

# Apresentação Final

## Diabetes – PIMA

Aprendizagem Estatística em Altas Dimensões  
(MAE5904-MAE0501-IBI5904)

Ícaro Maia Santos de Castro  
Rayssa de Carvalho Roberto  
Rodrigo Aoyama Nakahara  
Rodrigo Marcel Araujo Oliveira  
Vitor Hugo Vieira de Lima

IME-USP

25 de Novembro de 2020

# Agenda

- 1 Introdução
  - Descrição dos Dados
  - Partição dos Dados
- 2 Análise Descritiva
  - Dados de Treino
- 3 Análise de *Missings* e Imputações
  - *Missings* na Partição 1
  - Estratégia de Imputação
- 4 Modelagem
  - Análise Discriminante
  - Regressão Logística
  - *Random Forest*
  - *Support Vector Machine*
- 5 Comparação dos Modelos
- 6 Outras Implementações
- 7 Comentários Finais
- 8 Referências

# Objetivo e Desafios

## Objetivo

Prever se a pessoa possui ou não diabetes *mellitus*, com base em uma série de variáveis preditoras.

## Desafio

Classificação em banco de dados com muitos dados faltantes.

# Banco de Dados PIMA: Variáveis

## Variáveis

Variável	Explicação
Diabetes	Variável resposta categórica (1 se diabético, 0 se não diabético)
Pregnancies	Quantidade de gestações
Glucose	Concentração de glicose no plasma após 2 horas em um teste oral de tolerância a glicose
BloodPressure	Pressão arterial diastólica (mm Hg)
SkinThickness	Espessura da dobra da pele do tríceps (mm)
Insulin	Insulina sérica de 2-horas ( $\mu$ U/ml)
BMI	Índice de massa corporal (peso em kg/(altura em m) <sup>2</sup> )
DiabetesPedigreeFunction	Função “pedigree” de diabetes
Age	Idade (anos)

# Conjuntos de Treinamento e de Teste

## Partição 1: Imputação dos Dados

Total de 768 Observações (100%)

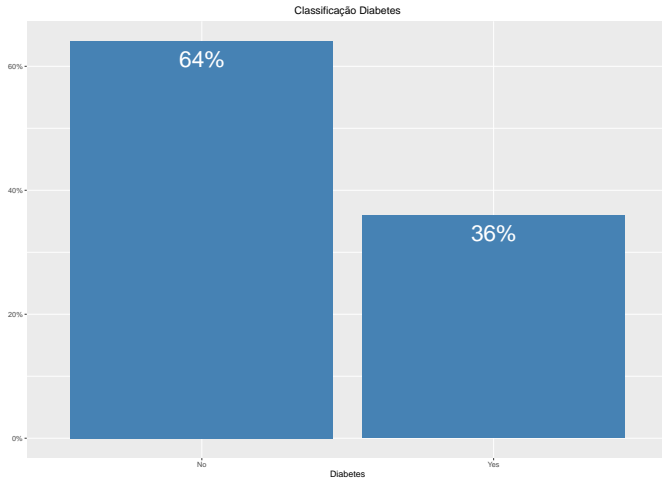
- Treinamento: 537 (70%)
- Teste: 231 (30%)
  - Teste: 161 (70% do Teste)
  - *Out-of-sample*: 70 (30% do Teste)

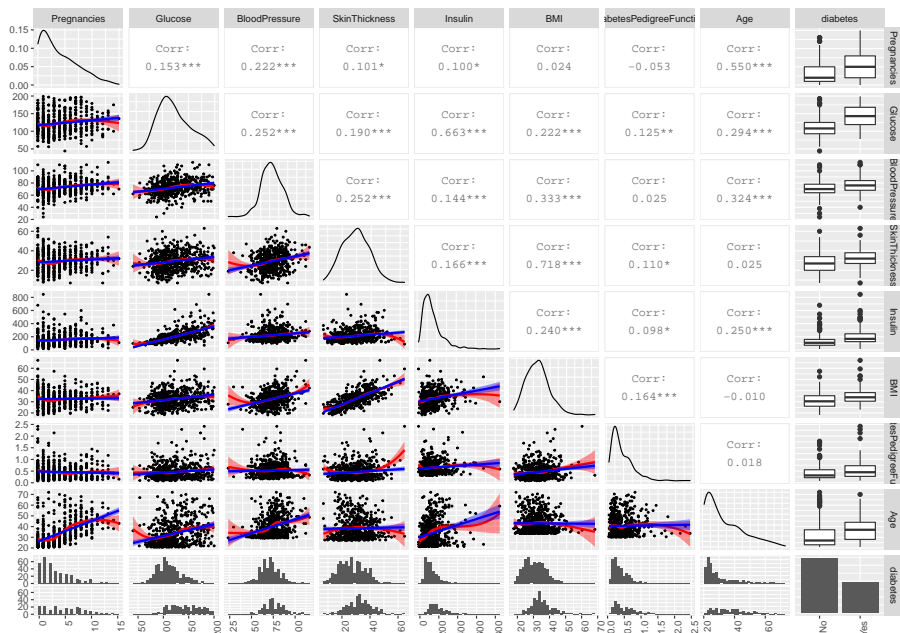
## Partição 2: Eliminação de *Missings*

Total de 382 Observações (100%)

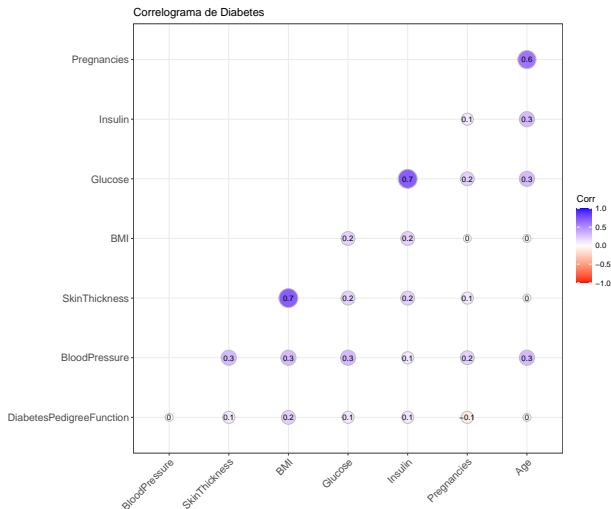
- Treinamento: 266 (70%)
- Teste: 116 (30%)
  - Teste: 89 (77% do Teste)
  - *Out-of-sample*: 27 (23% do Teste)

