco alto (1 em 3) pontuação de 15 a 20 (inclusive);

Risco baixo (1 em 100) – pontuação menor que 7;

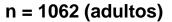
Risco muito alto (1 em 2) - pontuação maior que 20.

Coleta dos dados

Como foi a aplicação do findrisk







 $n\alpha = 43$ diabéticos tipo II



Tabulação dos dados

Alguns valores omissos (data missing) foram encontrados

Variável	Número de valores omissos
Idade	3
IMC	11
Circunferência	
da cintura	6
(mulheres)	
Circunferência	•
da cintura	3
(homens)	
Prática diária	15
de atividades físicas	15
Consumo de frutas	15
e legumes	15
Uso de medicamentos	15
anti-hipertensivos	15
Histórico de exame	17
com glicemia elevada	17
Histórico de familiar	
com diabetes	1 5
tipo I ou II	



n = 1062 (adultos) nα = 43 diabéticos tipo II

8 variáveis



<u>A</u> rqui	ivo <u>E</u> ditar	E <u>x</u> ibir <u>I</u> nse	erir <u>F</u> ormata	r Fe <u>r</u> ramen	itas <u>D</u> ados	Janela A
	· 🛅 • 1	O _# =0 [∢ ■		Ab X	o 6 ·
	Liberation	Sans 🔻	10 🔻	a a <u>a</u>	₹ ₹	= = [
H14		▼ f(x)	Σ =			
	Α	В	С	D	Е	F
1	Amostra	idade	imc	cintura_H	cintura_M	
2	n0001	0	0	0		
3	n0002	4	1	0		
4	n0003	3	DM	NA	3	
5	n0004	DM	1	DM	DM	
6	n0005	DM	1	1	NA	
7	n0006	2	2		4	
8	n0007	DM	0	DM	DM	
9	n0008	2	1	4	NA	
10	n0009	3	2	NA	0	
11	n0010	2	0	3	NA	
12	n0011	0	1	NA	0	
13						



ı

= 1019 amostras (observações)

Perfil dos Respondentes

Variável	Classe	Frequência	(%)
	Feminino	623	0,61
Gênero	Masculino	358	35%
	Não-informado	38	4%
	Menos de 45 anos	579	57%
Idade	Entre 45 - 54 anos	167	16%
ludue	Entre 55 - 64 anos	132	13%
	mais de 64 anos	139	14%
	Menor que 25 kg/m²	475	47%
IMC	Entre 25-30 kg/m ²	367	36%
	Superior a 30 kg/m ²	167	17%
Circunferência	Menos de 80 cm	233	38%
da cintura	Entre 80 - 88 cm	156	25%
(mulheres)	Mais de 88 cm	228	37%
Circunferência	Menos de 94 cm	147	41%
da cintura	Entre 94 - 102 cm	108	30%
(homens)	Mais de 102 cm	101	28%
Prática diária	Sim	467	46%
de atividades físicas	Não	545	54%
Consumo de frutas	Todo dia	640	63%
e legumes	Não todo dia	372	37%
Uso de medicamentos	Nunca fez uso	809	80%
anti-hipertensivos	Já fez uso	203	20%
Histórico de exame	Não	853	84%
com glicemia elevada	Sim	158	16%
Histórico de familiar	Não	473	47%
com diabetes	Sim, segundo grau	217	21%
tipo I ou II	Sim, primeiro grau	324	32%
	Baixo	389	38%
	Levemente elevado	339	33%
Findrisk	Moderado	162	16%
	Alto	120	12%
	Muito alto	9	1%



