CANTÍDIO DE MOURA CAMPOS NETO

**Análise inteligente de dados em um banco de dados de
procedimentos em cardiologia intervencionista**

Tese apresentada ao Instituto Dante
Pazzanese de Cardiologia – Entidade
Associada da Universidade de São
Paulo, para obtenção do título de
Doutor em Ciências.

Programa de Medicina, Tecnologia e
Intervenção em Cardiologia.

Orientador: Prof. Dr. Denys Emílio
Campion Nicolosi

Versão corrigida. Resolução CoPGr 6018/11, de 01 novembro de 2011. A
versão original está disponível na Biblioteca do IDPC.

**São Paulo
2016**

Dados Internacionais de Catalogação na Publicação (CIP)
Preparada pela Biblioteca do Instituto Dante Pazzanese de Cardiologia
©reprodução autorizada pelo autor

Campos Neto, Cantídio de Moura

Análise inteligente de dados em banco de dados de procedimentos em cardiologia intervencionista/Cantídio de Moura Campos Neto -- São Paulo, 2016.

Tese(doutorado)--Instituto Dante Pazzanese de Cardiologia Universidade de São Paulo

Área de Concentração: Medicina, Tecnologia e Intervenção em Cardiologia

Orientador: Prof. Dr. Denys Emílio Campion Nicolosi

Descritores: 1. Mineração de Dados. 2. Árvores de Decisão. 3. Cardiologia. 4. Doença das Coronárias 5. Stents

USP/IDPC/Biblioteca/64/16

*Dedico este trabalho à minha esposa Ana Maria e
às minhas filhas Gabriela e Júlia.
Elas souberam entender o meu desafio e me
apoiaram de forma irrestrita.*

AGRADECIMENTOS

Mais uma etapa finalizada e a constatação pessoal de que a limitação somos nós que a criamos.

Desde cedo, meus pais me ensinaram que o conhecimento assim como o amor e o respeito ao próximo estão no topo da escala de valores, e esta Tese é mais uma prova destes ensinamentos.

Tenho consciência de que o resultado de qualquer trabalho sempre vem do coletivo e este é o momento de agradecer às pessoas que me auxiliaram na realização deste trabalho.

Agradeço aos meus pais e à minha família pelos ensinamentos e pelo amor que fizeram parte de minha formação, o início de tudo.

À minha esposa Ana Maria e às minhas filhas Gabriela e Júlia pelo amor e pela compreensão com quem divido esta conquista.

À Gabriela Moura Campos, pelas orientações na elaboração gráfica deste trabalho.

Ao amigo e companheiro Prof. Denys Emilio Campion Nicolosi, um visionário, que, pelo seu espírito inovador, aceitou me orientar neste tema desafiante e atual.

Um agradecimento especial à Profa. Solange Oliveira Rezende, do Laboratório de Inteligência Computacional (LABIC) do Departamento de Ciências de Computação ICMC-USP-São Carlos, a quem conheci em 2012 numa visita ao IDPC. Em companhia de meu colega José Alves Ferreira, doutorando na época, fizemos uma visita ao LABIC, a convite da Profa. Solange, para tratar do seu tema de tese, e fiquei impressionado com a dinâmica da reunião e o convívio harmonioso do grupo. A possibilidade de

poder frequentar o LABIC foi um dos motivos de ter realizado este trabalho. Sou muito grato pela forma como fomos recebidos, pela sua generosidade, e pelas orientações recebidas.

Aos doutorandos do LABIC com quem aprendi muito, Rafael G. Rossi pelo apoio e pela paciência dispensada em atendimento às minhas dúvidas, e ao Fabiano Fernandes, por nossas conversas e *e-mails* trocados sobre mineração de dados.

Sem dados, não há mineração, esta é uma frase fundamental aprendida desde o início. Meus agradecimentos ao Serviço de Cardiologia Intervencionista do HCor liderado pelo Prof. José Eduardo Moraes Rego Sousa responsável pela criação e coordenação do Registro Desire, à Dra. Amanda Guerra de Moraes Rego Sousa, pelo incentivo e pela liderança do grupo de estudos do Desire, responsável pela produção de diversas publicações, em congressos e nos principais periódicos nacionais e internacionais.

Ao Dr. José de Ribamar Costa Jr. com quem muito aprendi e que, apesar de sua agenda, esteve sempre presente quando solicitado.

A toda equipe do Desire, responsável pela qualidade das informações, Dra. Adriana Costa Moreira, Dr. Ricardo Alves da Costa, Lucas Damiani, Luciana Alves e Juliana Ramalho, pelas valiosas sugestões.

Ao Programa de Pós-Graduação USP-IDPC, pela oportunidade oferecida.

Ao Prof. Júlio Cesar Rodrigues Pereira, pela qualidade das aulas ministradas em Bioestatística e ao João Ítalo pela assistência recebida durante e após o curso.

À Ana Simene, pelas orientações recebidas na fase de escrita e por suas sugestões.

A todos do Departamento de Bioengenharia, pelo apoio de sempre, em especial, ao amigo Reinaldo Akikubo, que foi um grande facilitador desde o início, à Emi Teles, na fase de cadastro junto ao Comitê de Ética em Pesquisa, e à Denize Ferrari, no editor de texto.

E, finalmente, meus agradecimentos aos médicos especialistas do Serviço de Hemodinâmica do IDPC, Dr. Alexandre Antonio Cunha Abizaid, Dra. Marinella Patrícia Centemero, Dr. José de Ribamar Costa Jr., Dr. Luiz Fernando Leite Tanajura e Dr. Ricardo Alves da Costa, pela avaliação das regras geradas pelos modelos de classificação, uma importante etapa do processo de KDD.

Tudo passa...

NORMALIZAÇÃO ADOTADA

Esta tese está de acordo com as seguintes normas, em vigor no momento desta publicação:

Referências: adaptado de *International Committee of Medical Journals Editors* (Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação. *Guia de apresentação de dissertações, teses e monografias*. Elaborado por Anneliese Carneiro da Cunha, Maria Julia de A. L. Freddi, Maria F. Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3a ed. São Paulo: Divisão de Biblioteca e Documentação; 2011.

Abreviaturas dos títulos dos periódicos de acordo com *List of Journals Indexed in Index Medicus*.

SUMÁRIO

LISTA DE SIGLAS E ABREVIATURAS

LISTA DE CLASSIFICAÇÕES E DEFINIÇÕES

LISTA DE FIGURAS

RESUMO

ABSTRACT

1 INTRODUÇÃO.....	2
1.1 O processo de aprendizagem e a criação de um modelo	5
1.2 Histórico	12
2 OBJETIVOS	21
2.1 Objetivo principal.....	21
2.2 Objetivo secundário	22
3 MATERIAIS E MÉTODOS.....	24
3.1 Casuística	24
3.2 Processo de <i>Knowledge-Discovery in Databases</i> KDD	30
3.2.1 Pré-processamento.....	30
3.2.2 Mineração de dados	36
3.2.2.1 A classificação	37
3.2.2.2 Medida de desempenho do classificador.....	39
3.2.2.3 Classificação por Árvore de Decisão	42
3.2.2.4 Indução de Árvores de Decisão	44
3.2.2.5 Construção da Árvore de Decisão	47
3.2.2.6 Indução por regras de classificação.....	55
3.3 Ferramentas para mineração dos dados	57
3.3.1 WEKA	58
3.3.2 Orange.....	61
4 RESULTADOS E DISCUSSÃO.....	64
4.1 Seleção dos atributos.....	65
4.2 Indução e avaliação dos modelos	68
4.3 Criação de duas novas classes	73
4.3.1 Classe MFP	73

4.3.2 Classe MACE_REFERENCIA.....	73
4.4 Detalhe do tipo de evento dentro da classe	78
4.4.1 Classe: MFP_FIRST e Classe: MACE_FIRST.....	78
4.5 Extração das regras pelas árvores de decisão	81
4.5.1 Regras para a classe MACE.....	83
4.5.2 Regras para a classe MFP.....	85
4.5.3 Regras para a classe MACE_REFERENCIA.....	86
4.6 Regras de classificação	88
5 CONCLUSÃO	98
6 ANEXOS.....	104
6.1 ANEXO A - Lista dos atributos do Registro Desire extraídos da base de dados com suas definições e descrições	104
6.2 ANEXO B - Dataset Desire: análise dos atributos categóricos.....	106
6.3 ANEXO C - (Modelos P-1, P-2 e P-3) Classe: MACE.....	111
6.4 ANEXO D - <i>Dataset</i> com as variáveis numéricas e categóricas.	120
6.5 ANEXO E - Seleção dos 20 atributos mais significativos pelo critério da razão de ganho (GR) para as classes MACE e MFP.	122
6.6 ANEXO F - Avaliação das regras extraídas pelos especialistas de domínio.....	123
7 REFERÊNCIAS	144

LISTA DE SIGLAS E ABREVIATURAS

AC	Angioplastia Coronária
ACC	<i>American College of Cardiology</i>
AD	Árvore de Decisão
AUC	<i>Area Under Curve</i>
AHA	<i>American Heart Association</i>
AM	Aprendizado de Máquina
ARFF	<i>Attribute Relation File Format</i> , sigla usada para designar arquivo contendo dados no formato de lista com exemplos que partilham atributos comuns
C4. 5	Algoritmo de classificação para indução de árvore de decisão
CART	<i>Classification and Regression Trees</i> (algoritmo de classificação para indução de árvore de decisão)
CD	Artéria coronária direita
Cx	Artéria circunflexa
CIR	Cirurgia
CSV	<i>Comma Separated Values</i>
DA	Artéria Descendente Anterior
DAC	Doença Arterial Coronária
DESIRE	<i>Drug-Eluting Stents in the Real World</i>
DML	Diâmetro Mínimo da Luz do Vaso
ECAM	Evento Cardíaco Adverso Maior
FPR	<i>False Positive Rate</i> (Falsos positivos)
FVE	Função Ventricular Esquerda
HDL	<i>High Density Lipoproteins</i> (Lipoproteínas de alta densidade)
ID3	<i>Iterative Dichotomiser</i> (algoritmo de classificação para indução de árvore de decisão).
ICP	Intervenção Coronária Percutânea
IA	Inteligência Artificial