# Nos Dados de Treinamento (Sem *Data Snooping*)



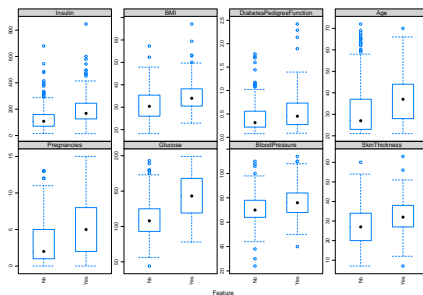
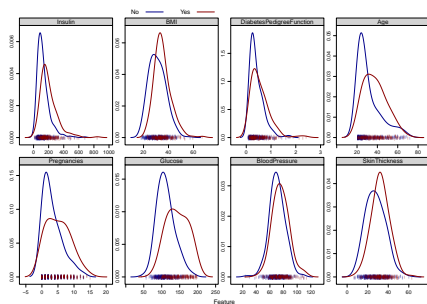


# Correlograma

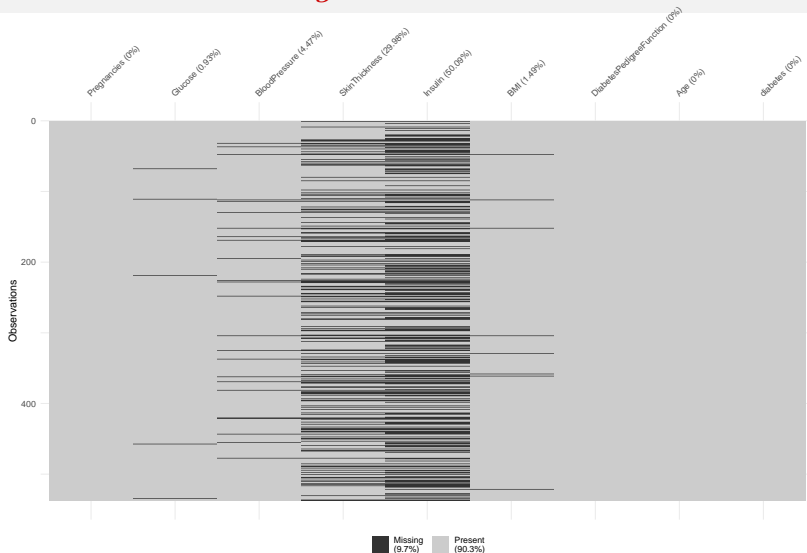




# Distribuição por Categoria de Diabetes

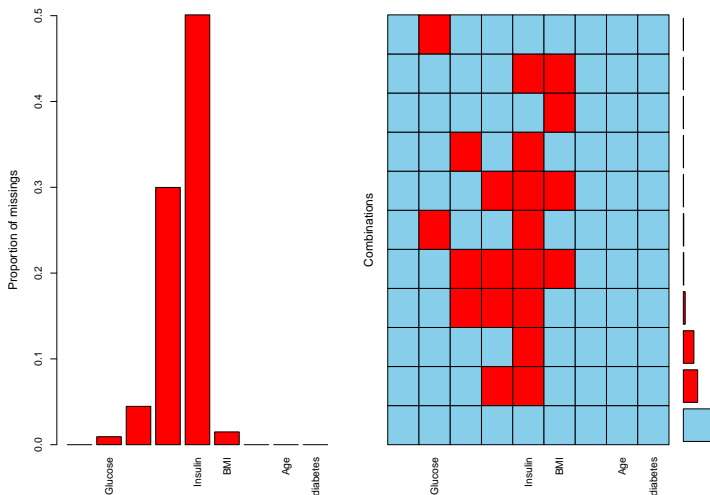


# Padrões nos *Missings*?



# Padrões nos *Missings*?

Figura: Frequência de *Missings* e Plot de Combinações



# Variáveis com Mais *Missings*



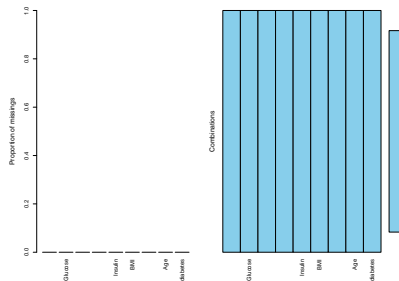
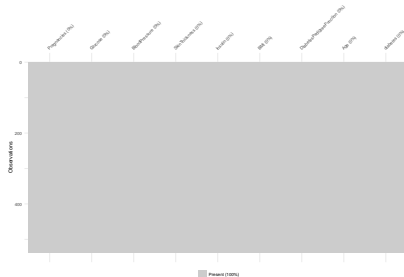
# Estratégia de Imputação (Partição 1)

## Pacote mice

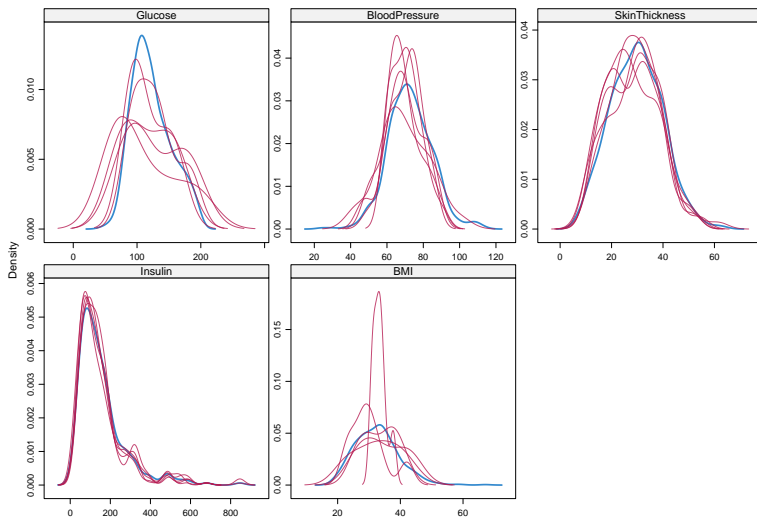
*Multivariate Imputation by Chained Equations* (MICE)

Método **pmm** (*predictive mean matching*)

