



THE DEVELOPER'S CONFERENCE

Trilha – Saúde 4.0

Wesllei Felipe Heckler

Graduado em Ciência da Computação

Software Developer e Scrum Master na CIGAM Software de Gestão

Utilizando Machine Learning para prever tendências de abandono em um Programa de Reabilitação Pulmonar

Roteiro



- Doenças Respiratórias Crônicas
- Reabilitação Pulmonar
- Análise Preditiva
 - Pré-processamento
 - Modelagem
 - Validação
 - Desenvolvimento da Ferramenta
- Resultados

Doenças Respiratórias Crônicas



THE
DEVELOPER'S
CONFERENCE

685.031 óbitos
por DRC no Brasil
entre 2003 e 2013

2015
3,2 milhões de
óbitos por DPOC

Crescimento de
11,6%
em comparação
com 1990



2015
400.000 óbitos
por Asma

2015
44% dos brasileiros apresentam
sintomas de doenças respiratórias

2017
DPOC foi a 4º
maior causa de morte
no mundo



2020
Passará a ser a 3º
maior causa de morte
no mundo

Doenças Respiratórias Crônicas



- O que são as doenças respiratórias crônicas?
- Causas/Fatores de Risco
- Sintomas
- Consequências

Reabilitação Pulmonar



- Reduzir os sintomas e melhorar a qualidade de vida dos pacientes
- Abordagem multidisciplinar
- Tratamento complementar
- Adaptado às necessidades de cada paciente

Reabilitação Pulmonar



- Programa de Reabilitação Pulmonar da Universidade Feevale
 - Criado em 2002
 - Dados armazenados em uma planilha Excel
 - Dificuldade na análise dos dados
 - Abandono

Dataset



- 566 pacientes atendidos até o início de 2018
- 356 atributos sobre os pacientes
 - Perfil: doença, idade, gênero, altura, peso...
 - Indicadores de saúde: número de vezes que o paciente tossiu, pressão no peito, ocorrência de dispneia...
 - Indicadores de qualidade de vida: limitação de atividades, qualidade do sono, nível de disposição...
 - Desempenho em 10 testes de carga máxima

Objetivos



- Predição de Abandono
- Extrair padrões sobre os dados

Técnicas



- Linguagem de Programação R
- Algoritmos de Machine Learning
 - Support Vector Machine
 - Decision Tree
 - Random Forest
- 10-fold cross validation

Técnicas

- Métricas de Validação
 - Sensibilidade
 - Especificidade
 - Area Under The ROC Curve
 - Acurácia
 - Precisão

Fases da Análise Preditiva



THE
DEVELOPER'S
CONFERENCE

Definição de Objetivos e Técnicas

Pré-processamento

Modelagem

Validação

Desenvolvimento da Ferramenta

Seleção de Atributos



- Doença
- Percentual de Espirometria VEF1
- Gênero
- Idade
- Número de internações no último ano
- Presença de Hipertensão Arterial Sistêmica
- Presença de Diabetes
- Presença de Cardiopatias
- Outras comorbidades
- Escala de Dispneia (gravidade da falta de ar)

Seleção de Atributos



➤ Indicador de Abandono

Limpeza de dados



- Valores NA
- Preenchimento
 - Avaliação de especialista
 - Valor padrão
 - Média, Moda, Mediana...
 - Predição

Limpeza de Dados



- Preenchimento manual com apoio da especialista
 - Doença (com base no percentual de espirometria)
 - Indicadores de comorbidades
- Preenchimento automático com valor padrão
 - Percentual de espirometria VEF1
 - Idade (valor padrão e média)
 - Número de internações no último ano
 - Escala de dispneia

Transformação de Dados



- Gênero
- Indicadores de comorbidades
- Padronização do atributo Doença
 - Redução de 50 para 27 rótulos

Modelagem



THE
DEVELOPER'S
CONFERENCE



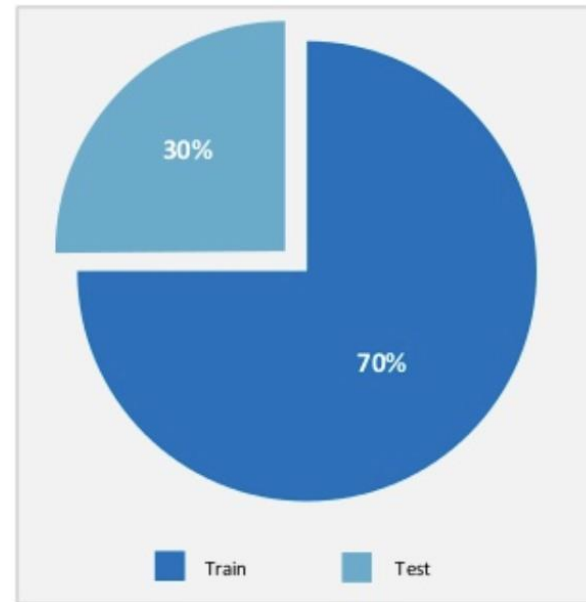
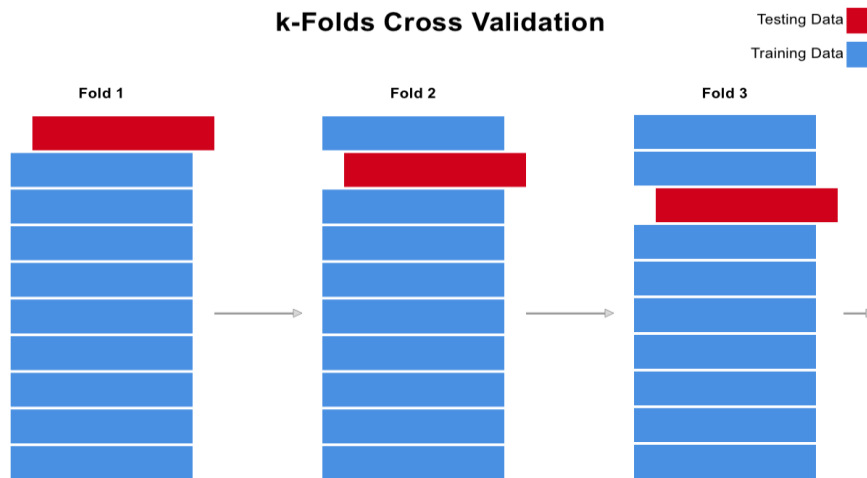
Completo