IAM	Infarto Agudo do Miocárdio
IP	Intervenção Percutânea
IMC	Índice de Massa Corpórea
KDD	<i>Knowledge Discovery in Database</i> , sigla usada para designar a descoberta de conhecimento em base de dados
LDL	<i>Low Density Lipoproteins</i> (Lipoproteínas de baixa densidade)
MD	Mineração de Dados
MIX	Indica que foram implantados stents de primeira e segunda geração no mesmo procedimento index
RLA	Revascularização da Lesão-Alvo
RM	Revascularização do Miocárdio
ROC	<i>Receiver Operating Characteristic</i>
SCA	Síndrome Coronária Aguda
SF	Stents farmacológicos
SGBDR	Sistema Gerenciador de Banco de Dados Relacional
SQL	<i>Structured Query Language</i>
TCE	Tronco de Artéria Coronária Esquerda
TPR	<i>True Positive Rate</i> (verdadeiros positivos)
UCI	Universidade da Califórnia – Irvine
Weka	<i>Waikato Enviroment for Knowledge Analysis</i>

LISTA DE CLASSIFICAÇÕES E DEFINIÇÕES

Angina Estável	Classificação dada pela <i>Canadian Cardiovascular Society</i> – CCS (S ou N).
ARFF	<i>Attribute Relation File Format</i> , sigla usada para designar arquivo contendo dados no formato de lista e com os atributos encabeçando as colunas.
AUC	<i>Area Under Curve</i> (área sob a curva ROC medida que corresponde ao desempenho do algoritmo classificador).
CK-MB	Enzima que ajudam no fornecimento de energia das células cardíacas. Valores acima de 3 vezes o limite superior ao normal é um dos critérios para se determinar o IAM.
CSL	<i>Comma Separated Value</i> (valores de atributos separados por vírgula).
Dataset	Conjunto de exemplos retirados de uma base de dados composto pelos atributos que descrevem cada um dos exemplos.
DCE	Doença coronária estável definida pela apresentação clínica que pode ser angina estável ou isquemia silenciosa (S ou N).
Diabetes	(Insulino dependente, Não insulino dependente, uso de medicamento ou não diabético).
ETL	<i>Extract Transform and Load</i> (sigla que representa o processo de extração de dados, seguida pela ação de transformação e de armazená-lo em um repositório de dados).
FVE	Função ventricular esquerda foi avaliada pelo cálculo da fração de ejeção (%FE) e classificada como normal (%FE $\geq$ 55%), disfunção discreta (40% $\leq$ %FE < 55%), disfunção moderada (30% $\leq$ %FE < 40%) e disfunção grave (%FE < 30%).
GR	<i>Gain Ratio</i> (Razão de ganho, um dos critérios de seleção dos atributos para construção da árvore de decisão).
IAM	Infarto agudo do miocárdio.
IAM_Q	Infarto agudo do miocárdio com onda Q (evento).
IAM_NQ	Infarto agudo do miocárdio sem onda Q (evento).

KNN	<i>K-Nearest Neighbor.</i>
LAD	<i>Left Anterior Descending Artery.</i>
LCX	<i>Left Circumflex Artery.</i>
Lesão alvo	A classificação do tipo da lesão-alvo segue definição do ACC/AHA (A, B1, B2 e C).
Lesão longa	Lesões cujo comprimento excede a 30 mm.
MACE	<i>Major Adverse Cardiac Event</i> (atributo-classe que pode ter dois valores S ou N, que corresponde à ocorrência ou não de um evento no paciente na fase hospitalar e na fase pós-hospitalar ao longo dos treze anos do Registro Desire).
MACE_REFERENCIA	Atributo-classe que pode ter dois valores S ou N, que corresponde à ocorrência ou não de um evento ao paciente, excluindo o IAM_NQ na fase hospitalar e no primeiro ano da fase pós-hospitalar do Registro Desire
MCARD	Morte cardíaca (evento).
MFP	<i>Mace First Period</i> (atributo-classe que pode ter dois valores S ou N, que corresponde à ocorrência de um evento ao paciente desde a fase hospitalar até o primeiro ano da fase pós-hospitalar do Registro Desire).
Nova lesão	Lesão resultante da progressão da doença coronariana (evento).
Obesidade	Considerada para valores de IMC > 30.
Off-label	Demais pacientes que não satisfazem a condição On-label.
On-label	Pacientes com lesão única, em artérias coronárias nativas de diâmetro entre 2,5 mm e 3,75 mm, e comprimento que não exceda a 30 mm.
Procedimento Index	O primeiro procedimento de implante do <i>stent</i> realizado no paciente.
RCA	<i>Right Coronary Artery.</i>
RFC	<i>Random Forrest Classification.</i>
Subset	Subconjunto de atributos extraídos de um dataset e com o mesmo número de exemplos.
TLR	<i>Treated Lesion Restenosis</i> (evento).
Trombose do stent	Classificação da trombose do stent segundo o <i>Academic Research Consortium</i> (definitiva, provável ou possível).

Tuplas	Em matemática, tupla é uma lista ordenada de elementos, como, por exemplo: (100, 200, 300, 400). Em ciência da computação, para banco de dados, é uma função finita que mapeia nomes de campos para um conjunto de valores inter-relacionados.
Vaso de fino calibre	Classificação para vasos com diâmetro $\leq 2,5$ mm.
View	Uma view pode ser definida como uma tabela virtual composta por linhas e colunas de dados vindos de uma ou mais tabelas relacionadas por meio de uma consulta.

LISTA DE FIGURAS

Figura 1	Aterosclerose, o efeito obstrutivo da doença nas artérias com redução do fluxo sanguíneo que pode levar a um quadro agudo de obstrução total (infarto agudo do miocárdio).....	2
Figura 2	Imagem de um stent convencional antes e após sua expansão (Bare metal stent).....	4
Figura 3	Ilustração de um stent farmacológico Biomatrix (Fármaco: BiolimusA9™).	4
Figura 4	Configuração básica de um problema de aprendizagem.....	6
Figura 5	Etapas operacionais do processo de KDD.	8
Figura 6	Diagrama que mostra o relacionamento da Mineração de Dados com várias disciplinas.....	10
Figura 7	Tarefas de mineração de dados.	11
Figura 8	Desire, total de pacientes, territórios, vasos, lesões tratadas e stents implantados. Nº de lesões tratadas por paciente e nº de stents implantados por paciente.....	26
Figura 9	Características clínicas dos pacientes tratados.	27
Figura 10	Características angiográficas dos pacientes do Registro Desire.	27
Figura 11	Desire, pacientes com histórico cardiovascular e sua classificação segundo o IMC agrupado pelo sexo.....	28
Figura 12	Desire, incidência do Diabetes segundo o sexo.	28
Figura 13	Desire, incidência de Hipertensão e Hiperlipidemia nos pacientes segundo o sexo.	29
Figura 14	Desire, distribuição das idades dos pacientes para o sexo masculino e feminino.	29
Figura 15	Desire, nº de pacientes que tiveram ao menos uma ocorrência de ECAM e sua incidência segundo o sexo.....	30

Figura 16	Tela da ferramenta Orange (mostra o objeto: Attribute Statistics) que permite visualizar os valores mínimo, máximo, média e desvio padrão, mediana e os quartis dos atributos quantitativos (Ex.: IMC).....	33
Figura 17	Trecho da View com detalhe da criação do atributo [FUNC_RENAL].....	34
Figura 18	Classificação como tarefa para mapear um conjunto de atributos x com um atributo-classe y (PAM: Pressão arterial média em mmHg).	37
Figura 19	Classificador: a indução supervisionada do modelo e dedução da classe para novos exemplos.	38
Figura 20	Metodologia 5-Fold para testar um modelo induzido.	39
Figura 21	Medidas objetivas para avaliação de um classificador.	40
Figura 22	Espaço ROC definido por TPR <i>versus</i> FPR.	42
Figura 23	Exemplo de uma AD para as variáveis registradas em um ambulatório fictício.	43
Figura 24	Exemplo da representação gráfica da divisão de classes de uma árvore de decisão.	47
Figura 25	Subset exemplo das condições meteorológicas para jogo de tênis, transcrito do livro <i>Data Mining</i> de Ian H. Witten.	48
Figura 26	Quatro inícios de árvores possíveis para os exemplos da tabela da Figura 25.....	49
Figura 27	Medida de entropia para um problema de classificação binária com valores entre 0 e 1.....	51
Figura 28	Regras de decisão que pressupõem a existência de uma condição para a classificação.	56
Figura 29	Tela de entrada da ferramenta WEKA versão 3.7.11.	58
Figura 30	Tela de trabalho do WEKA para escolha do algoritmo classificador, mostrando relatório de saída.....	60
Figura 31	Orange, diagrama dos componentes do processo de classificação pela árvore de decisão e por regras.	62
Figura 32	Eventos ocorridos na fase hospitalar.	66

Figura 33	Distribuição dos eventos pós-hospitalar ao longo dos 13 anos.....	67
Figura 34	Exemplo da relação dos itens e as funções associadas à mineração dos dados.....	69
Figura 35	Exemplo da relação dos itens e a descrição do processamento.	69
Figura 36	Registro do primeiro ECAM do paciente na fase hospitalar e pós-hospitalar (13 anos).	70
Figura 37	Resultado da classe MACE no período 13 anos, para os modelos P-1, P-2 e P-3.....	71
Figura 38	Distribuição dos primeiros eventos ocorridos até 1 ANO após o procedimento index.....	73
Figura 39	Distribuição dos eventos quando se retirou os pacientes com IAM_NQ da fase hospitalar e inseriu um segundo evento destes pacientes, caso hovesse ocorrido dentro do 1º ANO.....	74
Figura 40	Modelos resultantes da classe MFP com período até o 1º ANO, para os modelos P-4, P-5 e P-6.....	75
Figura 41	Modelos resultantes da classe MACE_REFERENCIA dentro do 1º ANO, para os modelos P-7, P-8 e P-9.....	77
Figura 42	Resultado da classe MFP_FIRST com período até o 1º ano, para o modelo P-10.	79
Figura 43	Modelo P-10, apresentação da árvore de decisão no modo texto. Eventos ocorridos até 1 ANO.....	79
Figura 44	Modelo P-11, apresentação da árvore de decisão no modo texto. Eventos ocorridos 13 ANOS.	80
Figura 45	Trecho da árvore de decisão extraída do dataset Desire (modelo P-9).	83
Figura 46	Regras extraídas do modelo P-2 (13 ANOS).....	84
Figura 47	Regras extraídas do modelo P-5 (1 ANO).	85
Figura 48	Modelo P-8: Árvore de decisão no modo texto.	86
Figura 49	Principais regras extraídas do modelo P-8.	87