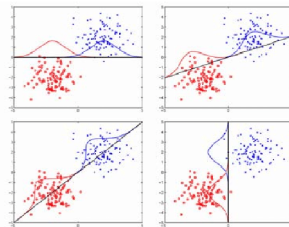
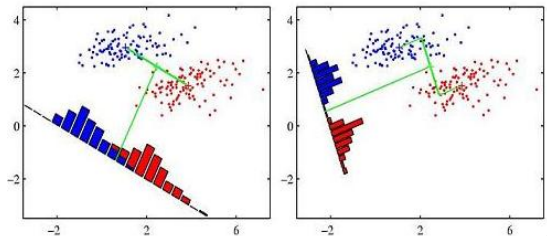
<https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html>



# Estratégia de Imputação (Partição 1)



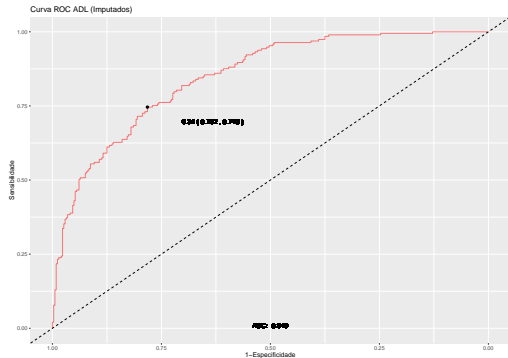
## Análise Discriminante



# Análise Discriminante Linear (Partição 1)

Predito	Referência	
	No	Yes
No	99	28
Yes	9	25

- Validação cruzada ( $k = 15$ )
- Dados imputados





---

Desempenho do Modelo

---

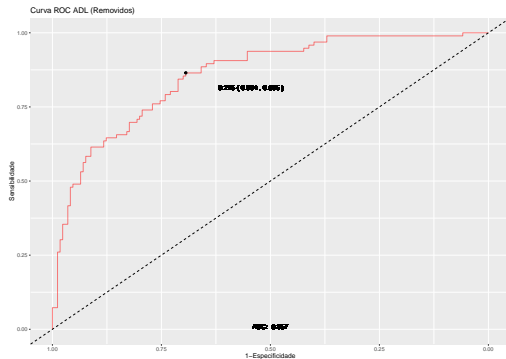
<b>Accuracy</b>	0.7702	<b>Sensitivity</b>	0.9167
<b>95% CI</b>	(0.6974, 0.8327)	<b>Specificity</b>	0.4717
<b>No Information Rate</b>	0.6708	<b>Pos Pred Value</b>	0.7795
<b>P-Value [Acc &gt;NIR]</b>	0.003815	<b>Neg Pred Value</b>	0.7353
		<b>Prevalence</b>	0.6708
<b>Kappa</b>	0.4274	<b>Detection Rate</b>	0.6149
		<b>Detection Prevalence</b>	0.7888
<b>Mcnemar's Test P-Value</b>	0.003085	<b>Balanced Accuracy</b>	0.6942

---

# Análise Discriminante Linear (Partição 2)

Predito	Referência	
	No	Yes
No	64	12
Yes	4	9

- Validação cruzada ( $k = 15$ )
- Dados sem *missings*



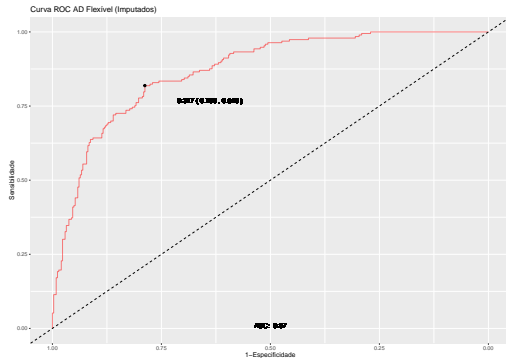
## Desempenho do Modelo

<b>Accuracy</b>	0.8202	<b>Sensitivity</b>	0.9412
<b>95% CI</b>	(0.7245, 0.8936)	<b>Specificity</b>	0.4286
<b>No Information Rate</b>	0.764	<b>Pos Pred Value</b>	0.8421
<b>P-Value [Acc &gt;NIR]</b>	0.12910	<b>Neg Pred Value</b>	0.6923
		<b>Prevalence</b>	0.7640
<b>Kappa</b>	0.4258	<b>Detection Rate</b>	0.7191
		<b>Detection Prevalence</b>	0.8539
<b>Mcnemar's Test P-Value</b>	0.08012	<b>Balanced Accuracy</b>	0.6849

# Análise Discriminante Flexível (Partição 1)

Predito	Referência	
	No	Yes
No	95	25
Yes	13	28

- Validação cruzada ( $k = 15$ )
- Dados imputados



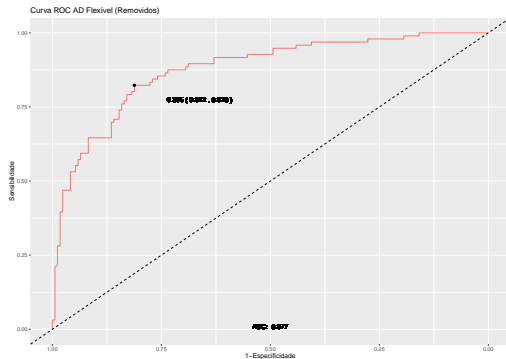
## Desempenho do Modelo

<b>Accuracy</b>	0.764	<b>Sensitivity</b>	0.8796
<b>95% CI</b>	(0.6907, 0.8272)	<b>Specificity</b>	0.5283
<b>No Information Rate</b>	0.6708	<b>Pos Pred Value</b>	0.7917
<b>P-Value [Acc &gt;NIR]</b>	0.006397	<b>Neg Pred Value</b>	0.6829
		<b>Prevalence</b>	0.6708
<b>Kappa</b>	0.4329	<b>Detection Rate</b>	0.5901
		<b>Detection Prevalence</b>	0.7453
<b>Mcnemar's Test P-Value</b>	0.074353	<b>Balanced Accuracy</b>	0.7040