Risco por Gênero

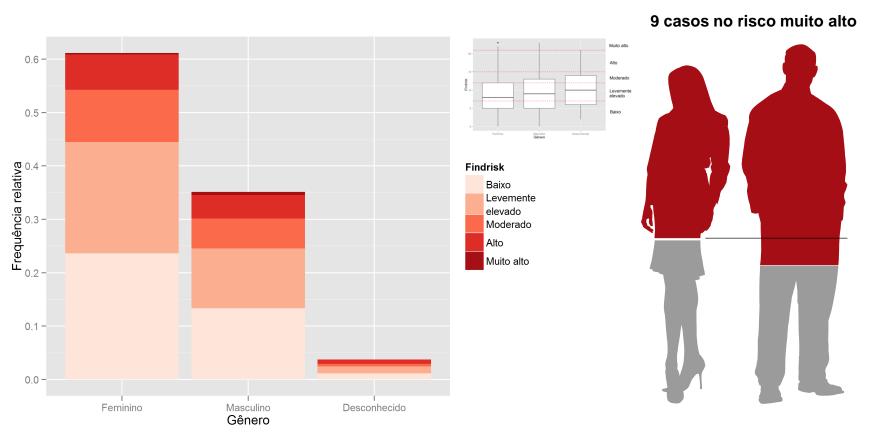


Gráfico B1 – Gráfico de barras da frequência relativa da variável Gênero particionado segundo as categorias do Findrisk.

Risco por Idade

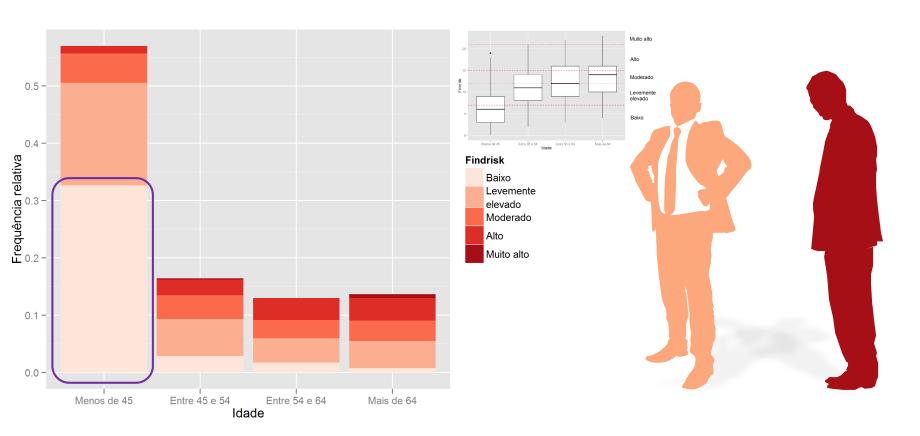


Gráfico B2 – Gráfico de barras da frequência relativa da variável Idade particionado segundo as categorias do Findrisk.



Risco por IMC

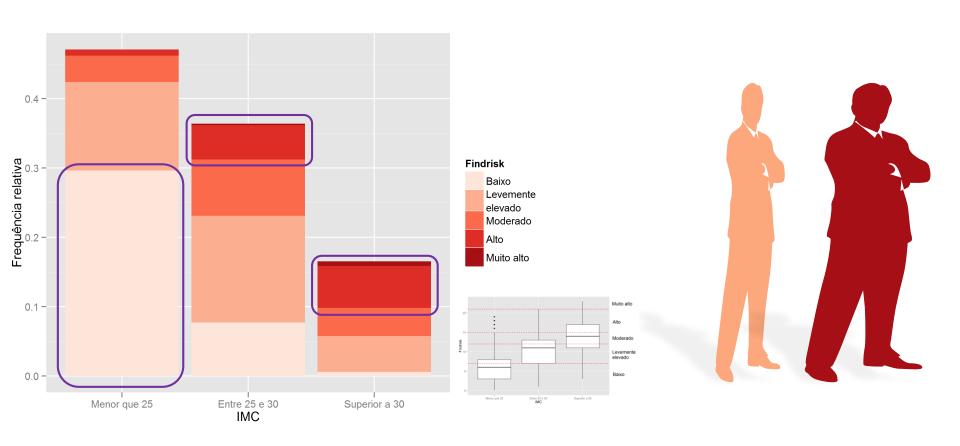


Gráfico B3 – Gráfico de barras da frequência relativa da variável IMC particionado segundo as categorias do Findrisk.