Sem escala de dispneia



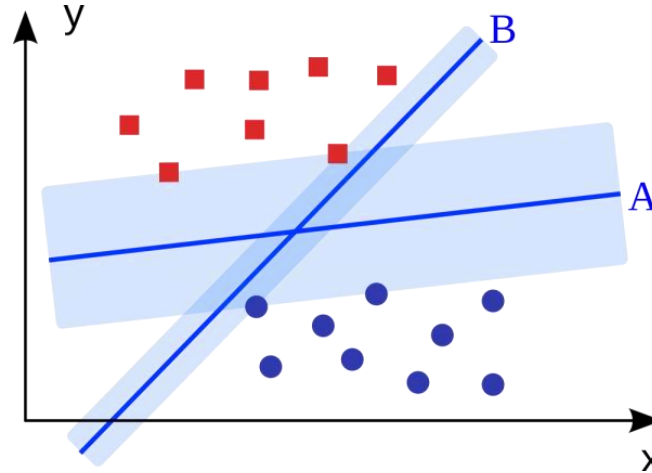
k-Folds Cross Validation



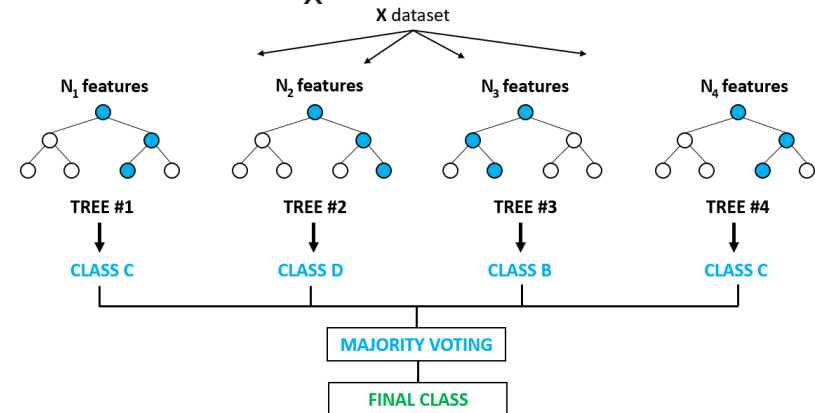
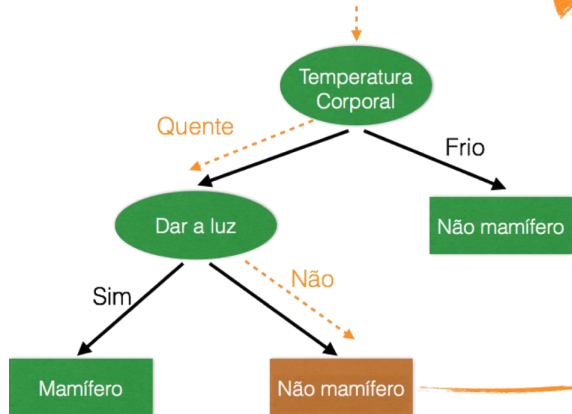
Algoritmos



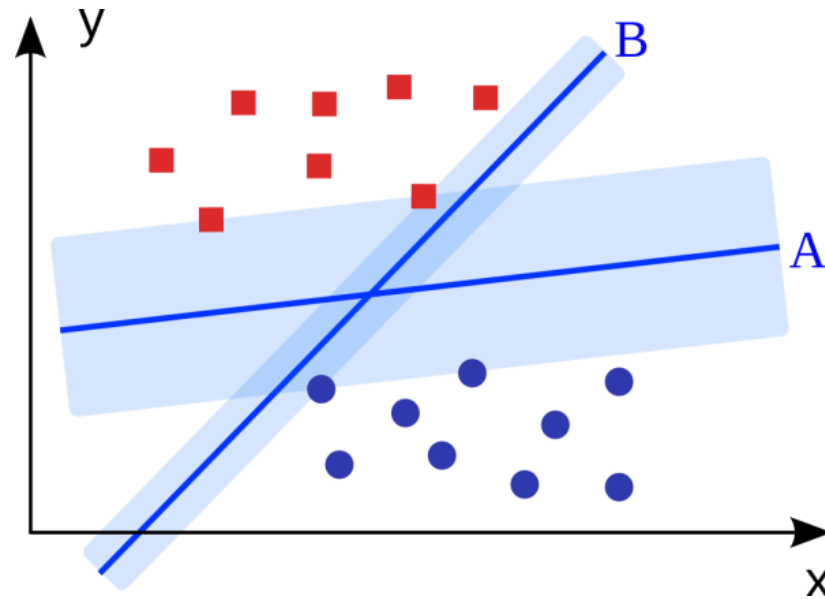
THE
DEVELOPER'S
CONFERENCE



Nome	Temperatura Corporal	Dar a Luz	Classe
Flamingo	Sangue Quente	Não	?