# Análise Discriminante Flexível (Partição 2)

Predito	Referência	
	No	Yes
No	61	11
Yes	7	10

- Validação cruzada ( $k = 15$ )
- Dados sem *missings*



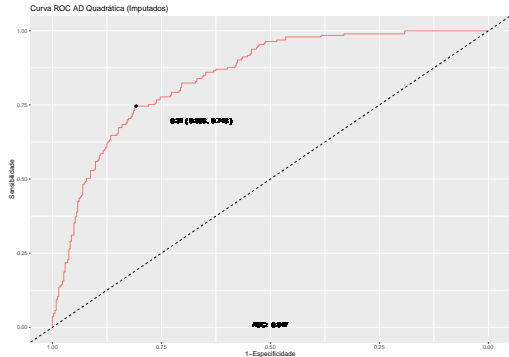
## Desempenho do Modelo

<b>Accuracy</b>	0.7978	<b>Sensitivity</b>	0.8971
<b>95% CI</b>	(0.6993, 0.8755)	<b>Specificity</b>	0.4762
<b>No Information Rate</b>	0.764	<b>Pos Pred Value</b>	0.8472
<b>P-Value [Acc &gt;NIR]</b>	0.2709	<b>Neg Pred Value</b>	0.5882
		<b>Prevalence</b>	0.7640
<b>Kappa</b>	0.3996	<b>Detection Rate</b>	0.6854
		<b>Detection Prevalence</b>	0.8090
<b>Mcnemar's Test P-Value</b>	0.4795	<b>Balanced Accuracy</b>	0.6866

# Análise Discriminante Quadrática (Partição 1)

Predito	Referência	
	No	Yes
No	96	27
Yes	12	26

- Validação cruzada ( $k = 15$ )
- Dados imputados





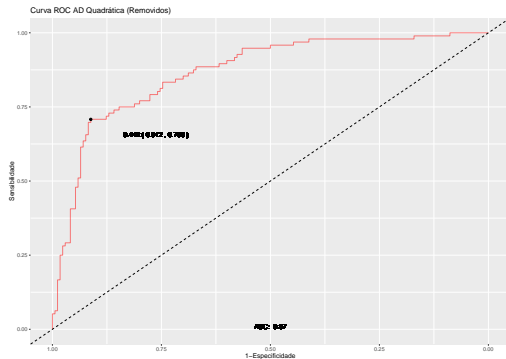
## Desempenho do Modelo

<b>Accuracy</b>	0.7578	<b>Sensitivity</b>	0.8889
<b>95% CI</b>	(0.6841, 0.8217)	<b>Specificity</b>	0.4906
<b>No Information Rate</b>	0.6708	<b>Pos Pred Value</b>	0.7805
<b>P-Value [Acc &gt;NIR]</b>	0.01039	<b>Neg Pred Value</b>	0.6842
		<b>Prevalence</b>	0.6708
<b>Kappa</b>	0.4089	<b>Detection Rate</b>	0.5963
		<b>Detection Prevalence</b>	0.7640
<b>Mcnemar's Test P-Value</b>	0.02497	<b>Balanced Accuracy</b>	0.6897

# Análise Discriminante Quadrática (Partição 2)

Predito	Referência	
	No	Yes
No	60	6
Yes	8	15

- Validação cruzada ( $k = 15$ )
- Dados sem *missings*



---

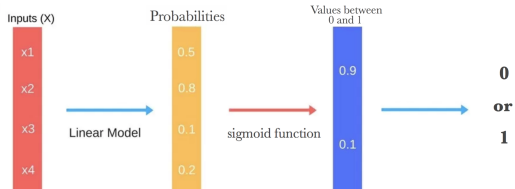
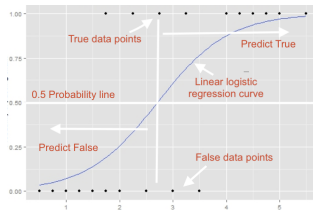
Desempenho do Modelo

---

<b>Accuracy</b>	0.8427	<b>Sensitivity</b>	0.8824
<b>95% CI</b>	(0.7502, 0.9112)	<b>Specificity</b>	0.7143
<b>No Information Rate</b>	0.764	<b>Pos Pred Value</b>	0.9091
<b>P-Value [Acc &gt;NIR]</b>	0.04778	<b>Neg Pred Value</b>	0.6522
		<b>Prevalence</b>	0.7640
<b>Kappa</b>	0.5776	<b>Detection Rate</b>	0.6742
		<b>Detection Prevalence</b>	0.7416
<b>Mcnemar's Test P-Value</b>	0.78927	<b>Balanced Accuracy</b>	0.7983

---

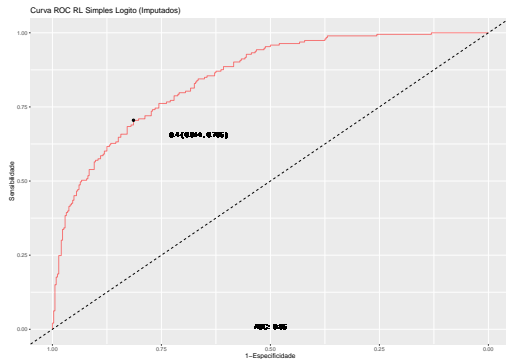
# Regressão Logística



# Regressão Logística Simples (Partição 1)

Predito	Referência	
	No	Yes
No	99	27
Yes	9	26

- Validação cruzada ( $k = 10$ )
- Dados imputados



---

Desempenho do Modelo

---

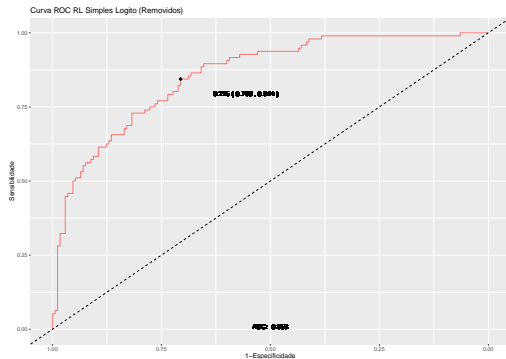
<b>Accuracy</b>	0.7764	<b>Sensitivity</b>	0.9167
<b>95% CI</b>	(0.7041, 0.8382)	<b>Specificity</b>	0.4906
<b>No Information Rate</b>	0.6708	<b>Pos Pred Value</b>	0.7857
<b>P-Value [Acc &gt;NIR]</b>	0.002203	<b>Neg Pred Value</b>	0.7429
		<b>Prevalence</b>	0.6708
<b>Kappa</b>	0.4458	<b>Detection Rate</b>	0.6149
		<b>Detection Prevalence</b>	0.7826
<b>Mcnemar's Test P-Value</b>	0.004607	<b>Balanced Accuracy</b>	0.7036