Risco por Circunferência da Cintura (M e H)

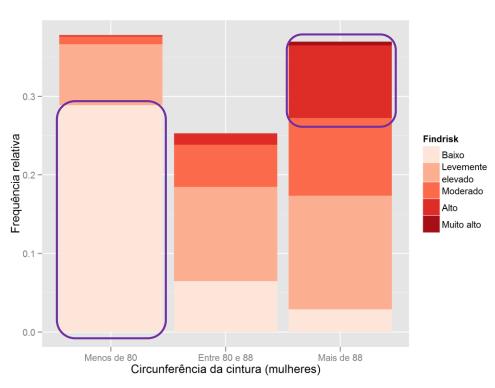


Gráfico B4 – Gráfico de barras da frequência relativa da variável Circunferência da cintura (mulheres) particionado segundo as categorias do Findrisk.

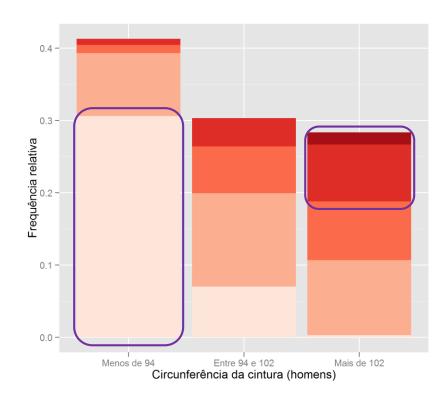


Gráfico B5 – Gráfico de barras da frequência relativa da variável Circunferência da cintura (homens) particionado segundo as categorias do Findrisk.





Risco por Circunferência da Cintura (M e H)

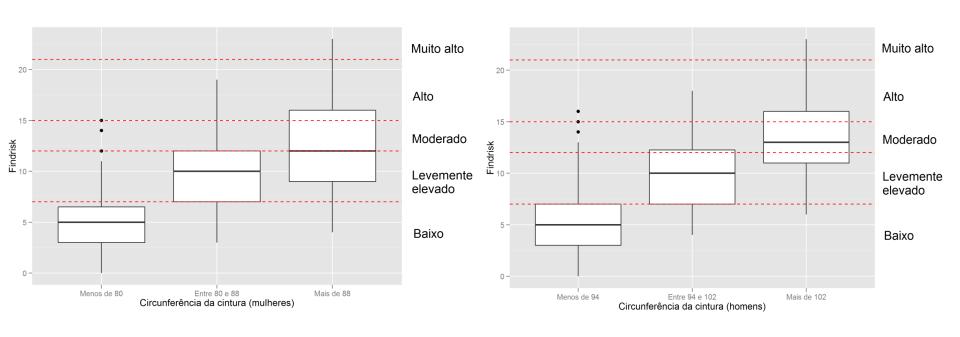


Gráfico B15 – Boxplot do escore indicado pelo Findrisk para cada categoria da variável Circunferência da cintura (mulheres).

Gráfico B16 – Boxplot do escore indicado pelo Findrisk para cada categoria da variável Circunferência da cintura (homens).





Risco por Atividade Física

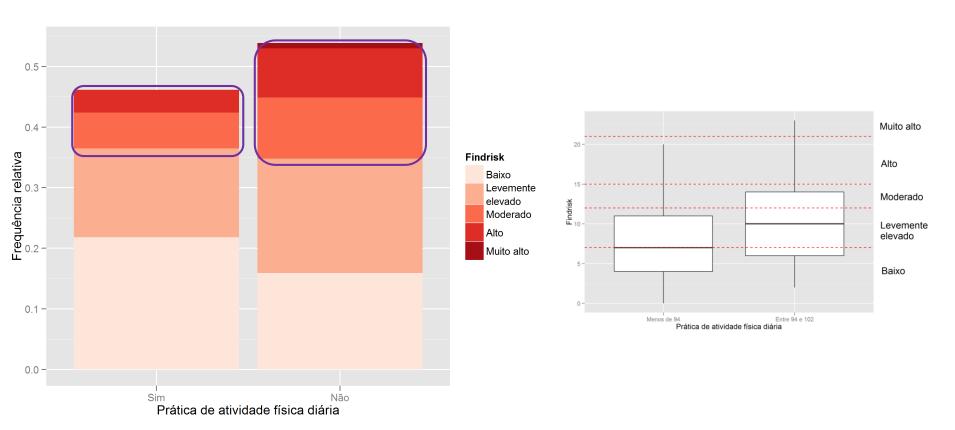


Gráfico B6 – Gráfico de barras da frequência relativa da variável Prática de atividade física diária particionado segundo as categorias do Findrisk.