Figura 50	Primeira parte do modelo de classificação pelo atributo MFP, composto por 6 partes. (algoritmo CN2).	91
Figura 51	Terceira parte do modelo de classificação pelo atributo MFP, composto por 6 partes. (algoritmo CN2).	92
Figura 52	Quadro-resumo das avaliações para as regras do modelo P-2, que considera a ocorrência de eventos ao longo dos 13 anos de acompanhamento.	94
Figura 53	Quadro-resumo das avaliações para as regras do modelo P-5, que considera a ocorrência de eventos no 1º ano de acompanhamento.	95
Figura 54	Quadro-resumo das avaliações para as regras do modelo P-8, que considera a ocorrência de eventos no 1º ano de acompanhamento e exclui os infartos não Q da fase hospitalar.	95
Figura 55	Quadro-resumo das avaliações para as regras de classificação, que considera a ocorrência de eventos no 1º ano de acompanhamento.	95
Figura 56	Desire, atributos relativos aos antecedentes.	104
Figura 57	Desire, atributos relativos ao procedimento.	104
Figura 58	Desire, atributos relativos aos dados angiográficos.	105
Figura 59	Desire, atributos relativos aos stents implantados.	105
Figura 60	Desire, atributos relativos aos eventos MACE.	105
Figura 61	Desire, variáveis categóricas referentes aos antecedentes.	106
Figura 62	Desire, variáveis categóricas referentes ao procedimento. ...	107
Figura 63	Desire, variáveis categóricas referentes aos dados angiográficos e <i>stents</i>	108
Figura 64	Desire, variáveis categóricas referentes aos eventos.	109
Figura 65	Desire, variáveis numéricas onde o diâmetro e o comprimento estão em milímetros e o valor residual de estenose em porcentagem.	110
Figura 66	Modelo P-1: apresentação parcial (1/3) da árvore de decisão no modo texto.	111

Figura 67	Modelo P-1: apresentação parcial (2/3) da árvore de decisão no modo texto.....	112
Figura 68	Modelo P-1: apresentação parcial (3/3) da árvore de decisão no modo texto.....	113
Figura 69	Modelo P-2: Árvore de decisão no modo texto.	114
Figura 70	Modelo P-3: apresentação parcial (1/2) da árvore de decisão no modo texto.....	115
Figura 71	Modelo P-3: apresentação parcial (2/2) da árvore de decisão no modo texto.....	116
Figura 72	Modelo P-4: Árvore de decisão no modo texto para classe MFP de eventos ocorridos até 1 ANO.	117
Figura 73	Modelo P-5: Árvore de decisão no modo texto para classe MFP de eventos ocorridos até 1 ANO.	118
Figura 74	Modelo P-6: Árvore de decisão no modo texto.	118
Figura 75	Modelo P-8: Árvore de decisão no modo texto.	119
Figura 76	Desire, dataset numérico composto por 70 atributos, denominado FULL_NUM.	120
Figura 77	Desire, dataset categórico composto por 65 atributos, denominado FULL_CATEG.	121
Figura 78	Para a classe MACE os 20 atributos mais significativos do dataset numérico pelo critério GR.	122
Figura 79	Para a classe MFP os 20 atributos mais significativos do dataset numérico pelo critério GR.	122
Figura 80	Avaliação das regras extraídas do modelo P-2 pelo especialista nº 1.....	123
Figura 81	Avaliação das regras extraídas do modelo P-5 pelo especialista nº 1.....	124
Figura 82	Avaliação das regras extraídas do modelo P-8 pelo especialista nº 1.....	125
Figura 83	Avaliação das regras extraídas do modelo CN2 pelo especialista nº 1.....	126
Figura 84	Avaliação das regras extraídas do modelo P-2 pelo especialista nº 2.....	127

Figura 85	Avaliação das regras extraídas do modelo P-5 pelo especialista nº 2.....	128
Figura 86	Avaliação das regras extraídas do modelo P-8 pelo especialista nº 2.....	129
Figura 87	Avaliação das regras extraídas do modelo CN2 pelo especialista nº 2.....	130
Figura 88	Avaliação das regras extraídas do modelo P-2 pelo especialista nº 3.....	131
Figura 89	Avaliação das regras extraídas do modelo P-5 pelo especialista nº 3.....	132
Figura 90	Avaliação das regras extraídas do modelo P-8 pelo especialista nº 3.....	133
Figura 91	Avaliação das regras extraídas do modelo CN2 pelo especialista nº 3.....	134
Figura 92	Avaliação das regras extraídas do modelo P-2 pelo especialista nº 4.....	135
Figura 93	Avaliação das regras extraídas do modelo P-5 pelo especialista nº 4.....	136
Figura 94	Avaliação das regras extraídas do modelo P-8 pelo especialista nº 4.....	137
Figura 95	Avaliação das regras extraídas do modelo CN2 pelo especialista nº 4.....	138
Figura 96	Avaliação das regras extraídas do modelo P-2 pelo especialista nº 5.....	139
Figura 97	Avaliação das regras extraídas do modelo P-5 pelo especialista nº 5.....	140
Figura 98	Avaliação das regras extraídas do modelo P-8 pelo especialista nº 5.....	141
Figura 99	Avaliação das regras extraídas do modelo CN2 pelo especialista nº 5.....	142

RESUMO

Campos Neto CM. *Análise inteligente de dados em um banco de dados de procedimentos em cardiologia intervencionista* [Tese]. São Paulo: Instituto Dante Pazzanese de Cardiologia – Entidade Associada da Universidade de São Paulo; 2016.

O tema deste estudo abrange duas áreas do conhecimento: a Medicina e a Ciência da Computação. Consiste na aplicação do processo de descoberta de conhecimento em base de Dados (KDD – *Knowledge Discovery in Databases*), a um banco de dados real na área médica denominado Registro Desire. O Registro Desire é o registro mais longo da cardiologia intervencionista mundial, unicêntrico e acompanha por mais de 13 anos 6.377 pacientes revascularizados unicamente pelo implante de *stents farmacológicos*. O objetivo é criar por meio desta técnica um modelo que seja descritivo e classifique os pacientes quanto ao risco de ocorrência de eventos cardíacos adversos maiores e indesejáveis, e avaliar objetivamente seu desempenho. Posteriormente, apresentar as regras extraídas deste modelo aos usuários para avaliar o grau de novidade e de concordância do seu conteúdo com o conhecimento dos especialistas. Foram criados modelos simbólicos de classificação pelas técnicas da árvore de decisão e regras de classificação utilizando para a etapa de mineração de dados os algoritmos C4.5, Ripper e CN2, em que o atributo-classe foi a ocorrência ou não do evento cardíaco adverso. Por se tratar de uma classificação binária, os modelos foram avaliados objetivamente pelas métricas associadas à matriz de confusão como acurácia, sensibilidade, área sob a curva ROC e outras. O algoritmo de mineração processa automaticamente todos os atributos de cada paciente exaustivamente para identificar aqueles fortemente associados com o atributo-classe (evento cardíaco) e que irão compor as regras. Foram extraídas as principais regras destes modelos de modo indireto, por meio da árvore de decisão ou diretamente pela regra de classificação, que apresentaram as variáveis mais influentes e preditoras segundo o algoritmo de mineração. Os modelos permitiram entender melhor o domínio de aplicação, relacionando a influência de detalhes da rotina e as situações associadas ao procedimento médico. Pelo modelo, foi possível analisar as probabilidades da ocorrência e da não ocorrência de eventos em diversas situações. Os modelos induzidos seguiram uma lógica de interpretação dos dados e dos fatos com a participação do especialista do

domínio. Foram geradas 32 regras das quais três foram rejeitadas, 20 foram regras esperadas e sem novidade, e 9 foram consideradas regras não tão esperadas, mas que tiveram grau de concordância maior ou igual a 50%, o que as tornam candidatas à investigação para avaliar sua eventual importância. Tais modelos podem ser atualizados ao aplicar novamente o algoritmo de mineração ao banco com os dados mais recentes. O potencial dos modelos simbólicos e interpretáveis é grande na Medicina quando aliado à experiência do profissional, contribuindo para a Medicina baseada em evidência.

Descritores: Mineração de dados; C4.5; Árvore de decisão; Inteligência artificial; Cardiologia; KDD; Banco de dados; Doença das coronárias.

ABSTRACT

Campos Neto CM. Intelligent data analysis in an interventional cardiology procedures database [Thesis]. São Paulo: Institute Dante Pazzanese Cardiology - Associated Entity of the University of São Paulo; 2016.

The main subject of this study comprehends two areas of knowledge, the Medical and Computer Science areas. Its purpose is to apply the Knowledge Discovery Database-KDD to the DESIRE Registry, an actual Database in Medical area. The DESIRE Registry is the oldest world's registry in interventional cardiology, is unicentric, which has been following up 6.377 resvascularized patients for more than 13 years, solely with pharmacological stent implants. The goal is to create a model using this technique that is meaningful to classify patients as the risk of major adverse cardiac events (MACE) and objectively evaluate their performance. Later present rules drawn from this model to the users to assess the degree of novelty and compliance of their content with the knowledge of experts. Symbolic classification models were created using decision tree model, and classification rules using for data mining step the C4.5 algorithms, Ripper and CN2 where the class attribute is the presence or absence of a MACE. As the classification is binary, the models were objectively evaluated by metrics associated to the Confusion Matrix, such as accuracy, sensitivity, area under the ROC curve among others. The data mining algorithm automatically processes the attributes of each patient, who are thoroughly tested in order to identify the most predictive to the class attribute (MACE), whom the rules will be based on. Indirectly, using decision tree, or directly, using the classification rules, the main rules of these models were extracted to show the more predictable and influential variables according to the mining algorithm. The models allowed better understand the application range, creating a link between the influence of the routine details and situations related to the medical procedures. The model made possible to analyse the probability of occurrence or not of events in different situations. The induction of the models followed an interpretation of the data and facts with the participation of the domain expert. Were generated 32 rules of which only three were rejected, 20 of them were expected rules and without novelty and 9 were considered rules not as expected but with a degree of agreement higher or equal 50%, which became candidates for an investigation to assess their possible importance. These models can be easily updated by reapplying

the mining process to the database with the most recent data. There is a great potential of the interpretable symbolic models when they are associated with professional background, contributing to evidence-based medicine.

Descriptors: Data mining; *C4.5*; Decision tree; Cardiology; Artificial intelligence; KDD; Database; Coronary disease.

1 INTRODUÇÃO

1 INTRODUÇÃO

Segundo dados da Organização Mundial de Saúde, as doenças crônicas não transmissíveis (DCNT) são a maior causa de morte em todo o mundo. Apenas em 2008, elas foram responsáveis por 36 milhões de óbitos, ou seja, 63% do total global de mortes^{1,2}.

No grupo das DCNT, as doenças cardiovasculares (DCV) foram responsáveis por 48% das mortes, seguidas pelo câncer (21%), doenças respiratórias crônicas (12%) e diabetes (3%)^{1,2}.

Em nosso país, as DCV são responsáveis por quase 30% de todas as mortes registradas, o que coloca o Brasil entre os 10 países com maior índice por morte cardiovascular^{1,2}.