---

# Regressão Logística Simples (Partição 2)

Predito	Referência	
	No	Yes
No	64	11
Yes	4	10

- Validação cruzada ( $k = 10$ )
- Dados sem *missings*



---

Desempenho do Modelo

---

Accuracy	0.8315	Sensitivity	0.9412
95% CI	(0.7373, 0.9025)	Specificity	0.4762
No Information Rate	0.764	Pos Pred Value	0.8533
P-Value [Acc >NIR]	0.08127	Neg Pred Value	0.7143
		Prevalence	0.7640
Kappa	0.4717	Detection Rate	0.7191
		Detection Prevalence	0.8427
Mcnemar's Test P-Value	0.12134	Balanced Accuracy	0.7087

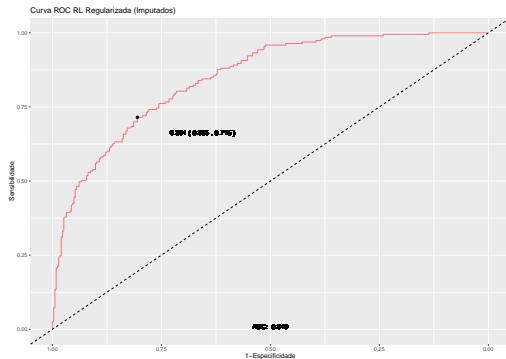
---



# Regressão Logística Regularizada (Partição 1)

Predito	Referência	
	No	Yes
No	99	28
Yes	9	25

- Validação cruzada ( $k = 10$ )
- Dados imputados



---

Desempenho do Modelo

---

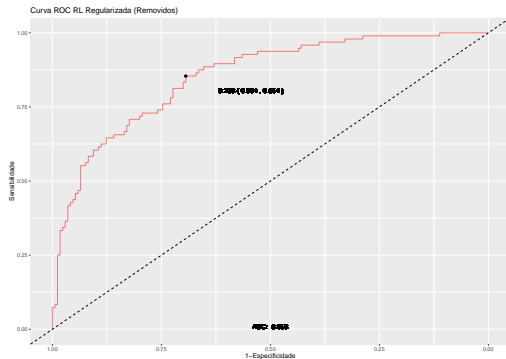
Accuracy	0.7702	Sensitivity	0.9167
95% CI	(0.6974, 0.8327)	Specificity	0.4717
No Information Rate	0.6708	Pos Pred Value	0.7795
P-Value [Acc >NIR]	0.003815	Neg Pred Value	0.7353
		Prevalence	0.6708
Kappa	0.4274	Detection Rate	0.6149
		Detection Prevalence	0.7888
Mcnemar's Test P-Value	0.003085	Balanced Accuracy	0.6942

---

# Regressão Logística Regularizada (Partição 2)

Predito	Referência	
	No	Yes
No	64	11
Yes	4	10

- Validação cruzada ( $k = 10$ )
- Dados sem *missings*



---

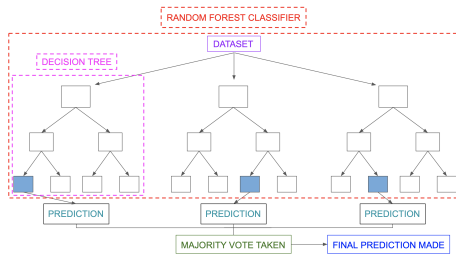
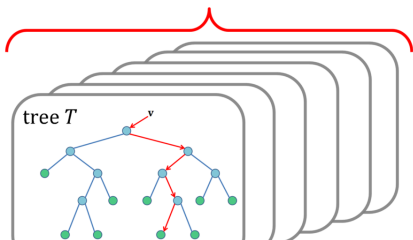
Desempenho do Modelo

---

Accuracy	0.8315	Sensitivity	0.9412
95% CI	(0.7373, 0.9025)	Specificity	0.4762
No Information Rate	0.764	Pos Pred Value	0.8533
P-Value [Acc >NIR]	0.08127	Neg Pred Value	0.7143
		Prevalence	0.7640
Kappa	0.4717	Detection Rate	0.7191
		Detection Prevalence	0.8427
Mcnemar's Test P-Value	0.12134	Balanced Accuracy	0.7087

---

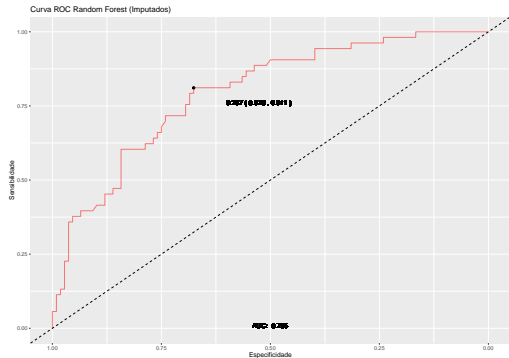
# Random Forest



# Random Forest (Partição 1)

Predito	Referência	
	No	Yes
No	93	28
Yes	15	25

- Validação cruzada ( $k = 10$ )
- Dados imputados



---

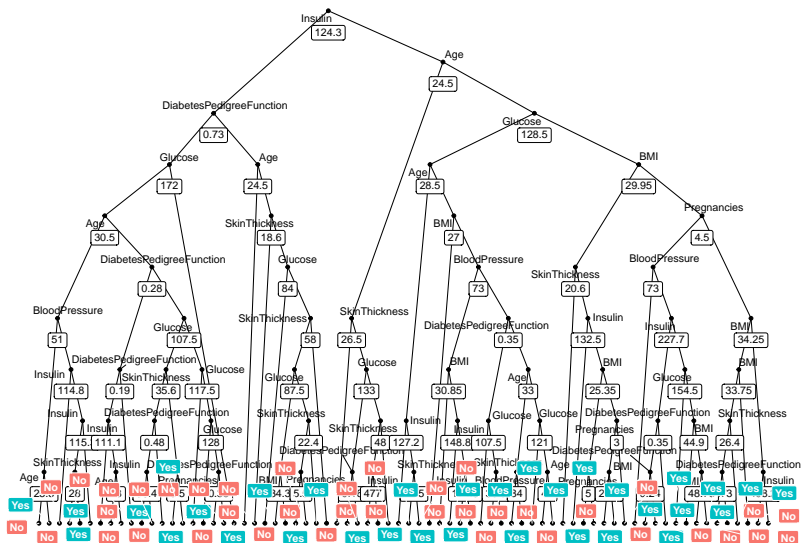
**Desempenho do Modelo**


---

Accuracy	0.7329	Sensitivity	0.8611
95% CI	(0.6576, 0.7995)	Specificity	0.4717
No Information Rate	0.6708	Pos Pred Value	0.7686
P-Value [Acc >NIR]	0.05368	Neg Pred Value	0.6250
		Prevalence	0.6708
Kappa	0.355	Detection Rate	0.5776
		Detection Prevalence	0.7516
Mcnemar's Test P-Value	0.06725	Balanced Accuracy	0.6664