Support Vector Machine



Support Vector Machine



➤ Parâmetros

- Kernel Linear

- Custo (0.02 a 0.5, variando em 0,02)

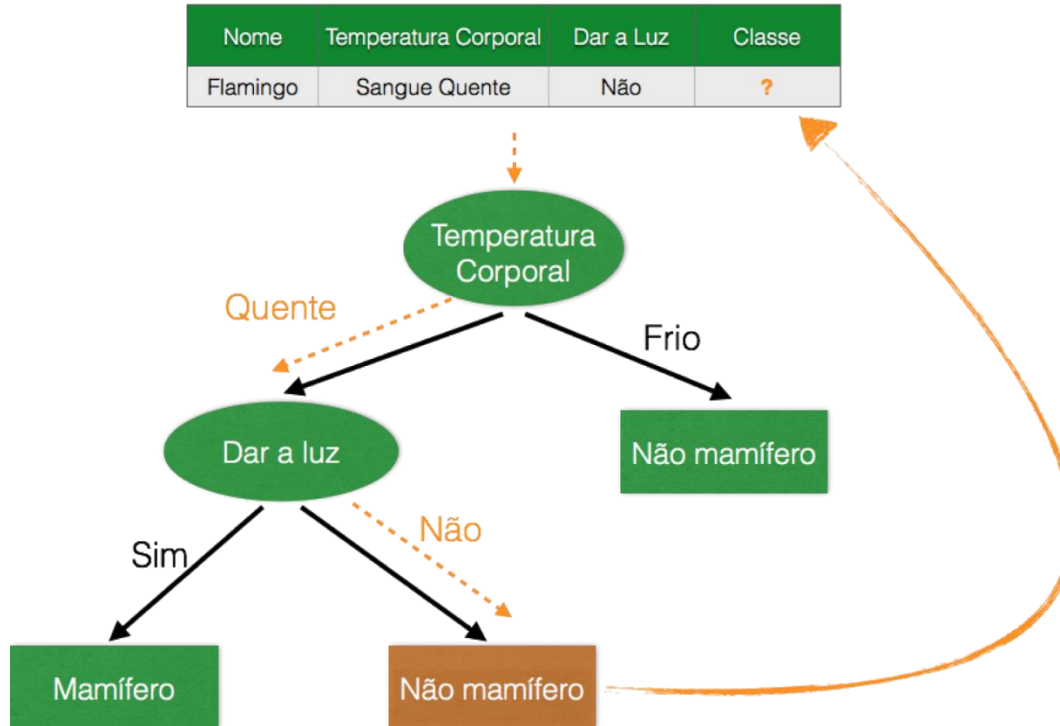
 - Dataset completo = 0.02

 - Dataset sem escala dispneia = 0.06

Decision Tree



THE
DEVELOPER'S
CONFERENCE



Decision Tree



➤ Parâmetros

➤ Ganho de informação (Entropia ou Gini)

➤ Entropia

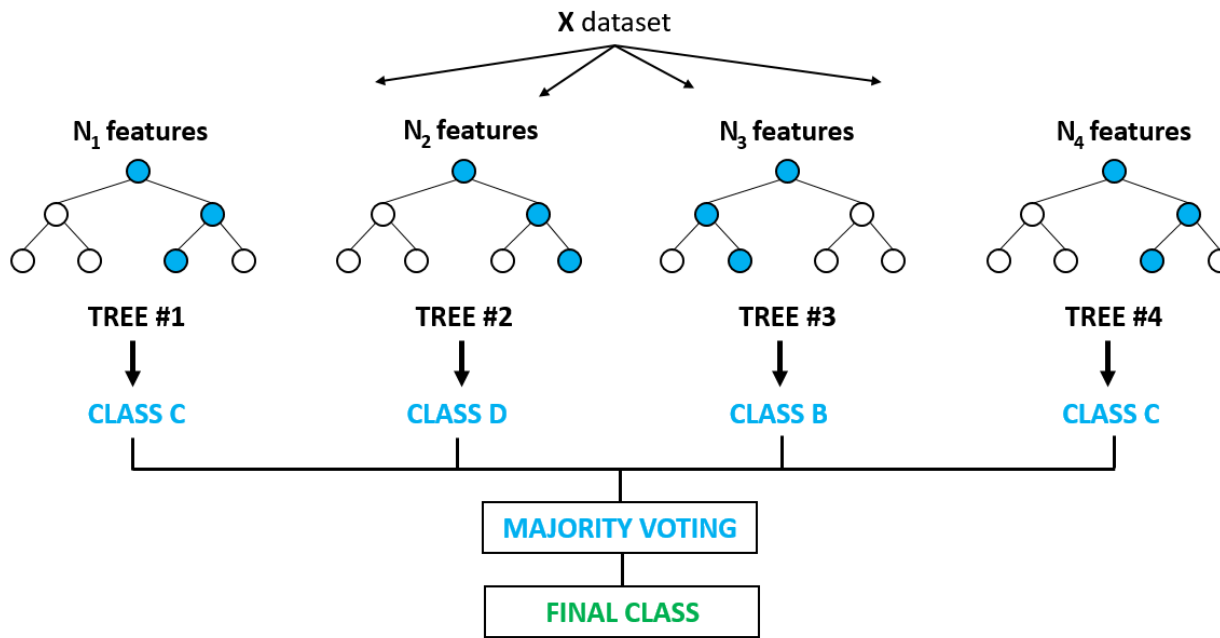
➤ Quantidade mínima de registros por nodo (entre 1 e 20)

➤ 14 registros

Random Forest



THE
DEVELOPER'S
CONFERENCE



Random Forest



➤ Parâmetros

- Número de preditores (entre 1 e 10)
 - Dataset completo = 9
 - Dataset sem escala dispneia = 2
- Número de árvores geradas
 - Dataset completo = 350
 - Dataset sem escala dispneia = 450