Risco por Consumo de Frutas e Legumes

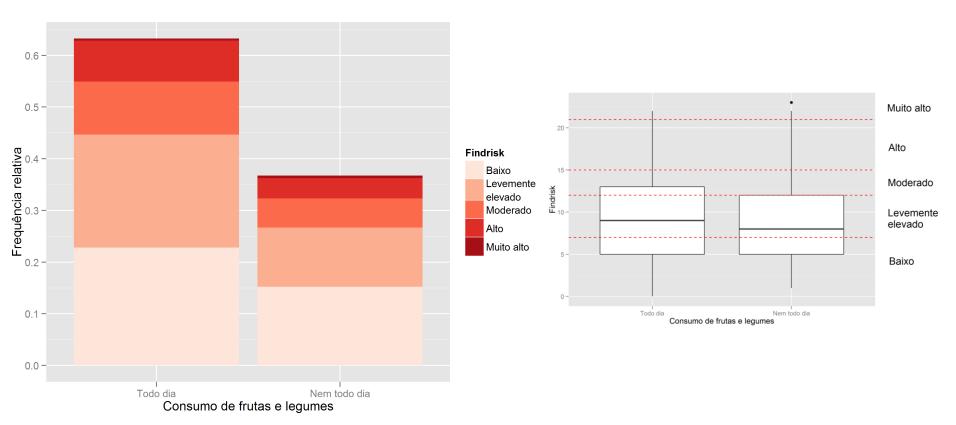


Gráfico B7 – Gráfico de barras da frequência relativa da variável Consumo de frutas e legumes particionado segundo as categorias do Findrisk.



Risco por Uso de Medicamento Anti-hipertensivo

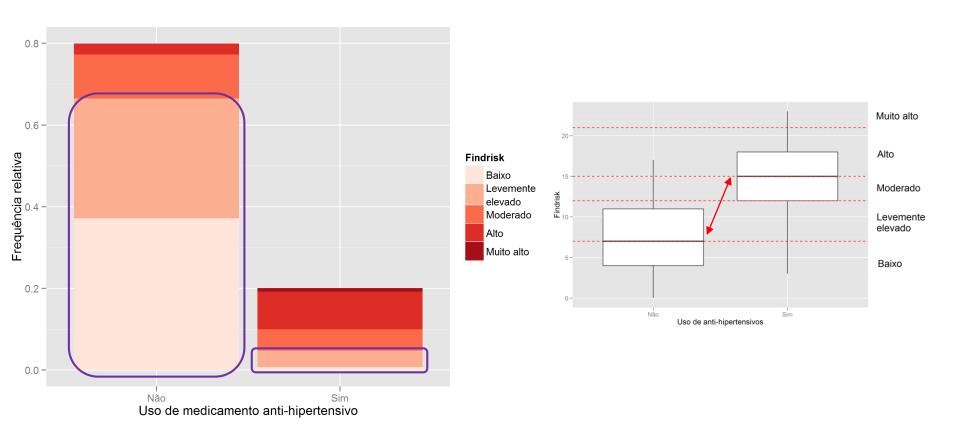


Gráfico B8 – Gráfico de barras da frequência relativa da variável Uso de medicamento anti-hipertensivo particionado segundo as categorias do Findrisk.