---

# Partição 1: Menor árvore com imputação 250

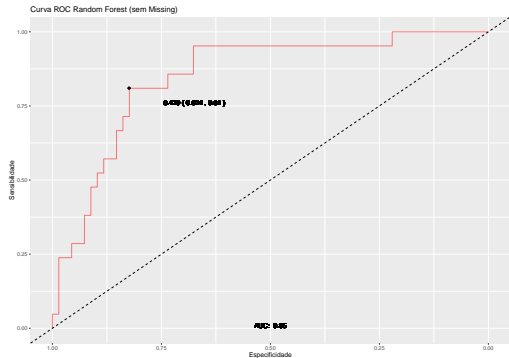




# Random Forest (Partição 2)

Predito	Referência	
	No	Yes
No	59	9
Yes	9	12

- Validação cruzada ( $k = 10$ )
- Dados sem *missings*



---

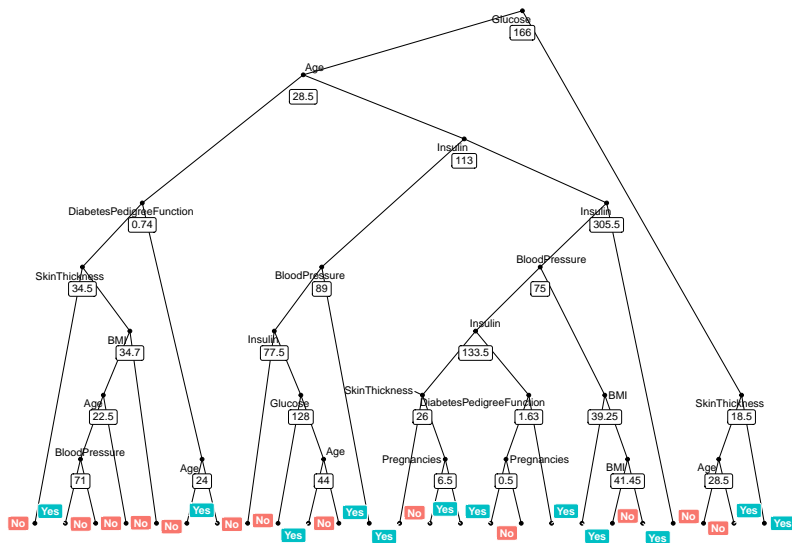
**Desempenho do Modelo**


---

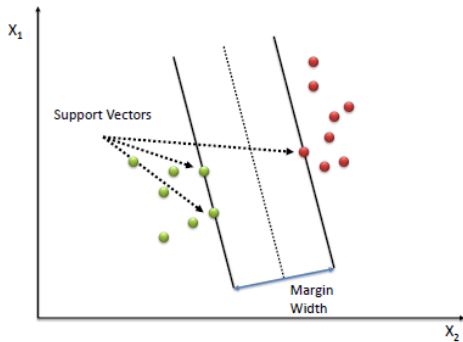
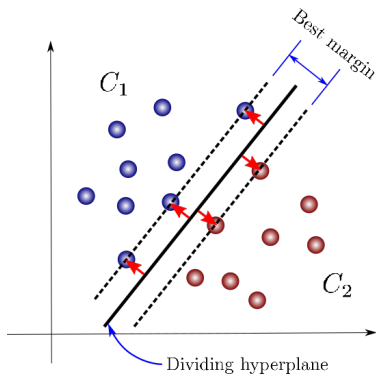
Accuracy	0.7978	Sensitivity	0.8676
95% CI	(0.6993, 0.8755)	Specificity	0.5714
No Information Rate	0.764	Pos Pred Value	0.8676
P-Value [Acc >NIR]	0.2709	Neg Pred Value	0.5714
		Prevalence	0.7640
Kappa	0.4391	Detection Rate	0.6629
		Detection Prevalence	0.7640
Mcnemar's Test P-Value	1.0000	Balanced Accuracy	0.7195