Comparação de Desempenho



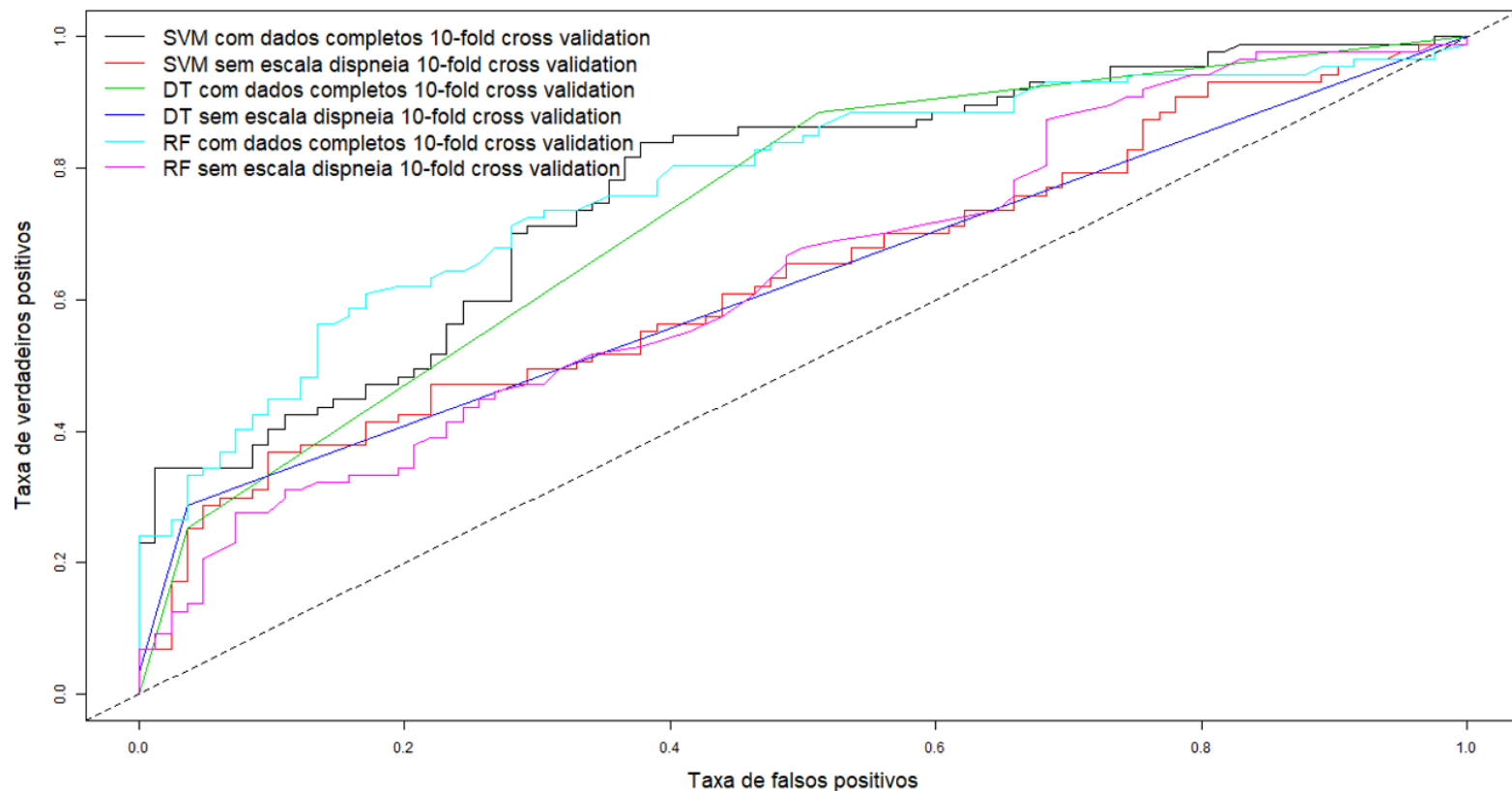
THE
DEVELOPER'S
CONFERENCE

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	AUC
SVM com dados completos	0.6746	0.7561	0.5977	0.6392	0.7684
SVM sem escala dispnea	0.6095	0.9024	0.3333	0.5606	0.6362
DT com dados completos	0.6923	0.4878	0.8851	0.8	0.7350
DT sem escala dispnea	0.6154	0.9634	0.2874	0.5603	0.6260
RF com dados completos	0.6982	0.7683	0.6322	0.6632	0.7735
RF sem escala dispnea	0.5740	0.8415	0.3218	0.5391	0.6370

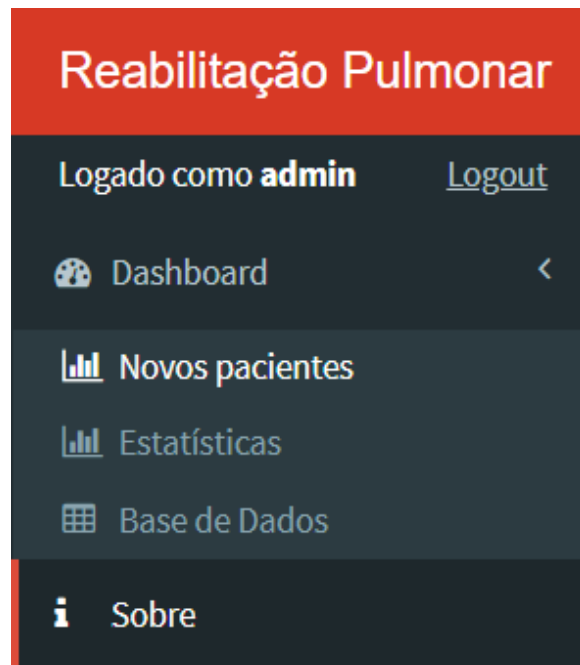
Comparação de Desempenho



THE
DEVELOPER'S
CONFERENCE



Ferramenta



Novos Pacientes



THE
DEVELOPER'S
CONFERENCE

Reabilitação Pulmonar

Logado como **admin** [Logout](#)