As DCV são aquelas que afetam o coração e as artérias, como o infarto agudo do miocárdio (IAM), acidente vascular cerebral (AVC), arritmias cardíacas, isquemias e anginas. A principal característica das DCV é a presença da aterosclerose, causada pelo acúmulo de placas de gordura nas artérias, como mostrado na Figura 1, cuja evolução ao longo dos anos termina por restringir ou até bloquear a passagem do sangue e o aporte de oxigênio aos tecidos^{1,2}.

Fonte: <http://hmsportugal.wordpress.com/2011/10/24/o-que-e-a-angina-de-peito>

Figura 1 - Aterosclerose, o efeito obstrutivo da doença nas artérias com redução do fluxo sanguíneo que pode levar a um quadro agudo de obstrução total (infarto agudo do miocárdio).

No conjunto das DCV, o IAM vem preocupando a sociedade e as autoridades de saúde pública, pelo impacto causado na taxa de mortalidade e no número de internações em todo o país, sendo as maiores causadoras de atendimento nas unidades de emergência³.

O único tratamento para a revascularização do miocárdio de pacientes que haviam sofrido infarto era a técnica cirúrgica com o implante de pontes venosas e/ou arteriais. Em setembro de 1977, o Dr. Andréas R. Gruentzig realiza, na Suíça, a primeira dilatação de uma obstrução significativa nas artérias coronárias, utilizando pequenos balões por acesso percutâneo, sem incisão cirúrgica⁴. A partir de então, nasce a Intervenção Coronária Percutânea (ICP) como alternativa para a desobstrução das lesões coronarianas, porém o índice de reestenose ainda era alto ocorrendo em 20% a 40% das desobstruções inicialmente dilatadas com sucesso^{5,6}.

A solução para boa parte dos problemas relativos à angioplastia com balão vem com os stents metálicos, quando, em 1993, dois importantes estudos comparando o *stent* Palmaz-Schatz com a AC estabeleceram que o implante eletivo do *stent* era a forma de intervenção mais adequada, pois minimizava os problemas relativos à AC, como o remodelamento negativo do vaso pós AC, as oclusões agudas, as dissecções coronarianas e as cirurgias de emergência⁷.

O tratamento coronário percutâneo atualmente emprega dois tipos de *stents*: os convencionais, também conhecidos como *Stent Metálico Simples*, e os *Stents Farmacológicos* (SF).

O *Stent Metálico Simples* é uma prótese de aço inoxidável autoexpansível, como mostra a Figura 2, que é implantado no local da lesão da artéria coronária para reestabelecimento do fluxo sanguíneo.

Figura 2 - Imagem de um stent convencional antes e após sua expansão (*Bare metal stent*).

Os *Stents* Farmacológicos (SF) possuem uma plataforma equivalente, porém são revestidos com fármacos que têm a função de inibir localmente a reação inflamatória devido à presença do dispositivo metálico. O processo de liberação da droga no organismo ocorre dentro de determinado período, reduzindo significativamente a reestenose pós-procedimento (Figura 3)⁸.

Os benefícios trazidos pelo SF foi um agente de mudança no uso dos *stents*, pois ampliou a possibilidade do tratamento de lesões antes consideradas controversas devido à elevada incidência de reestenose, permitindo, assim, o tratamento de pacientes cada vez mais complexos⁸.

Fonte: <http://www.biosensors.com/intl/products-technology-biomatrix>

Figura 3 - Ilustração de um stent farmacológico Biomatrix (Fármaco: BiolimusA9™).

1.1 O processo de aprendizagem e a criação de um modelo

Em Ciências da Computação, a Inteligência Artificial (IA) é a área que se preocupa em desenvolver sistemas computacionais inteligentes, isto é, sistemas que exibem características, as quais associamos com a inteligência no comportamento humano, como compreensão da linguagem, aprendizado, resolução de problemas, etc.⁹.

O Aprendizado de Máquina (AM) é uma subárea da IA dedicado ao desenvolvimento de algoritmos e técnicas que permitem ao computador aprender e que se preocupa com o raciocínio indutivo, extraindo regras e padrões de grandes conjuntos de dados (programas que aprendem por experiência).

Aprender a classificar os objetos como pertencentes a um determinado conjunto ou a uma determinada classe é uma característica da inteligência que tem despertado o interesse dos pesquisadores da área de Ciências da Computação. Identificar as características comuns de um conjunto de objetos que são representativos de sua classe é o foco principal da classificação quando realizada por uma pessoa ou por um algoritmo.

Para introduzir o problema da aprendizagem, considere o departamento de crédito de um banco que recebe milhares de solicitações de empréstimo por dia e precisa automatizar seu processo de análise de crédito.

Assim como em outras situações do mundo real, o banco não tem a fórmula mágica que determine se o crédito deva ser concedido ou não. Por outro lado, ele tem uma grande quantidade de registros sobre as análises de crédito já realizado, logo, ele pode usar este histórico de informações para buscar descobrir uma boa fórmula de aprovação de crédito.

Estes dados são a base para a construção de um modelo de aprovação de crédito que poderá ser usado para futuros clientes, por meio de um aprendizado supervisionado. O diagrama de blocos apresentado na Figura 4 formaliza este processo de aprendizado de máquina.

No primeiro bloco, tem-se a função-alvo $f: X \rightarrow Y$ que é desconhecida e representa a fórmula ideal de aprovação de crédito (X representa todas as possíveis entradas das informações dos clientes e Y representa todas as possíveis saídas, que, para o caso do crédito, são apenas duas: aprovado ou reprovado).

No segundo bloco, tem-se o conjunto de dados D do tipo entra-saída representados pela notação $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)$ que correspondem aos dados históricos dos clientes e suas respectivas decisões sobre a aprovação do crédito.

Finalmente, o algoritmo de aprendizado seleciona a função g de um conjunto de fórmulas candidatas H (conjunto de hipóteses) como a que mais se aproxima de f .

Logo, quando um novo cliente solicitar crédito, o banco irá apoiar sua decisão em g (hipótese produzida pelo algoritmo de aprendizado) e não em f (a função-alvo ideal que permanece desconhecida).

Fonte: Adaptada do livro Learning from Data; Abu Mustafa.

Figura 4 - Configuração básica de um problema de aprendizagem.

A essência do problema de aprendizado se resume ao fato de que existe uma função-alvo a ser aprendida e que é desconhecida. Há um

conjunto de dados gerados por esta função (dados históricos) que serão usados pelo algoritmo de aprendizado para encontrar, dentre um conjunto de funções (hipóteses), a que melhor se aproxima do alvo.

Atualmente, encontramos-nos superexpostos a uma quantidade enorme de dados e informações, e o problema que se apresenta é a incapacidade do homem em assimilar tanta informação, criando, assim, uma lacuna na análise, compreensão e extração de novos conhecimentos.

Tal necessidade levou ao aparecimento de uma metodologia que se utiliza de tecnologias de extração automática de conhecimento, denominada “Descoberta de Conhecimento em Base de Dados”, conhecida pela sigla inglesa KDD (*Knowledge Discovery in Databases*), a qual foi apresentada, em 1996, por Fayyad¹⁰, e tornou-se uma área de pesquisa amplamente estudada, e que se utiliza de uma nova geração de teorias computacionais e ferramentas necessárias para auxiliar na extração de informações úteis.

Segundo a definição dada por Fayyad¹⁰: “KDD - Descoberta de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutido nos dados”.

Rezende¹¹ destaca o significado individual de cada termo como:

- Dados: Conjunto de fatos em um repositório;
- Padrões: Refere-se a alguma abstração de um subconjunto de dados numa linguagem descritiva de conceitos;
- Processo: Composto por etapas bem definidas como preparo dos dados, busca por padrões e avaliação do conhecimento adquirido;
- Válidos: As abstrações realizadas devem possuir algum grau de certeza; - Novos: Devem acrescentar novas informações sobre os dados;
- Úteis: O conhecimento adquirido deve agregar valor e ser incorporado à rotina;
- Compreensíveis: Serem descritos em uma linguagem que possa ser compreendida pelos usuários.

Importante destacar que este processo é centrado na interação entre as diversas classes de usuários, ou seja, especialistas com conhecimento em diferentes domínios, e seu sucesso depende, em parte, dessa interação.

Fayyad descreve o KDD como sendo um processo interativo, iterativo e composto por fases bem determinadas, sendo a parte central e mais importante conhecida como Mineração de Dados (MD) que, muitas vezes, empresta o nome para o processo como um todo.

Quanto às suas etapas, o processo de KDD pode ser dividido em três fases:

- Pré-processamento;
- Extração de Padrões e
- Pós-processamento, conforme apresentado na Figura 5.

Figura 5 - Etapas operacionais do processo de KDD.

O processo de KDD não é trivial já que alguma técnica de busca ou inferência é envolvida, ou seja, não é um processo de computação direta. Embora, inicialmente, os passos devam seguir a sequência apresentada, este é um processo iterativo, o que implica em reentradas em diversos pontos na busca pelo melhor modelo. Cada etapa tem sua importância e depende do resultado da sua precedente, que terá impacto no resultado final.

A primeira etapa, a de Pré-processamento, compreende a seleção do conjunto de dados contendo todas as possíveis variáveis e os registros que farão parte da análise, que pode implicar no desenvolvimento de um programa específico para atender as diversas peculiaridades da carga dos

dados. Outra atividade é a limpeza dos dados redundantes e inconsistentes, a recuperação de dados incompletos e a identificação dos dados discrepantes. Ocorre, também, a redução das variáveis e a transformação de seus valores, como a normalização, categorização e combinação de outros.

O produto final é um arquivo no formato tabular cujas colunas representam as variáveis ou os atributos e as linhas os registros de casos históricos adequadamente preparados para serem processados pelo algoritmo de mineração de dados. A participação do especialista de domínio para o qual se está desenvolvendo o trabalho é fundamental nesta etapa.

A etapa de Mineração de Dados é onde se define a tarefa (o que se espera encontrar nos dados) e os métodos (algoritmos) que serão utilizados de acordo com os objetivos do estudo a fim de se obter uma resposta ao problema, ou seja, descobrir os padrões que interessam.

Na fase de pós-processamento é que são avaliados de modo objetivo e subjetivo os resultados gerados, identificando as regras potencialmente interessantes para, posteriormente, serem validadas com o especialista de domínio e com o usuário final.