---

## Partição 2: Menor árvore sem imputação: 250



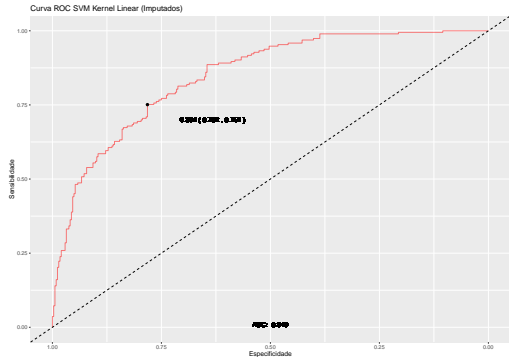
# Support Vector Machine



# SVM Kernel Linear (Partição 1)

Predito	Referência	
	No	Yes
No	98	27
Yes	10	26

- Validação cruzada ( $k = 10$ )
- Dados imputados



---

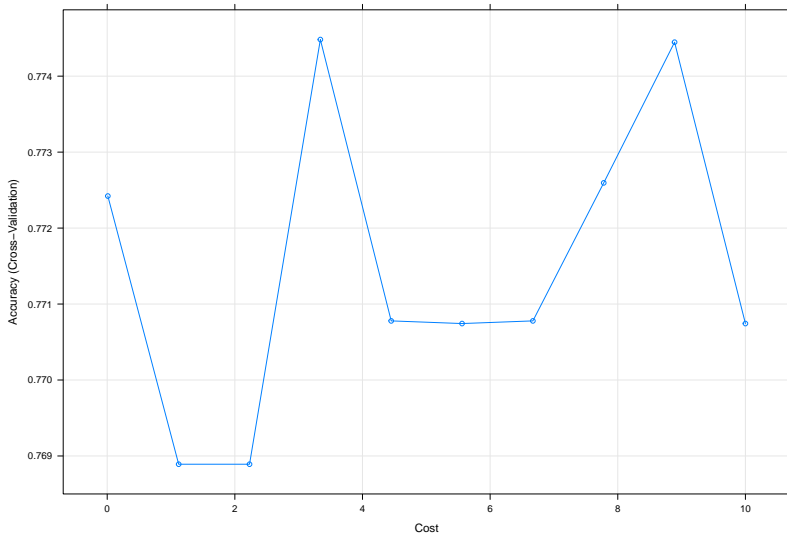
### Desempenho do Modelo

---

Accuracy	0.7702	Sensitivity	0.9074
95% CI	(0.6974, 0.8327)	Specificity	0.4906
No Information Rate	0.6708	Pos Pred Value	0.7840
P-Value [Acc >NIR]	0.003815	Neg Pred Value	0.7222
		Prevalence	0.6708
Kappa	0.4334	Detection Rate	0.6087
		Detection Prevalence	0.7764
McNemar's Test P-Value	0.008529	Balanced Accuracy	0.6990

---

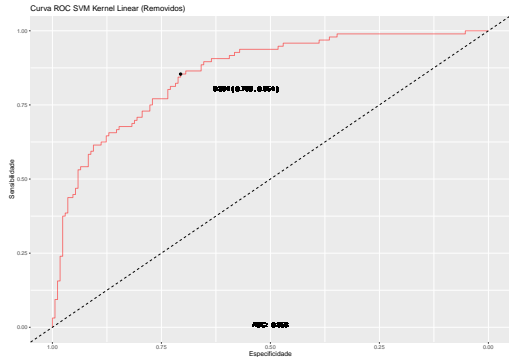
SVM Kernel Linear – Acurácia vs Valores de Cost (Imputados)



# SVM Kernel Linear (Partição 2)

Predito	Referência	
	No	Yes
No	64	12
Yes	4	9

- Validação cruzada ( $k = 10$ )
- Dados sem *missings*





---

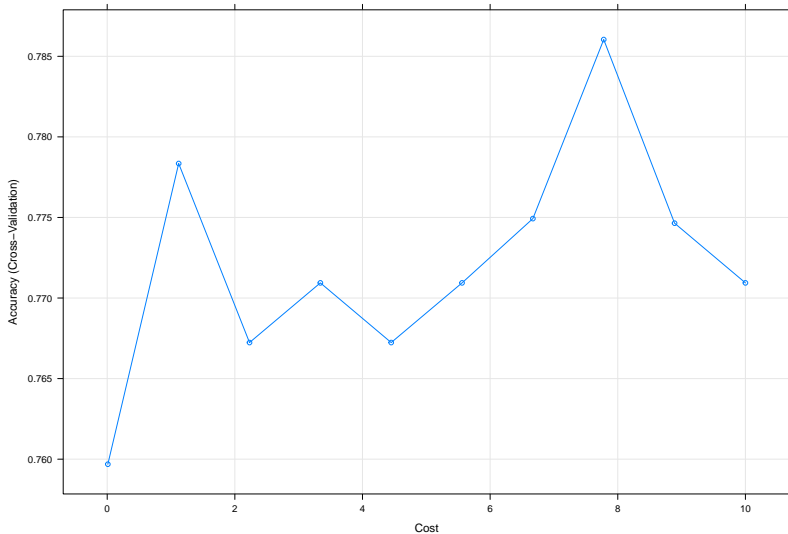
**Desempenho do Modelo**


---

Accuracy	0.8202	Sensitivity	0.9412
95% CI	(0.7245, 0.8936)	Specificity	0.4286
No Information Rate	0.764	Pos Pred Value	0.8421
P-Value [Acc >NIR]	0.12910	Neg Pred Value	0.6923
		Prevalence	0.7640
Kappa	0.4258	Detection Rate	0.7191
		Detection Prevalence	0.8539
Mcnemar's Test P-Value	0.08012	Balanced Accuracy	0.6849

---

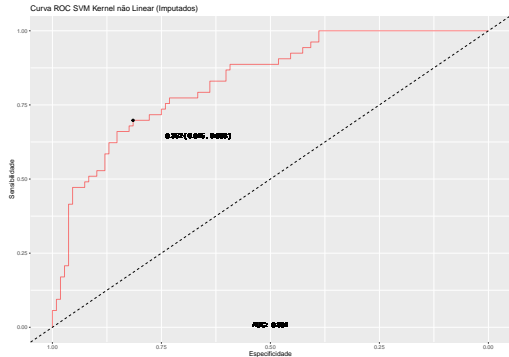
SVM Kernel Linear – Acurácia vs Valores de Cost (Removidos)



# SVM *Kernel* Não-Linear (Partição 1)

Predito	Referência	
	No	Yes
No	99	26
Yes	9	27

- Validação cruzada ( $k = 10$ )
- Dados imputados



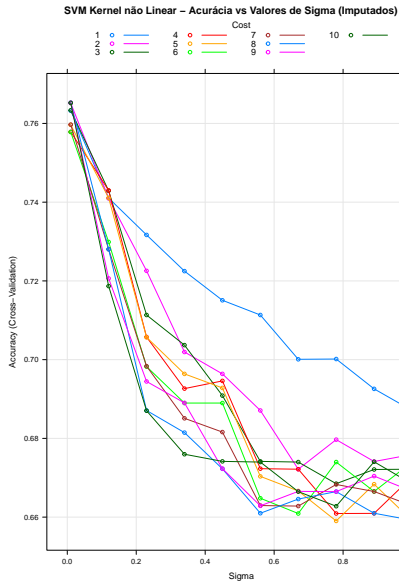
---

### Desempenho do Modelo

---

Accuracy	0.7826	Sensitivity	0.9167
95% CI	(0.7109, 0.8437)	Specificity	0.5094
No Information Rate	0.6708	Pos Pred Value	0.7920
P-Value [Acc >NIR]	0.001230	Neg Pred Value	0.7500
		Prevalence	0.6708
Kappa	0.464	Detection Rate	0.6149
		Detection Prevalence	0.7764
McNemar's Test P-Value	0.006841	Balanced Accuracy	0.7131