[Dashboard](#)

[Novos pacientes](#)

[Estatísticas](#)

[Base de Dados](#)

[Sobre](#)

Selecionar tabela com dados de novos pacientes

Selecionar

tabela-novos.csv

Upload complete

Selecionar técnica para predição de tendências de abandono

Support Vector Machine

Acurácia: $0.67 \frac{(VP+VN)}{(VP+FP+FN+VN)}$

Sensibilidade: $0.58 \frac{(VP)}{(VP+FN)}$

Especificidade: $0.76 \frac{(VN)}{(FP+VN)}$

Precisão: $0.7 \frac{(VP)}{(VP+FP)}$

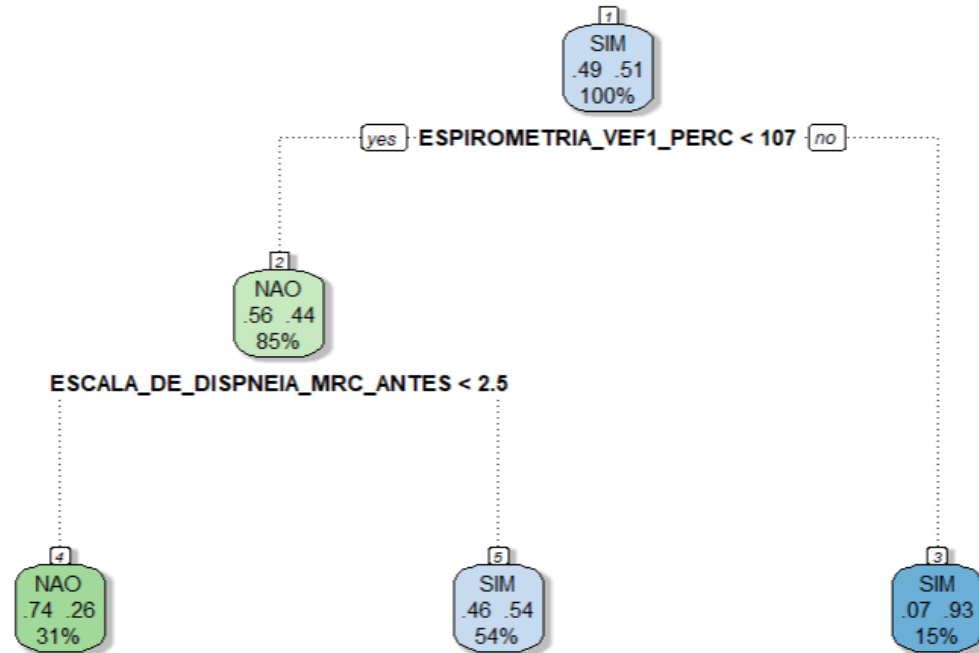
Matriz de Confusão

	NAO		SIM	
NAO	159	(VP)	69	(FP)
SIM	116	(FN)	222	(VN)

Tabela de Novos Pacientes

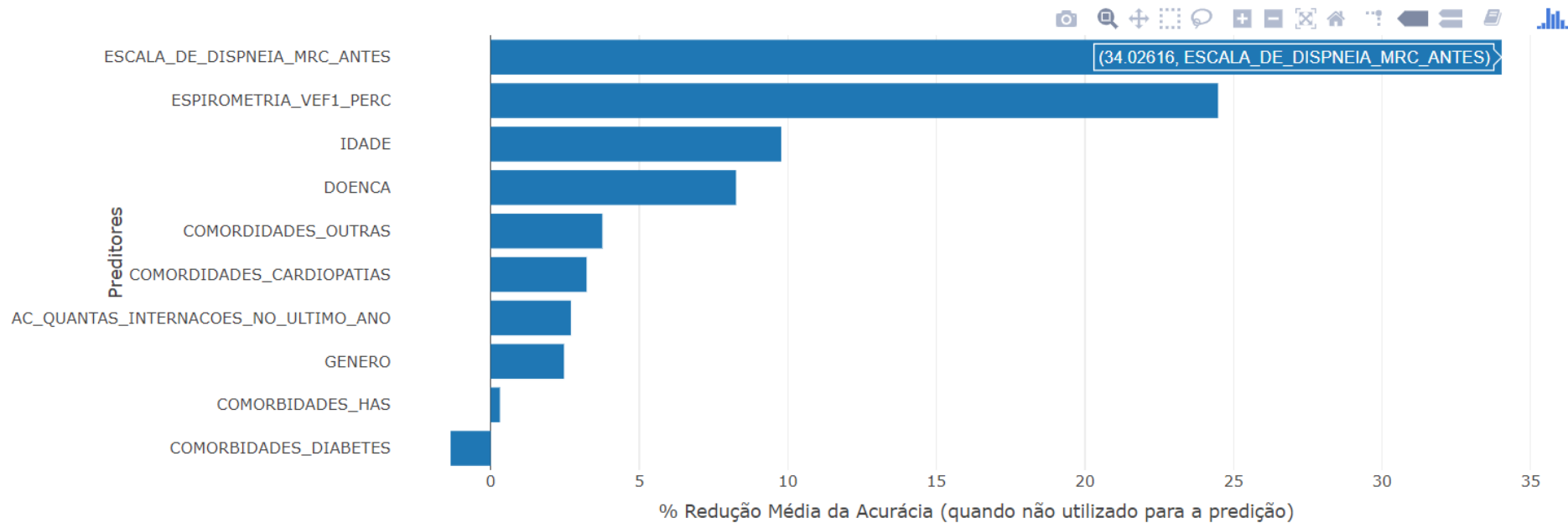
TENDÊNCIA DE ABANDONO	DOENÇA	ESPIROMETRIA VEF1 %	GÊNERO	IDADE	INTERNAÇÕES NO ÚLTIMO ANO	COMORBIDADES HAS	COMORBIDADES DIABETES	COMORBIDADES CARDIOPATIAS	COMORBIDADES OUTRAS	ESCALA DE DISPNEIA MR ANTE
NAO	ASMA E BRONQUIEC	42	MASCULINO	85	0	NAO	NAO	NAO	NAO	
SIM	ASMA	31	FEMININO	74	0	SIM	NAO	NAO	NAO	
SIM	ASMA E BRONQUIEC	150	MASCULINO	76	0	NAO	NAO	NAO	NAO	
SIM	BQT	2.11	FEMININO	58	0	NAO	NAO	NAO	NAO	
SIM	BQT	83.8	FEMININO	63	4	SIM	SIM	NAO	NAO	
SIM	ASMA E BRONQUIEC	68	MASCULINO	57	0	NAO	NAO	NAO	NAO	
SIM	ASMA E BRONQUIEC	47	MASCULINO	67	0	SIM	NAO	NAO	NAO	
SIM	ASMA +	67	FEMININO	76	0	SIM	NAO	NAO	NAO	