Em relação a sua origem, pode-se dizer que a Mineração de Dados se baseia em ideias como amostragem, teste de hipóteses advindas da estatística; algoritmos de busca, técnicas e modelagem e teorias de aprendizagem da inteligência artificial; aprendizado de máquina e reconhecimento de padrões conforme apresentado na Figura 6. Há outras áreas de apoio que desempenham importante papel como sistemas de banco de dados (*armazenamento, indexação e processamento de consultas*), computação paralela (*desempenho computacional no processamento de quantidades massivas de dados*) e distribuída (para contornar o grande volume, quando os dados não podem ser processados localmente)^{9,12}.

Figura 6 - Diagrama que mostra o relacionamento da Mineração de Dados com várias disciplinas.

As tarefas de MD subdividem-se em duas categorias: **Preditivas** e **Descritivas**.

Enquanto as **Preditivas** procuram adquirir conhecimento a partir de um conjunto de dados históricos de treinamento, para, depois, fazer a predição para novas amostras, as **Descritivas** buscam identificar padrões de comportamento comuns que sintetizem as relações subjacentes entre os dados. A Figura 7 mostra a divisão das tarefas de MD em função da atividade, em que a Classificação e Regressão pertencem à atividade Preditiva, e a Associação, *Clustering*, Sumarização, e outras pertencem à atividade Descritiva¹¹.

Contudo, não existe uma separação clara entre estes dois modelos, pois alguns modelos preditivos podem ser considerados descritivos, uma vez que eles podem ser compreensíveis e interpretáveis^{10,13}.

Fonte: Reproduzida do livro Sistemas Inteligentes; Rezende SO.

Figura 7 - Tarefas de mineração de dados.

A Classificação objetiva encontrar modelos que descrevam classes entre os dados apresentados (denominado conjunto de treinamento), os quais devem estar corretamente e previamente categorizados, ou seja, trata-se de um aprendizado supervisionado. O modelo construído pode ser utilizado para prever a classe de objetos que ainda não foram classificados.

A Regressão é similar à Classificação, porém usada quando o atributo-classe apresenta valores numéricos (contínuos). Exemplo: estimar a pressão arterial do paciente conhecendo sua idade, índice de massa corporal e sexo.

Clustering visa identificar e aproximar os registros similares em subconjuntos ou *clusters* de modo que os elementos de um *cluster* compartilham de propriedades comuns distinguindo-os de outros elementos pertencentes a outros *clusters*. O objetivo é maximizar as similaridades *intracluster* e minimizar as similaridades *intercluster*. Exemplo: segmentação de mercado para determinado tipo de produto. Diferentemente da Classificação, esta tarefa não necessita que os registros estejam previamente categorizados, ou seja, trata-se de um aprendizado não supervisionado.

Regra de Associação consiste em identificar quais atributos estão relacionados, descrevendo associações entre as variáveis no mesmo item. Caracteriza o quanto a presença de um conjunto de atributos nos registros de uma base de dados implica na presença de algum outro conjunto de atributos nos mesmos registros, apresentam-se na forma: SE atributo X ENTÃO atributo Y.

A Sumarização procura encontrar uma descrição compacta para um conjunto de dados. Exemplo: média e desvio padrão.

1.2 Histórico

A “Descoberta de Conhecimento em Bases de Dados – KDD” refere-se ao uso de métodos e às ferramentas para analisar um grande volume de dados com o objetivo de auxiliar na extração de relações de interesse pelos especialistas do domínio.

Sua aplicação já é realidade em diversas áreas, embora na Medicina tenha sido relativamente limitada. Nos últimos anos, o termo “*mineração de dados*”, uma das etapas da Descoberta de Conhecimento – KDD tem aparecido com mais frequência na literatura em aplicações médicas¹⁴.

A mineração de dados na clínica médica pode ser vista como uma nova ferramenta de suporte ao diagnóstico, ao prognóstico e à conduta do paciente por meio da construção de modelos descritivos e preditivos por meio das informações disponíveis em seus registros (bases de dados)¹⁴.

Na fase da revisão bibliográfica, foram encontrados diversos trabalhos na literatura nacional referentes ao uso da mineração de dados em Medicina que compreende um período iniciado nos anos 2000.

Ferro e col.¹⁵ relatam, em 2001, dois estudos de caso, o primeiro de um banco de dados natural (n=417) para prever doenças coronarianas (UCI Irvine) e outra base de dados real (n=385) de uma clínica de reprodução humana. Em ambos, mostram como o processo de KDD pode ser aplicado para a Medicina, sendo que, na fase da mineração dos dados,

foi empregado o método de aprendizado de máquina simbólico indutivo supervisionado (árvore de decisão pelo algoritmo C4.5). Buscava-se a extração de novos conhecimentos possivelmente interessantes e a construção de classificadores para apoio à tomada de decisão. Segundo os autores, a aplicação do KDD para pequenas bases de dados é chamada de Análise Inteligente dos Dados, como foi o caso deste artigo.

Steiner e cols.¹⁶, em 2006, propõem aplicar a MD para induzir modelos para a classificação de 118 pacientes com o objetivo de distinguir entre câncer e cálculo como a causa de obstrução para a bile no fígado, eles propõem uma análise exploratória prévia antes da fase de mineração de dados comparando cinco técnicas entre redes neurais, regressão logística e árvore de decisão. Em suas conclusões, destacam que, dentre todas as técnicas adotadas, apenas a árvore de decisão deixa claro ao usuário quais são os atributos que estão discriminando os padrões (compreensibilidade) e de que forma está ocorrendo.

Cavalcante¹⁷, em 2009, relata os resultados da mineração de dados por regra de associação de 85 pacientes portadores de DAC, no hospital Monte Sinai, por meio de um questionário fechado com os fatores de risco descritos pelo *Framingham Heart Study*, demonstrando que, pela análise do estilo de vida destes pacientes, há uma convergência aos percentuais já homologados.

Maciel e cols., em 2015¹⁸, aplicam a árvore de decisão (Weka-J4.8) para a triagem de pacientes (n=10.997) em uma unidade de pronto atendimento quanto ao risco de vida (eletivo, baixo, médio e alto), conseguindo um modelo com 59,33% de acurácia. Segundo o autor, apesar de baixa, a acurácia não foi a única métrica para o sucesso, pois o fato do modelo ser interpretável permitiu o entendimento de que atributo é mais determinante para cada classe de risco, assim como os intervalos de valores.

Outros trabalhos relatam o uso da mineração de dados na área da saúde pública para identificar padrões de características materno-fetais na predição da mortalidade infantil¹⁹ ou com o objetivo de extrair informações

epidemiológicas para a classificação etiológica de hepatites virais²⁰, outro trabalho que usa a mineração de dados para a classificação automática de uma carteira de beneficiários de um grupo de saúde buscando uma ação preventiva referente ao Diabetes Mellitus Tipo II²¹.

Entre os artigos internacionais, é possível perceber que a aplicação da mineração em bases de dados médicas, para testes e avaliações de modelos, despertou o interesse da área de Ciências da Computação, por meio da inteligência artificial e, mais especificamente, do aprendizado de máquina que trata do desenvolvimento de algoritmos voltados à aprendizagem e ao reconhecimento de padrões. Nota-se, ainda, o interesse despertado pelo tema em países que fogem ao eixo convencional da América do Norte e Europa ocidental, como Índia, Irã, Eslovênia e Chipre.

Jaisankar & Gayathri da School of Computing Science and Engineering do Vellore Institute of Technology India²² apresenta uma ampla revisão de artigos, comentando 47 trabalhos publicados no período entre 2005 e 2013 sobre o uso de diversas técnicas de IA voltadas ao diagnóstico e prognóstico das doenças cardíacas. Este artigo apresenta um enfoque técnico, voltado aos diferentes algoritmos propostos, que, para serem avaliados e comparados, utilizam, em sua maioria, bases de dados de referência disponíveis no site da UCI (*University of California, Irvine – machine learning repository*).

Trabalhos nesta área empregam a metodologia da mineração de dados para a extração de regras e/ou para classificação por diferentes técnicas cujos modelos resultantes são avaliados e comparados por meio de medidas objetivas oriundas da matriz de confusão, como a acurácia, sensibilidade, especificidade, e pelo valor da área sob a curva ROC (*receiver operating characteristic*).

Para a tarefa de mineração de dados descritiva utilizando as Regras de Associação, Karaolis e cols., em seu artigo “*Association Rule Analysis for the Assessment of the Risk of Coronary Heart Disease Events*”²³, empregam a mineração de dados para avaliar os fatores de risco para a doença cardíaca coronária seguindo um protocolo pré-estabelecido que incluía os

pacientes que deram entrada no hospital Paphos de Nicosia, Chipre, com, pelo menos, um dos seguintes eventos em sua história clínica: Infarto agudo do miocárdio (IAM), Intervenção percutânea coronária (IP) ou Cirurgia de revascularização do miocárdio com ponte de safena (CIR).

Com um conjunto de dados com 265 casos de pacientes que tiveram IAM, 160 para IP e 152 para CIR, as variáveis independentes foram compostas pelos fatores clínicos, como: idade, sexo, fumante, pressão sistólica e diastólica, histórico de hipertensão, história familiar, diabetes e pelos fatores bioquímicos: Colesterol total, HDL, LDL, triglicérides e glicose.

Deste trabalho, são apresentadas três regras por grupo, resultado da mineração por regras de associação, utilizando o clássico algoritmo “Apriori” da ferramenta Weka:

Grupo com evento de IAM:

- 1- A maioria dos pacientes com IAM é do sexo masculino;
- 2- 72% dos homens com alta taxa de glicose são fumantes;
- 3- 70% dos pacientes com níveis de HDL anormal tem histórico de hipertensão.

Grupo com evento de IP:

- 1- 72% dos pacientes hipertensos com histórico familiar de doença coronária são fumantes;
- 2- A maioria dos fumantes com histórico familiar e hipertensos é do sexo masculino;
- 3- 68% dos pacientes na faixa de 61 a 70 anos têm histórico de hipertensão.

Grupo com evento de CIR:

- 1- O sexo não influencia se o paciente for fumante com alto nível de glicose e hipertenso;
- 2- A maioria dos pacientes fumantes com histórico de hipertensão e que tem a pressão alta é homem;
- 3- Os pacientes da regra acima cuja taxa de colesterol e LDL está dentro da normalidade tem uma redução de 13% no evento.

Segundo o autor, os novos pacientes podem ter seus fatores de risco modificáveis monitorados e reduzidos, o que irá contribuir para a redução dos eventos cardíacos.

No mesmo hospital, Karaolis²⁴, em *Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees*, avaliou um grupo maior de pacientes coronarianos segundo o mesmo protocolo, utilizando Regras de Classificação e Árvore de Decisão.

Investigou como a mineração de dados pela técnica da árvore de decisão pode ajudar na avaliação de risco da doença coronariana (DC). O objetivo foi identificar os fatores de risco mais importantes com base nas regras de classificação extraídas.