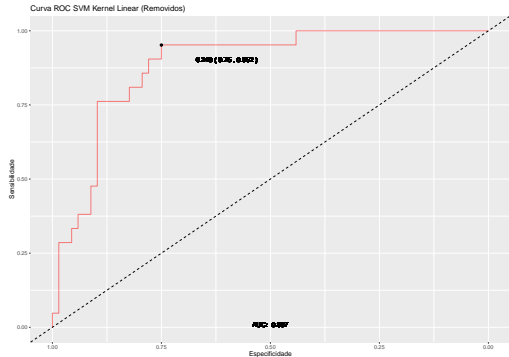
---



# SVM *Kernel* Não-Linear (Partição 2)

Predito	Referência	
	No	Yes
No	63	13
Yes	5	8

- Validação cruzada ( $k = 10$ )
- Dados sem *missings*



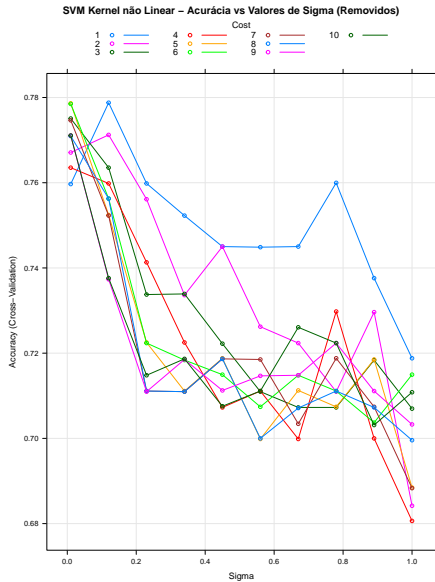
---

### Desempenho do Modelo

---

Accuracy	0.7978	Sensitivity	0.9265
95% CI	(0.6993, 0.8755)	Specificity	0.3810
No Information Rate	0.764	Pos Pred Value	0.8289
P-Value [Acc >NIR]	0.27088	Neg Pred Value	0.6154
		Prevalence	0.7640
Kappa	0.354	Detection Rate	0.7079
		Detection Prevalence	0.8539
McNemar's Test P-Value	0.09896	Balanced Accuracy	0.6537

---





# Comparação dos Modelos

**Métricas:** Acurácia, Sensibilidade, Especificidade e Acurácia Balanceada

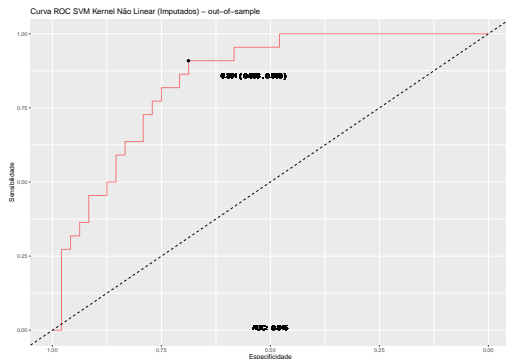
Modelo	Imputados				Removidos			
	Acuracia	Sensibilidade	Especificidade	Acurácia Balanceada	Acurácia	Sensibilidade	Especificidade	Acurácia Balanceada
AD Linear	0.77	0.91	0.47	0.69	0.82	0.94	0.43	0.68
AD Flexível	0.76	0.88	0.53	0.70	0.80	0.90	0.48	0.69
AD Quadrática	0.76	0.89	0.49	0.69	0.84	0.88	0.71	0.8
RL Simples	0.78	0.91	0.49	0.70	0.83	0.94	0.48	0.71
RL Regularizada	0.77	0.91	0.47	0.69	0.83	0.94	0.48	0.71
Random Forest	0.73	0.86	0.47	0.67	0.80	0.87	0.57	0.72
SVM Linear	0.77	0.91	0.49	0.7	0.82	0.94	0.43	0.68
SVM Não Linear	0.78	0.92	0.51	0.71	0.80	0.93	0.38	0.65

# SVM *Kernel* Não-Linear (Partição 1)

**Teste:** Desempenho sobre *out-of-sample*

Predito	Referência	
	No	Yes
No	38	8
Yes	10	14

- Validação cruzada ( $k = 10$ )
- Dados imputados



## Desempenho do Modelo

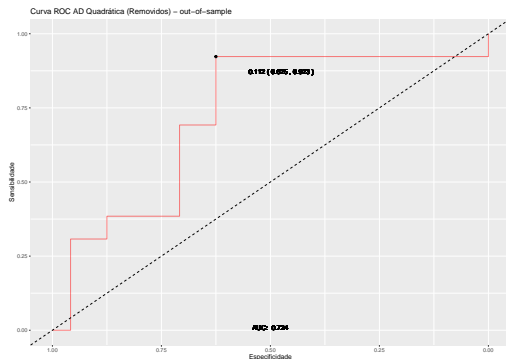
Accuracy	0.7429	Sensitivity	0.7917
95% CI	(0.6244, 0.8399)	Specificity	0.6364
No Information Rate	0.6857	Pos Pred Value	0.8261
P-Value [Acc >NIR]	0.1846	Neg Pred Value	0.5833
		Prevalence	0.6857
Kappa	0.4177	Detection Rate	0.5429
		Detection Prevalence	0.6571
Mcnemar's Test P-Value	0.8137	Balanced Accuracy	0.7140

# Análise Discriminante Quadrática (Partição 2)

**Teste:** Desempenho sobre *out-of-sample*

Predito	Referência	
	No	Yes
No	17	5
Yes	7	8

- Validação cruzada ( $k = 10$ )
- Dados sem *missings*



### Desempenho do Modelo

Accuracy	0.6757	Sensitivity	0.7083
95% CI	(0.5021, 0.8199)	Specificity	0.6154
No Information Rate	0.6486	Pos Pred Value	0.7727
P-Value [Acc >NIR]	0.4384	Neg Pred Value	0.5333
		Prevalence	0.6486
Kappa	0.3127	Detection Rate	0.4595
		Detection Prevalence	0.5946
Mcnemar's Test P-Value	0.7728	Balanced Accuracy	0.6619

# Outras implementações...



Dúvidas, comentários ou sugestões?

### Envie-nos um e-mail

- [icaromsc@usp.br](mailto:icaromsc@usp.br)
- [rayscarvalho@usp.br](mailto:rayscarvalho@usp.br)
- [nakahara@usp.br](mailto:nakahara@usp.br)
- [rodrigo.marcel.oliveira@usp.br](mailto:rodrigo.marcel.oliveira@usp.br)
- [vitorhugo@usp.br](mailto:vitorhugo@usp.br)

## Referências:

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*. In: Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer New York.

Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2 ed. Springer New York.