Risco por Histórico de Exame com GLICEMIA elevada

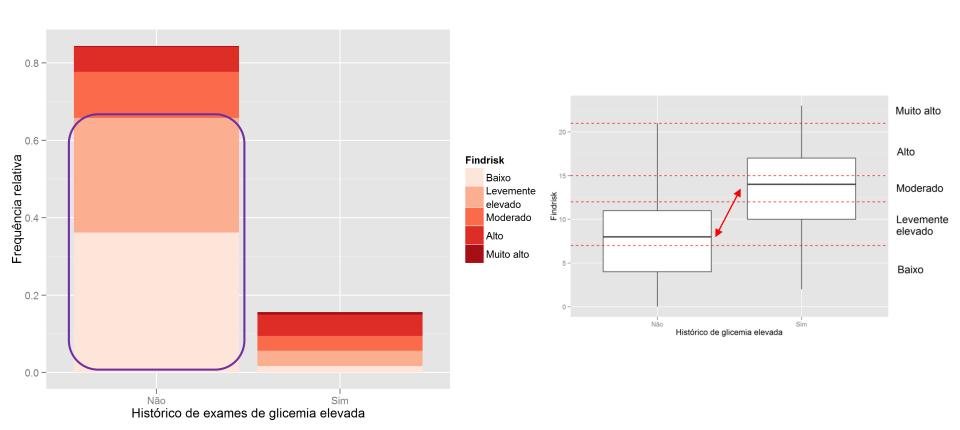


Gráfico B9 – Gráfico de barras da frequência relativa da variável Histórico de exames de glicemia elevada particionado segundo as categorias do Findrisk.



Risco por Histórico de Familiar com Diabetes (hereditariedade)

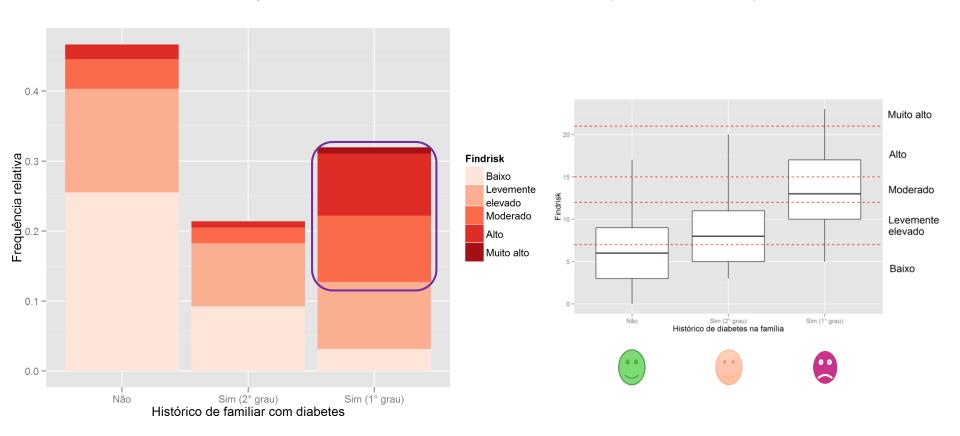


Gráfico B10 – Gráfico de barras da frequência relativa da variável Histórico de familiar com diabetes particionado segundo as categorias do Findrisk.



Relação entre Variáveis

Associação ou Não Independência entre as Variáveis

- IDADE
- IMC
- CIRCUNFERÊNCIA DA CINTURA
- ATIVIDADE FÍSICA (> 30min)
- CONSUMO DE LEGUMES /FRUTAS
- USO DE MEDICAMENTO (ANTI-HIPERTENÇÃO)
- GLICEMIA ELEVADA (EXAME SAÚDE)
- FATOR HEREDITARIEDADE (DIABETES I ou II)

Teste X²

H0: Não existe evidência de Associação entre as Variáveis

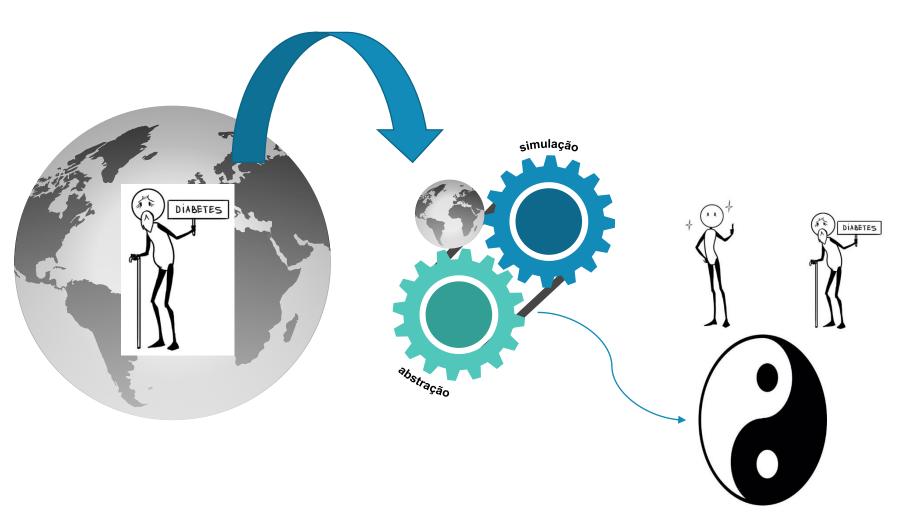
H1: Existe	evidencia d	de Associa	çao entre	e as var	ıaveıs (d	OK)