A análise foi realizada empregando tarefa de classificação pela técnica da Árvore de Decisão pelo algoritmo C4.5, que foi escolhida pelo autor por ser uma técnica amplamente utilizada, e por gerar resultados confiáveis e por ser um classificador com grande acurácia.

Os resultados dos modelos dos três grupos foram avaliados pelas medidas de acurácia, sensibilidade e especificidade, com máxima acurácia de 66% para IAM, 75% para IP e 75% para CIR, embora diferentes fatores de risco se destacassem em cada um dos grupos, os fatores mais importantes extraídos pela análise das regras de classificação foram sexo, idade, hábito de fumar, pressão sanguínea e colesterol.

Dentre as regras extraídas para o grupo CIR: cita que a cirurgia ocorre em pessoas com idade entre 51 e 60 anos com histórico de diabetes e que o histórico familiar não é um fator de risco importante dentro do grupo.

Dentro do grupo IP: entre os diabéticos, o número de intervenções percutâneas aumenta com a idade. Para o grupo IAM: a porcentagem dos pacientes com idade entre 51 e 60 anos, não fumante com histórico de hipertensão e com infarto é, aproximadamente, a mesma dos que eram fumantes e não tiveram infarto.

Rajalaxmi e Abdullah²⁵ empregaram a técnica de classificação para testar um modelo de mineração de dados usando o algoritmo conhecido como *Random Forest Classifier* (RFC) para exploração e predição dos registros de pacientes com doença coronariana provenientes de um repositório de dados para aprendizado de máquina da UCI (Hospedado no site da *UCI Machine Learning Dataset* da Universidade da Califórnia), com o objetivo de avaliar sua acurácia comparativamente a outras técnicas.

Os autores compararam seus resultados com os de A. Reena e col.²⁶ que utilizaram a técnica de árvore de decisão sobre o mesmo repositório de dados e os mesmos atributos fornecidos por Cleveland, mostrando que o desempenho do RFC foi melhor.

Ainda em cima da mesma base de dados, de Cleveland, Anbrasi e cols.²⁷ apresentaram um trabalho sobre redução de atributos, com a utilização de um algoritmo genético para determinar, dentre os 13 atributos analisados, quais seriam os 6 atributos mais representativos para o diagnóstico da doença coronariana, com a intenção de poder reduzir, assim, o número de informações e, conseqüentemente, o número de exames que o paciente deve se submeter.

Sivanesh e cols.²⁸, do departamento de Ciência da Computação e Engenharia da *Anna University os Technology*, em *Decision Support for Medical Diagnosis Using Data Mining*, observa que o diagnóstico médico por natureza é complexo e advém de um processo cognitivo difuso, portanto, métodos computacionais classificadores como a árvore de decisão tem grande potencial para auxiliar nos sistemas médicos de suporte à decisão.

Alizadehsani e cols.²⁹ utilizam a mineração de dados em uma base de 303 pacientes que foram atendidos em um hospital de Teerã com dor torácica, contendo 37 atributos indicativos de doença arterial coronariana.

O objetivo de seu trabalho foi descobrir um meio de identificar a artéria lesionada (LAD, LCX e RCA) quando não há evidências eletrocardiográficas suficientes e somente baseado em 16 atributos demográficos, 14 atributos relativos a sintoma e exame clínico, e 7 características eletrocardiográficas. O atributo “classe” foi o resultado do cateterismo que avaliou angiograficamente o estreitamento na artéria. (DAC se obstrução maior do que 50% ou normal caso menor).

Três algoritmos de classificação foram usados: C4.5 que, segundo o autor, é um dos melhores algoritmos para árvore de decisão, o Naive Bayes, que é um dos métodos Bayesianos que pressupõe independência entre os atributos, e o KNN (K-Nearest Neighbor) algoritmo de classificação simples de aprendizado supervisionado.

Para comparar os modelos gerados, o autor também utilizou o critério da acurácia, sensibilidade e especificidade por serem os mais utilizados na área médica.

Os resultados confirmaram a árvore de decisão com o C4.5 como sendo o melhor método de classificação dentre os três, apresentando uma acurácia de 74,2% para a artéria LAD, 63,76% para a LCX e 68,33% para a RCA. O efeito dos 37 atributos na estenose da artéria foi analisado pelo autor, que constatou o maior peso da variável dor torácica típica e idade na estenose destes vasos.

Parte considerável dos artigos encontrados na pesquisa pelos descritores Data Mining, C4.5, Árvore de Decisão, Cardiologia, KDD, Doença Coronária, Fatores de Risco Cardiovascular e Modelos Preditivos utilizavam o mesmo repositório de dados (*Heart Disease Data Set*) pertencente ao Centro de Aprendizado de Máquina e Sistemas Inteligentes da Universidade da Califórnia – UCI, com isto, diferentes algoritmos puderam ser comparados, principalmente em função da acurácia, pois utilizavam a mesma base de dados.

O *UCI Machine Learning Repository*³⁰ foi criado em 1987 por alunos de pós-graduação da Universidade da Califórnia e consiste em uma coleção de banco de dados que são usados pela comunidade de aprendizado de máquina para análise empírica dos algoritmos de mineração de dados, e o *Heart Disease Data Set* é um deles.

Pelo levantamento bibliográfico realizado, nota-se que o emprego da metodologia de KDD em banco de dados reais na área médica tem sido pouco explorado e, se forem consideradas as diversas subespecialidades, nota-se que ainda há muito por fazer.

Para a revascularização do miocárdio, o tratamento pela Intervenção Coronária Percutânea (ICP), uma técnica alternativa à cirurgia cardíaca, estabeleceu-se como um tratamento seguro, eficaz e com tempo de internação dos pacientes muito reduzidos. Cada vez mais utilizada, hoje com os stents farmacológicos apresenta índices de reestenose abaixo dos 5%.

A busca contínua das equipes de intervencionistas na identificação das causas dos eventos cardíacos relacionados aos dispositivos implantados e no aprimoramento desta técnica foram os fatores que motivaram o emprego do processo de KDD, em uma base de dados real, em um único hospital, com pacientes tratados unicamente com stents farmacológicos.

Desenvolver um modelo simbólico, de fácil interpretação para os médicos intervencionistas, que descreva a inter-relação entre as variáveis de maior influência, armazenadas no banco de dados, através da mineração de dados é o tema central deste trabalho.

2 OBJETIVOS

2 OBJETIVOS

Considerando o potencial do processo de descoberta do conhecimento (KDD) e sua consolidação em diversas áreas, reconhecidamente uma metodologia moderna e eficaz na análise de grandes volumes de dados.

Considerando a existência de um banco de dados real de intervenção percutânea para revascularização coronária de pacientes tratados unicamente com o implante de stents farmacológicos e acompanhados por mais de 13 anos pela mesma equipe de especialistas (Registro Desire); e levando em conta as considerações apresentadas, tendo em vista que a área médica ainda foi pouco explorada pela metodologia do KDD e que o banco de dados do Registro Desire já foi muito estudado neste período, sempre utilizando os conhecimentos estatísticos como suporte, formula-se a seguinte hipótese:

“É possível extrair do Banco de Dados do Desire um modelo que descreva a inter-relação entre as variáveis de maior influência no desfecho de um evento cardíaco adverso e classifique estes pacientes quanto ao risco de sofrer este evento”.

Uma alternativa de estudo para o Registro Desire, além da metodologia usualmente empregada de análise estatística, tal modelo, por ser descritivo, poderia reter o conhecimento adquirido, e ser útil para o treinamento de futuros médicos e como auxiliar no apoio à decisão médica.

2.1 Objetivo principal

O objetivo principal deste trabalho é aplicar a metodologia de KDD ao banco de dados do Registro Desire e, pela técnica da mineração de dados, construir modelos simbólicos de classificação supervisionados, obter um modelo descritivo que classifique os pacientes pelo risco da ocorrência de

eventos cardíacos adversos maiores e indesejáveis (ECAM), e avaliar seu desempenho objetivamente pelas métricas pertinentes.

2.2 Objetivo secundário

O objetivo secundário é extrair as principais regras destes modelos e realizar uma investigação junto ao especialista de domínio, para avaliar, segundo seu conhecimento, o grau de novidade das regras geradas.

3 Materiais e Métodos

3 MATERIAIS E MÉTODOS

Este estudo utilizou como fonte principal de informação o banco de dados do Registro Desire⁸, que é um registro unicêntrico, prospectivo e acompanha os pacientes submetidos à revascularização coronariana unicamente pelo implante de stents farmacológicos e que pertence ao Serviço de Cardiologia Intervencionista liderado pelo Prof. Eduardo Sousa no Hospital do Coração (HCor) da Associação do Sanatório Sírio.

A equipe médica do Registro caracteriza-se pela grande experiência na área da intervenção percutânea, tanto pelo pioneirismo como pelo número de casos tratados.

Este trabalho foi submetido ao Comitê de Ética e Pesquisa do Instituto Dante Pazzanese de Cardiologia, conforme Protocolo nº 4490.

3.1 Casuística

O Registro Desire foi criado em 2002 para avaliar o desempenho dos stents farmacológicos utilizados na revascularização percutânea dos pacientes com indicação de ICP, tratados neste centro hospitalar, com o objetivo de avaliar a evolução clínica tardia destes pacientes⁸.

Consiste de um registro com inclusão consecutiva de pacientes, cujos critérios de inclusão e exclusão estão descritos a seguir.

Critério de inclusão: O Registro incluiu todos os pacientes com, pelo menos, uma lesão com estenose maior ou igual a 50%, anatomia favorável para a ICP e que tivessem subvenção de planos de Saúde Suplementar ou de recursos próprios^{8,31,32}.

Critério de exclusão: Pacientes com implantes mistos (stents farmacológicos e não farmacológicos), indicação para cirurgia de emergência ou contraindicação para o uso de qualquer um dos antiplaquetários previstos pelo protocolo medicamentoso^{8,31,32}.

Todas as informações coletadas do paciente nas diferentes fases do tratamento e, posteriormente, no acompanhamento tardio, são registradas, primeiramente, em fichas específicas que irão compor o prontuário físico. Após a alta, as fichas são sistematicamente revisadas por um cardiologista, responsável para a validação dos dados, e, posteriormente, encaminhada para a digitação.

O cuidado com a qualidade da informação foi uma preocupação da equipe do Registro Desire que, desde o seu início, armazenou os dados em banco de dados (Microsoft®-SQLServer) por meio de um sistema, desenvolvido especificamente para este fim, composto por diferentes módulos, como o de cadastro com funcionalidades para verificar os limites permitidos entre as diferentes variáveis quantitativas, apresentar, no formato de lista, os conteúdos das variáveis qualitativas, alertar sobre a obrigatoriedade de determinados dados e realizar uma verificação de consistência ao final do preenchimento. Outros módulos do sistema permitem o seguimento tardio dos pacientes, a emissão de relatórios com aplicação de filtros, a exportação dos dados para serem analisados estatisticamente, etc.