Novos Pacientes



Novos Pacientes

Random Forest

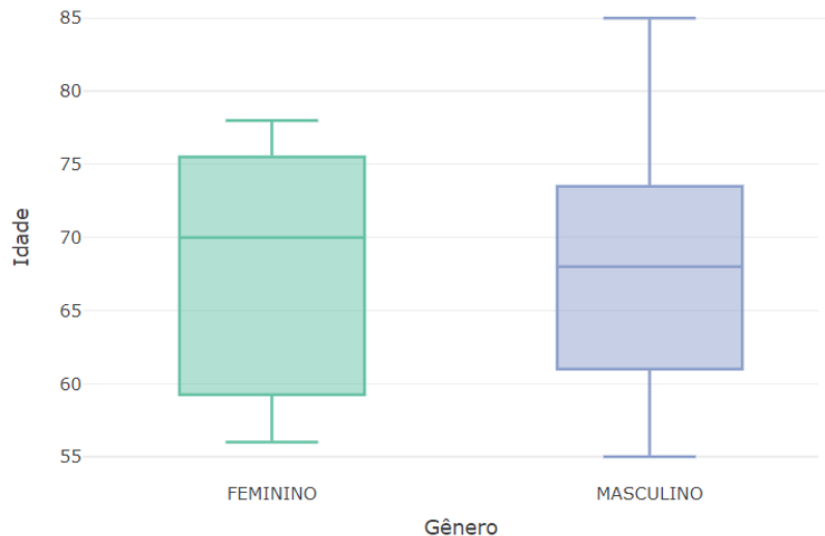


Novos Pacientes

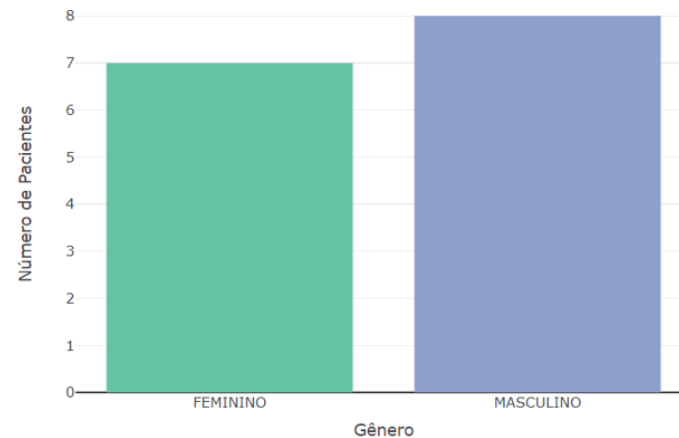


THE
DEVELOPER'S
CONFERENCE

Idade x Gênero



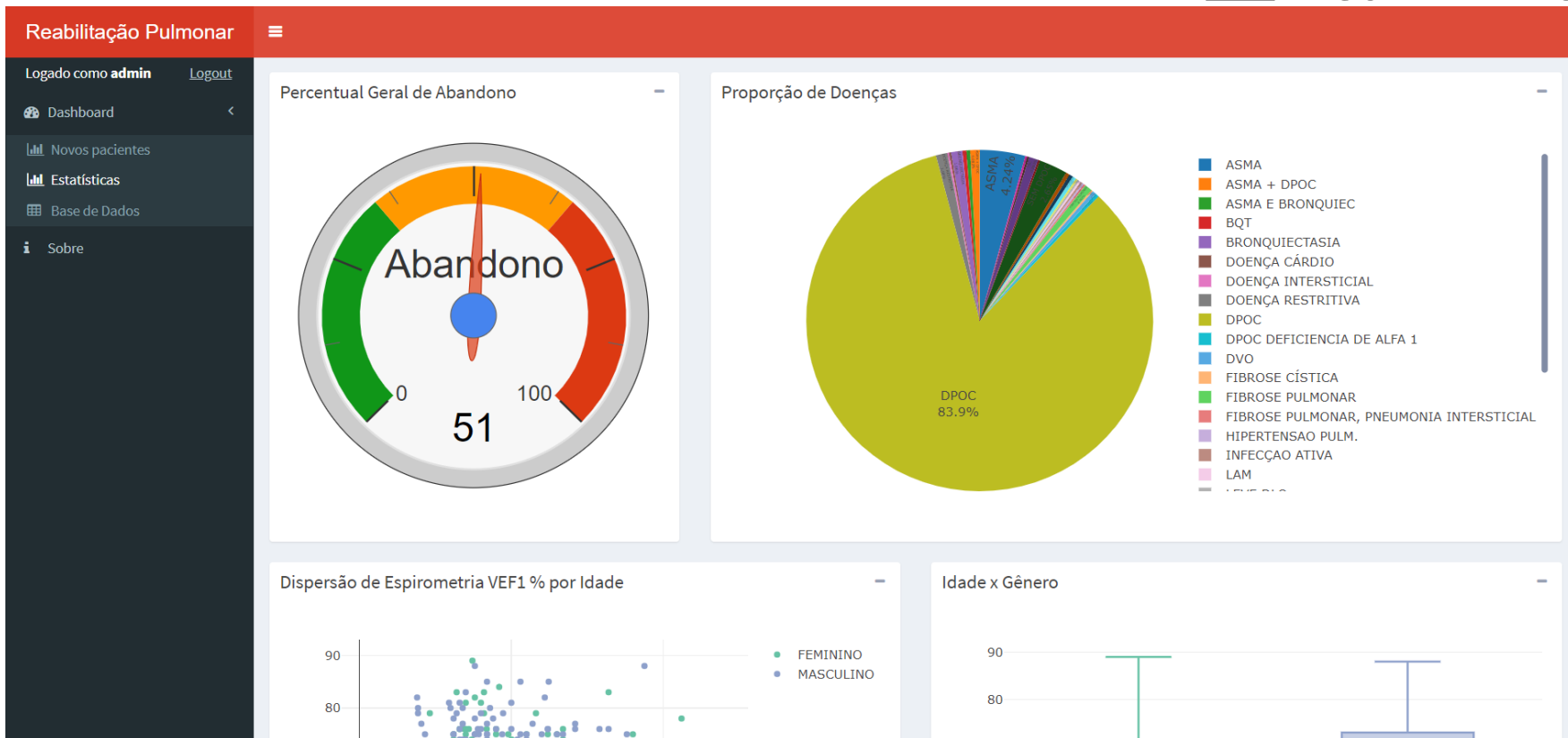
Pacientes por Gênero



Estatísticas



THE
DEVELOPER'S
CONFERENCE

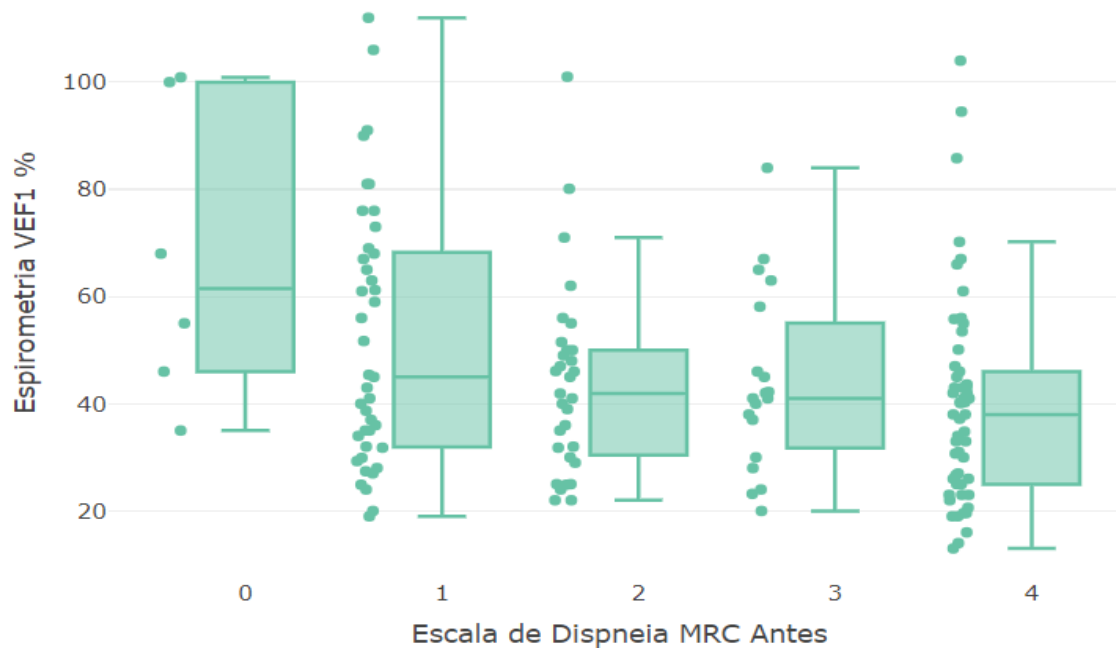


Estatísticas



THE
DEVELOPER'S
CONFERENCE

Escala de Dispneia MRC Antes por Espirometria VEF1 % (Gênero = Feminino) —

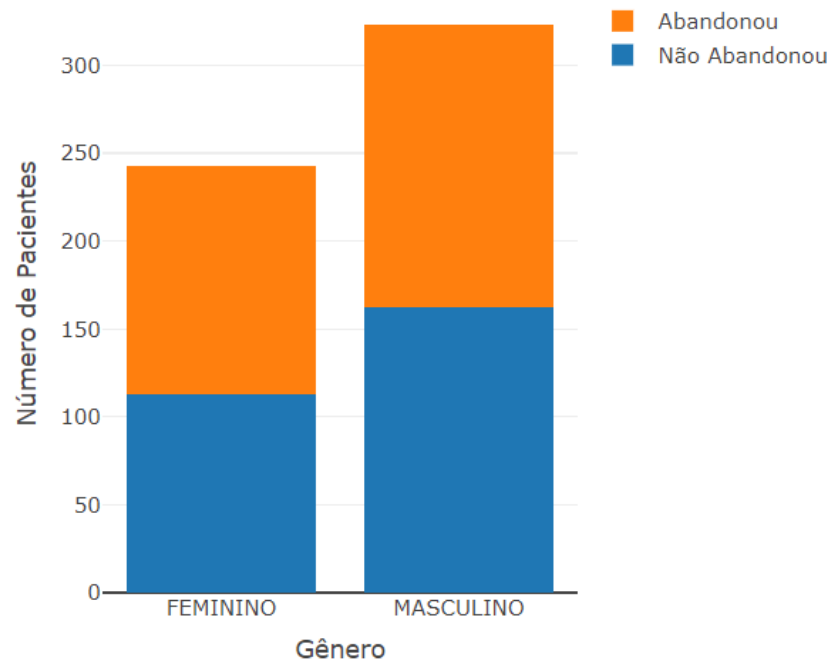


Estatísticas



THE
DEVELOPER'S
CONFERENCE

Abandono por Gênero

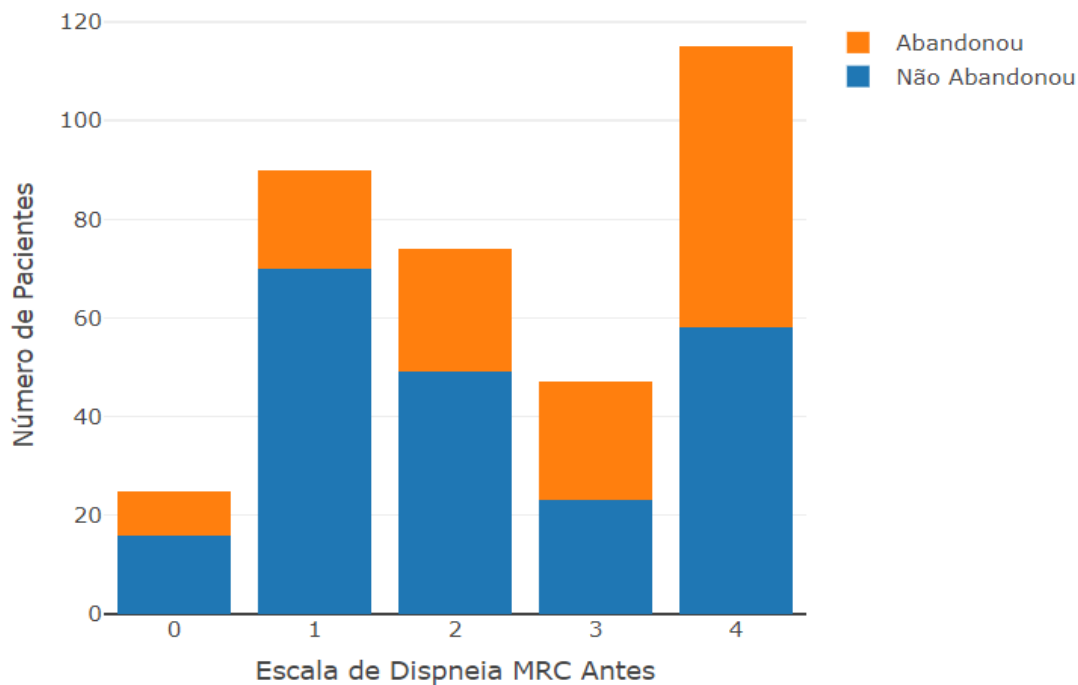