	1	2	3	4	5	6	7	8
1	-	-	-	-	-	-	-	-
2		-	-	-	-	-	-	-
3			-	-	-	-	-	-
4	ok			-	-	-	-	-
5	ok				-	-	-	-
6	ok	ok	ok*		ok	-	-	-
7							-	-
8								-

$$X^{2} = \sum_{i=1}^{n} \frac{\left(o_{i} - e_{i}\right)^{2}}{e_{i}}$$



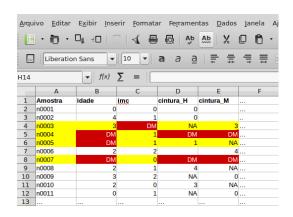
Conceituação de Modelo



Modelo de Regressão Logística

Variáveis Regressoras (explicativas)

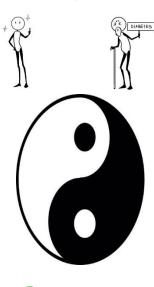
- Quantitativas
- Qualitativas
- Escore!





Variável Resposta

- · Qualitativas dicotômicas
- Distribuição Binomial



0 1

risco de desenvolvimento da Diabetes do tipo II em um período de 10 anos

Criação do Modelo

Regressão Logística



Modelo de Regressão Logística

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

0 (risco de desenvolvimento da Diabetes do tipo II em um período de 10 anos)

x é uma característica (imc, idade, histórico de glicemia....)

β são os parâmetros do modelo (pesos que afetam cada uma das variáveis)

Modelo de Regressão Logística

Resolvendo Problemas

- Na base de dados existia apenas 39 casos de diabetes tipo II (resposta=1) e
 939 casos sem diagnóstico da doença (resposta=0).
- Se toda a amostra fosse considerada não seria possível a estimação dos parâmetros do modelo.
- Com isso foram amostrados 200 controles dentro da própria amostra (PAULA, 2012).

Modelos Ajustados

- Modelo I: considera as oito perguntas do questionário e o gênero como variáveis explicativas; e
- Modelo II: considera o escore final do questionário como variável explicativa.

Modelo de Regressão Logística

Avaliação do modelo 1

- Apresentou um desvio de 91,539, que corresponde a um valor-p de aproximadamente 1 e é um indício de que o modelo está bem ajustado.
- Após <u>retirada de algumas observações</u>, os parâmetros estimados β0, β6, β7, β8 foram significativos.

Tabela 3. Estimativa, erro-padrão e valor-p dos coeficientes do Modelo I.

Coeficiente	Estimativa	Erro-padrão	Valor-p
Intercepto	-5,7884	0,8230	< 0, 0001
Anti-hipertensivos	1,2304	0,5483	0,0248
Hist. glicemia	3,9443	0,6142	0,0002
Hist. familiar (1° grau)	2,2888	0,6763	< 0,0001

 indivíduos que utilizam medicamentos antihipertensivos tem uma chance cerca de 3,4 vezes maior de serem diagnosticados com Diabetes Tipo II.

Tabela 3. Razão de chances para as variáveis do Modelo I.

Variável	Razão	Valor	
Anti-hipertensivos	Sim/Não	3,4225	
Hist. glicemia	Sim/Não	51,6383	
Hist. familiar (1° grau)	Sim/Não	9,8631	

Variável resposta tem distribuição binomial?

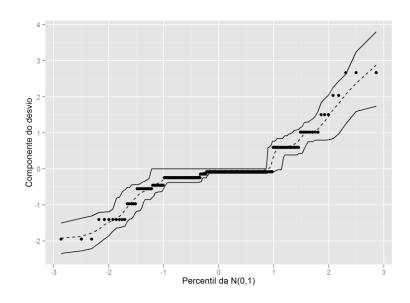


Gráfico C1 - Gráfico envelope para o Modelo 1.