Os pacientes foram tratados com os seguintes stents farmacológicos (SF) de primeira¹ e segunda² geração: Cypher¹, Taxus¹, Endeavor², Resolute², Biomatrix², Xcience V² e Promus², não havendo limitação quanto ao número de stents implantados nem quanto ao número de lesões tratadas. O uso de agentes antiplaquetários seguiu um protocolo antitrombótico com a administração de ácido acetil salicílico e clopidogrel antes e após o procedimento^{8,31,32}.

Os pacientes foram acompanhados posteriormente (por contato telefônico ou pela oportunidade da visita ao hospital) para que fossem registradas as ocorrências de eventos cardíacos adversos maiores ECAM,

como o infarto agudo do miocárdio (com onda Q e sem onda Q), revascularização de uma lesão previamente tratada ou de uma nova lesão e morte por causa cardíaca, com um seguimento de 98,2%^{31,32}.

Para este estudo, o banco de dados congelado resultou numa base composta por 6.377 pacientes, 9.575 vasos, 10.417 lesões tratadas e 11.146 stents implantados, com uma taxa de 1,63 lesões por paciente e 1,75 stents por paciente, conforme apresentado na Figura 8.

Características gerais		
Pacientes		6.377
Territórios		8.633
Vasos		9.575
Lesões		10.417
Stents		11.146
Lesões/paciente		1,63
Stents/paciente		1,75

Figura 8 - Desire, total de pacientes, territórios, vasos, lesões tratadas e stents implantados. N° de lesões tratadas por paciente e n° de stents implantados por paciente.

Pelas Figuras 9 e 10, são apresentadas as informações sobre o perfil da população em estudo, com suas características gerais, clínicas e angiográficas.

Características clínicas basais			
Variável		n=6.377	%
Média das idades , anos		64,66 ± 11,19	
Sexo masculino , %		4971	77,95
Diabetes melito , %		1817	28,49
em uso de insulina		329	5,16
Hipertensão , %		4875	76,45
Dislipidemia , %		4062	63,70
Tabagismo atual , %		548	8,59
História familiar de DAC , %		2109	33,07
IAM prévio , %		1301	20,42
Revasc prévia , %		2557	40,10
Insuficiência renal , %		366	5,74
Apresentação clínica , %			
Angina estável		1238	19,41
Angina instável		1594	25,00
IAM		1126	17,66
Isquemia silenciosa		2369	37,93

Figura 9 - Características clínicas dos pacientes tratados.

Características angiográficas			
Variável			%
Território miocárdico acometido (nº de territórios = 8.633), %		n=8.633	
DA		3567	41,32
Cx		2541	29,43
CD		2373	27,49
TCE		152	1,76
Características da lesão (nº de lesões = 10.417), %		n=10.417	
Localizadas em ponte de safena		511	4,91
Calcificação moderada ou grave		3878	37,23
Excentricidade		4351	41,77
Bifurcação		515	4,94
Lesão tipo B2 ou C		6246	66,39
Nº de stents implantados por lesão , %			
1		9462	90,83
2		833	8,00
3		89	0,85
4		9	0,07

Figura 10 - Características angiográficas dos pacientes do Registro Desire.

As Figuras a seguir apresentam alguns detalhes, com agrupamento pelo gênero, desta população em que o sexo masculino corresponde a, aproximadamente, 78% dos pacientes.

A Figura 11 mostra que 44% das mulheres tinham histórico cardiovascular contra 50% no grupo dos homens e, quanto ao índice de massa corpórea 70% dos pacientes, foram considerados normais ou acima do peso para ambos os sexos.

Figura 11 - Desire, pacientes com histórico cardiovascular e sua classificação segundo o IMC agrupado pelo sexo.

Dentre os conhecidos fatores de risco, o diabetes esteve presente em 31% das mulheres e em 28% dos homens (Figura 12), a hipertensão arterial entre as mulheres foi de 83% contra 74% entre os homens, e a percentagem de hiperlipidemia entre os homens e mulheres foram equivalentes e próximas a 64% (Figura 13).

Figura 12 - Desire, incidência do Diabetes segundo o sexo.

Figura 13 - Desire, incidência de Hipertensão e Hiperlipidemia nos pacientes segundo o sexo.

Quanto à idade, a população em estudo apresentou um valor médio de $64,66 \pm 11,19$ anos, em que se pode verificar que as mulheres são mais idosas, com uma média de idade de $69 \text{ anos} \pm 11,16$ e apresentam uma distribuição diferente dos homens, os quais se aproximam mais de uma distribuição normal. Para os homens, a média de idade foi de $63,40 \pm 10,88$ anos (Figura 14).

Figura 14 - Desire, distribuição das idades dos pacientes para o sexo masculino e feminino.

Em relação à ocorrência dos eventos cardíacos adversos maiores (ECAMs), estes atingiram 24% dos pacientes, dos quais 5% do sexo feminino e 19% do sexo masculino (Figura 15).

Em números absolutos, houve 353 pacientes do sexo feminino com, pelo menos, um evento e 1.288 do sexo masculino, o que resultou numa incidência de eventos próxima a 25% por sexo.

Figura 15 - Desire, nº de pacientes que tiveram ao menos uma ocorrência de ECAM e sua incidência segundo o sexo.

3.2 Processo de *Knowledge-Discovery in Databases* KDD

3.2.1 Pré-processamento

Gerar um modelo aplicando a metodologia de mineração de dados tem uma implicação direta com o conjunto de dados a ser estudado, principalmente na qualidade e no conteúdo de cada atributo que compõe seus registros. O conjunto de dados é a matéria-prima para indução dos modelos e representa toda a experiência do domínio de onde se quer extrair o conhecimento.

O fato da fonte de dados ser de um único centro e restrita a uma mesma equipe de especialistas contribuiu favoravelmente na fase de pré-processamento, pois não houve a necessidade de captação de dados em outras fontes nem os problemas relacionados à duplicidade de registros.

A estrutura de armazenamento dos dados é composta por diversas tabelas relacionadas entre si, as quais correspondem às diferentes fases do protocolo, identificadas como pertencentes aos seguintes subgrupos: ANTECEDENTES que descrevem os dados antropométricos, o histórico e os fatores de risco do paciente; PROCEDIMENTO com informações do primeiro procedimento ou procedimento índice, o quadro clínico do paciente no momento da chegada ao hospital, medicação e resultado dos biomarcadores pré e pós-procedimento; ANGIOGRÁFICOS referem-se às

características dos vasos e das lesões tratadas; *STENTS* com dados relativos aos dispositivos implantados (apenas *stents* farmacológicos); *EVENTOS ADVERSOS* com informações sobre a ocorrência de eventos, como infarto agudo do miocárdio com onda Q e sem onda Q, revascularização de uma lesão previamente tratada ou de uma nova lesão por intervenção percutânea ou procedimento cirúrgico e morte cardíaca. Tais eventos cardíacos maiores serão os atributos-alvos de nosso estudo.

O Anexo A apresenta todas as variáveis disponíveis agrupadas segundo a fase da coleta, com seu tipo, os valores e com uma breve descrição.

O arquivo ARFF (*Attribute Relation File Format*) é um padrão de entrada, para as ferramentas de mineração de dados. Ele tem o formato texto, e é dividido entre cabeçalho e itens: o cabeçalho é composto pelos nomes dos atributos, seu tipo (numérico ou categórico) e seus valores possíveis, os itens referem-se ao conteúdo dos atributos representativos de cada paciente.

Inicialmente, foram extraídos todos os atributos disponíveis na base de dados do Registro Desire para, depois, aplicar as etapas de filtragem e a redução da dimensionalidade por meio da seleção de subconjuntos de atributos para serem submetidos aos algoritmos de mineração.

A seleção e extração dos dados dos pacientes foram realizadas por meio de uma *view*, desenvolvida em linguagem SQL (*Structured Query Language*) nativa do banco e que teve como entrada os atributos distribuídos e armazenados nas diferentes tabelas do Desire, e, como saída, as variáveis ordenadas e seus valores devidamente processados. No processo de KDD, esta etapa, conhecida pela sigla ETL (*Extract, Transform and Load*), é uma das primeiras e das mais importantes, pois é o momento de integrar as diferentes fontes de dados e realizar as transformações necessárias.

O resultado desta *view* foi importado para uma planilha Excel e, posteriormente, convertido ao formato CSV (*comma-separated values*), separando, assim, os valores das diversas variáveis pelo delimitador “ponto e vírgula” para, finalmente, compor os itens do arquivo ARFF.

Outra importante etapa do pré-processamento é a limpeza, em que são identificados os valores faltantes (*missing values*), os valores fora do domínio (*outliers*), os valores inconsistentes resultantes do processo de coleta, da limitação dos instrumentos de medida e por outras causas. O objetivo da limpeza é evitar que os valores incorretos ou ausentes se apresentem como ruído e atrapalhe o processamento dos registros pelo algoritmo de mineração, refletindo, assim, no modelo final.

Esta etapa teve a participação do especialista que verificou todos os atributos categóricos, apresentados no formato de tabelas e que se encontram dispostos no Anexo B. A análise das variáveis numéricas foi realizada pela própria ferramenta, que, após a leitura do arquivo ARFF, forneceu a estatística destas variáveis, como os valores mínimo, máximo, média e desvio padrão, conforme pode ser visto na Figura 16.

Figura 16 - Tela da ferramenta Orange (mostra o objeto: Attribute Statistics) que permite visualizar os valores mínimo, máximo, média e desvio padrão, mediana e os quartis dos atributos quantitativos (Ex.: IMC).

Ainda na fase do pré-processamento, algumas variáveis numéricas foram categorizadas, uma vez que há algoritmos de classificação que lidam melhor com variáveis categóricas do que com as numéricas. Embora a ferramenta Weka possibilite a discretização, optou-se por centralizar esta etapa em um único lote de programação e realizar esta função pela *view*.

Outros atributos foram criados com o objetivo de agregar mais informação, como, por exemplo, o valor médio dos diâmetros dos vasos tratados que considera o diâmetro médio dos vasos cujas lesões foram desobstruídas no procedimento índice; a informação da geração dos *stents* que considera se os dispositivos implantados em determinado paciente, no procedimento índice, pertenciam à mesma geração ou a gerações distintas.