Estatísticas



THE
DEVELOPER'S
CONFERENCE

Abandono por Escala de Dispneia Antes



Base de Dados



THE
DEVELOPER'S
CONFERENCE

Reabilitação Pulmonar

Logado como **admin** [Logout](#)

[Dashboard](#)

[Novos pacientes](#)

[Estatísticas](#)

[Base de Dados](#)

[Sobre](#)

Baixar tabela completa com dados de pacientes.

Importar tabela com dados de novos pacientes

Selecionar Nenhuma tabela selecionada.

Tabela de Pacientes

ABANDONOU	DOENÇA	ESPIROMETRIA VEF1 %	GÊNERO	IDADE	INTERNAÇÕES NO ÚLTIMO ANO	COMORBIDADES HAS	COMORBIDADES DIABETES	
All	All	All	All		All	All	All	
SIM	DPOC	45	MASCULINO	65	1	NAO	NAO	N/A
SIM	ASMA	150	FEMININO	63	-1	NAO	NAO	N/A
SIM	DPOC	34	MASCULINO	74	0	NAO	NAO	N/A
SIM	DPOC	150	MASCULINO	80	0	NAO	NAO	N/A
SIM	DPOC	60	MASCULINO	72	1	NAO	NAO	N/A
SIM	DPOC	150	MASCULINO	75	1	NAO	NAO	N/A
SIM	DPOC	150	MASCULINO	76	0	NAO	NAO	N/A
SIM	DPOC	34	MASCULINO	77	2	NAO	SIM	N/A
SIM	DPOC	83	MASCULINO	51	0	NAO	NAO	N/A

Limitações



- Quantidade de registros
 - Utilizar somente 10-fold cross validation
 - Aumentar dataset

Resultados



- Random Forest: 69,82% de acurácia
- Ferramenta
- Estratégias para redução de abandonos
- Oportunidade de exploração de novos padrões



THE DEVELOPER'S CONFERENCE



/weslleheckler



weslleheckler@gmail.com