Modelo de Regressão Logística

Avaliação do modelo 1

Após retirada de algumas obs?

Tabela 2. Valores-p para os testes de razão de verossimilhanças para cada uma das variáveis explicativas.

Variáveis	р
Gênero	85,83%
Idade	21,81%
IMC	30,18%
Circunferência da cintura	31,86%
Prática de atividades físicas	28,71%
Consumo de frutas e legumes	36,00%
Uso de medicamentos anti-hipertensivos	2,71%
Histórico de glicemia elevada	<0,01%
Histórico familiar de diabetes	0,34%

Modelo de Regressão Logística

Avaliação do modelo 2

- A mesma técnica apresentada no slide anterior.
- Além disso foi avaliado especificidade (falsos positivos) e sensibilidade (verdadeiros positivos).
- Sensibilidade e Especificidade servem para verificar a performance da classificação.
- Medida KS = 64.6 %, Kolmogorov-Smirnof (Conover, 1971). Essa medida analisa a probabilidade predita pelo modelo que maximiza a sensibilidade e a especificidade.

Tabela A18 – Tabela com as medidas de predição para o Modelo II

Escore	Probabilidade de	Sensibilida	Especificida			
Total	Corte	de	de	KS	VPP	VPN
0	0,8%	100,0%	0,0%	0,0%	100,0%	-
1	1,0%	2,5%	100,0%	2,5%	100,0%	16,7%
2	1,1%	5,0%	100,0%	5,0%	100,0%	17,0%
3	1,4%	9,5%	100,0%	9,5%	100,0%	17,7%
4	1,6%	16,0%	100,0%	16,0%	100,0%	18,8%
5	2,0%	21,5%	97,4%	18,9%	97,7%	19,5%
6	2,4%	29,0%	97,4%	26,4%	98,3%	21,1%
7	2,8%	37,0%	97,4%	34,4%	98,7%	23,2%
8	3,4%	45,5%	94,9%	40,4%	97,8%	25,3%
9	4,0%	52,0%	94,9%	46,9%	98,1%	27,8%
10	4,8%	59,5%	92,3%	51,8%	97,5%	30,8%
11	5,7%	63,5%	89,7%	53,2%	96,9%	32,4%
12	6,8%	70,5%	89,7%	60,2%	97,2%	37,2%
13	8,1%	77,0%	82,1%	59,1%	95,7%	41,0%
14	9,6%	82,5%	82,1%	64,6%	95,9%	47,8%
15	11,3%	88,5%	66,7%	55,2%	93,2%	53,1%
16	13,3%	91,0%	61,5%	52,5%	92,4%	57,1%
17	15,5%	93,0%	56,4%	49,4%	91,6%	61,1%
18	18,1%	96,0%	48,7%	44,7%	90,6%	70,4%
19	21,0%	98,0%	43,6%	41,6%	89,9%	81,0%
20	24,2%	98,5%	33,3%	31,8%	88,3%	81,3%
21	27,8%	99,0%	20,5%	19,5%	86,5%	80,0%
22	31,6%	100,0%	7,7%	7,7%	84,7%	100,0%
23	35,7%	100,0%	2,6%	2,6%	84,0%	100,0%

Considerações Finais

Resumo do que realmente se destacou

Análise Descritiva

Na análise dos gráficos (barras e boxplot) percebeu-se que a variáveis gênero e consumo de frutas / legumes foram as únicas que aparentaram não ter diferenças entre as suas categorias.

Associação de Variáveis

 Pelo teste X² as variáveis idade e uso de medicamentos anti-hipertensivos foram as duas principais associações com as demais variáveis.

Análise Inferencial

- Pelo ajuste do Modelo I, as variáveis que melhor explicaram a probabilidade de se diagnosticar um indivíduo com Diabetes do tipo II são: histórico de glicemia elevada, histórico familiar de diabetes e uso de medicamentos anti-hipertensivos.
- Já pelo Modelo II, conclui-se que o Findrisk apresentou um bom desempenho na identificação de indivíduos já diagnosticados com Diabetes Milito do tipo II.