A Figura 17 ilustra um trecho da *view* com detalhe da criação da variável categórica [FUNC_RENAL] que é função do atributo numérico [ext_Clearance] com objetivo de substituir os atributos [InsuficienciaRenal] = [S,N] e [Creatinina], pois, segundo o especialista, esta variável representa melhor a função renal.

```

--dbo.fc_TrataCampo_ExportTese(p.InsuficienciaRenal, 'SimNaoCMB') as InsuficienciaRenal,
-- dbo.fc_TrataCampo_ExportTese(p.Creatinina, 'Num') as Creatinina,

(Especialista sugere introduzir nova variavel para função renal substituindo p.InsuficienciaRenal e
p.Creatinina)

(Especialista diz que clearance reflete melhor)

-- Valor Clearance
--dbo.fc_TrataCampo_ExportTese(p.ext_Clearance, 'Num') as CLEARANCE,

CASE
    when p.ext_Clearance >0 then dbo.fc_TrataCampo_ExportTese(p.ext_Clearance, 'Num')
    --p.ext_Clearance

ELSE '?'

END ASCLEARANCE,

-- Categorizar clearance em func_renal

CASE
    --when p.ext_Clearance >=90 then 'NORMAL'
    --when p.ext_Clearance >=60 and p.ext_Clearance <90 then 'DISCRETA'
    --when p.ext_Clearance >=30 and p.ext_Clearance <60 then 'MODERADA'
    --when p.ext_Clearance <30 then 'GRAVE'
--ELSE '?'

--Redução nas categorias do FUNC_RENAL (07/10/2014), sugerido pelo especialista

    when p.ext_Clearance >=60 and p.ext_Clearance <150 then 'NORMAL_DISCRETA'
    when p.ext_Clearance > 0 and p.ext_Clearance <60 then 'MODERADA_GRAVE'

ELSE '?'

END ASFUNC_RENAL,

```

Figura 17 - Trecho da View com detalhe da criação do atributo [FUNC_RENAL].

A seguir, são detalhados alguns destes atributos e a descrição de seus valores:

- **[ext_vaso_diam_medio]**, atributo numérico que contém o valor médio dos diâmetros dos vasos tratados no procedimento índex.

- **[vaso_fino]**, atributo categórico que indica se as lesões tratadas no procedimento índice se encontravam localizadas somente em vasos com diâmetro menor ou igual a 2,5 mm (S, N, MIX). MIX refere-se ao paciente com lesões localizadas em vasos de distintos diâmetros.
- **[ext_vaso_diam_menor]**, atributo numérico que contém o valor do diâmetro do menor vaso tratado no procedimento índice.
- **[ext_vaso_diam_medio_fino]**, variável categórica tipo S/N, que calcula o valor médio do diâmetro dos vasos tratados e classifica como [S] se todos os vasos tratados forem menor ou igual a 2,5 mm e [N] caso contrário.
- **[Lesao_longa]** variável categórica que informa se todas as lesões tratadas no procedimento índice tiveram comprimento maior do que 30 mm (S, N e MIX).
- **[OneTypeStent_Index]**, variável categórica que identifica se o paciente foi tratado por apenas um modelo de *stent*, isto é, se foi utilizado o mesmo SF em todas as lesões tratadas no procedimento índice cujo domínio é BIOMATRIX, CYPHER, ENDEAVOR, PROMUS, RESOLUT, TAXUS, XCIENCE.
- **[Stent_geracao]**, variável categórica que informa se todos os stents utilizados no procedimento índice foram da mesma geração ou não (NOVA, VELHA, MIX).

Na fase de pré-processamento, pode-se constatar que o KDD é um processo interativo e iterativo. Iterativo, pois, a cada vez que uma variável é recategorizada ou criada, um novo arquivo ARFF é gerado. Interativo, pela participação do especialista, uma vez que estes dados representam os fatos reais do universo observado e tem relação direta com a indução do modelo e com a qualidade do conhecimento que se pretende extrair.

3.2.2 Mineração de dados

A escolha da tarefa para mineração dos dados depende do que se quer encontrar na base de dados ou que tipo de regularidade.

Objetivou-se construir um modelo de conhecimento que classifique os pacientes do Desire relacionando-os à ocorrência de eventos do tipo ECAM e que também seja descritivo, permitindo a interpretação dos relacionamentos entre os atributos independentes e o atributo dependente ou atributo-classe, possibilitando ao médico ter uma visão da interdependência das variáveis mais significativas e entender o porquê de um determinado paciente pertencer a uma determinada classe. Posteriormente, este modelo será avaliado pelo especialista quanto à relevância das informações nele contida visando à sua aplicação como um modelo preditivo para estimar a ocorrência de ECAM quando apresentado a um novo caso.

Assim, optou-se por um aprendizado supervisionado, sobre o conjunto de dados do Desire, para serem classificados segundo uma classe binária (sim ou não) que se refere à ocorrência ou não de um evento cardíaco adverso maior (ECAM).

Dentro da tarefa de classificação, os modelos interpretáveis ou simbólicos são representados por regras de classificação, árvores de decisão ou fórmulas matemáticas, o que depende do algoritmo empregado na mineração. Por outro lado, há os modelos induzidos por algoritmos, como máquinas de vetores-suportes (SVM), redes neurais artificiais e “k-vizinhos” mais próximos que resultam em verdadeiras “caixas-pretas”, impossíveis de serem interpretados³³.

Pelos motivos apresentados, definiu-se pela tarefa de classificação que utilize algoritmos geradores de modelos com aprendizado simbólico induzido e supervisionado, entendendo por aprendizado indutivo o processo que, a partir de exemplos ou fatos, conduz a algumas generalizações³³.

3.2.2.1 A classificação

Classificação é a tarefa de aprendizado que objetiva determinar uma função $f(x)$ que relacione o conjunto de atributos independentes (x) de um objeto com uma das classes predefinidas y , ou melhor, a tarefa de classificação tem como objetivo encontrar um modelo para o atributo-classe que melhor represente sua relação com os valores dos outros atributos do objeto.

Semelhante ao processo de anamnese, que procura avaliar os sinais e sintomas, e correlacionar a um diagnóstico, a classificação analisa os atributos de acordo com as condições e características das classes pré-classificadas pelo modelo.

Figura 18 - Classificação como tarefa para mapear um conjunto de atributos x com um atributo-classe y (PAM: Pressão arterial média em mmHg).

Conforme ilustra o diagrama da Figura 18, uma entrada composta por uma coleção de registros é apresentada para uma tarefa de classificação. Cada registro descreve um paciente ou instância e é caracterizado pela tupla (x,y) , onde:

x representa o conjunto dos atributos de um objeto; e

y é o atributo-alvo ou atributo-classe denominado como categoria.

A Figura 19 mostra alguns registros ambulatoriais fictícios dispostos em tabela e que serão classificados por uma ferramenta de mineração de dados. A última coluna corresponde ao atributo “diagnóstico” (a classe) que contém a informação se o paciente foi classificado com “IAM” ou sem “IAM”.

Figura 19 - Classificador: a indução supervisionada do modelo e dedução da classe para novos exemplos.

O processo ocorre com a submissão de um conjunto de treinamento a um algoritmo de classificação que aprende com os exemplos e cria um modelo que, depois, é aferido por outro conjunto de teste. Após esta etapa, o modelo encontra-se avaliado objetivamente em relação ao seu desempenho como classificador e pode ser empregado para classificar novos exemplos.

Em sua apresentação do processo de KDD, Fayyad salienta que a escolha do algoritmo de mineração de dados depende do interesse da análise pelo usuário final, o qual, muitas vezes, está mais interessado em entender o modelo do que na sua capacidade preditiva¹⁰.

O conjunto de teste é independente do conjunto de treinamento, que pode ser um conjunto externo ou uma divisão do próprio conjunto de treinamento. A forma de teste a ser utilizada é o da validação cruzada ou *k-fold*, que consiste na divisão do conjunto total dos dados disponíveis para treinamento em *k* subconjuntos de mesmo tamanho, mutuamente exclusivo, sendo que um dos subconjuntos é utilizado para o teste e os (*k*-1) restantes são utilizados para estimar os parâmetros do modelo. Este processo é

realizado por k vezes, alternando de forma circular o conjunto de teste como mostra a Figura 20.

Figura 20 - Metodologia 5-Fold para testar um modelo induzido.

A validação cruzada fornece uma boa estimativa sobre o desempenho do modelo, uma vez que estamos testando em diferentes partições dos exemplos e favorecendo a sua generalização, evitando, com isto, o fenômeno conhecido como *overfitting*, quando o modelo fica demasiadamente ajustado pela fase de treinamento (em aprendizado de máquina, *overfitting* ocorre quando um modelo estatístico descreve o erro aleatório ou ruído em vez da relação subjacente).

3.2.2.2 Medida de desempenho do classificador

O estudo atual trata de um classificador binário, com a classe positiva representada pela ocorrência do evento e a negativa, por sua ausência.

Cada vez que um classificador é apresentado para um novo exemplo, ele toma a decisão sobre a classe apropriada em função do modelo aprendido. Algumas vezes, ele está correto, outras vezes, não, assim, a predição pode assumir quatro possibilidades e que resulta numa matriz 2×2 denominada de matriz de contingência ou matriz de confusão. A verdadeira positiva (TP) e verdadeira negativa (TN) são classificações corretas. A falsa positiva (FP) ocorre quando a classe é erroneamente avaliada como positiva

e, na verdade, é negativa. A falsa negativa (FN) ocorre quando a classe é incorretamente classificada como negativa e, na verdade, é positiva.

Matriz de Confusão		
CLASSE	CLASSIFICADOR	
	+	-
REAL +	TP	FN
REAL -	FP	TN

← SENSIBILIDADE

← ESPECIFICIDADE

PRECISÃO

MEDIDAS	
Acurácia	$TP+TN / (TP+TN+FN+FP)$
Sensibilidade, TPR ou Recall	$TP / (TP + FN)$
Especificidade	$TN / (TN + FP)$
VPP (Valor preditivo +)	$TP / (TP + FP)$
VPN (Valor preditivo -)	$TN / (TN + FN)$
FPR = (1-Especificidade)	$FP / (TN+FP)$

TP: VERDADEIRA POSITIVA
TN :VERDADEIRA NEGATIVA
FP: FALSO POSITIVA
FN: FALSO NEGATIVA

TPR: True Positive Rate
FPR: False Positive Rate

Figura 21 - Medidas objetivas para avaliação de um classificador.

A Figura 21 apresenta o conjunto de medidas objetivas de um modelo de classificação composto pela matriz de confusão e as medidas derivadas desta matriz.

A seguir, são apresentadas as medidas mais usadas para avaliar um classificador binário e que fornecem informações objetivas sobre o seu desempenho:

- A **acurácia** é a taxa dos exemplos que são corretamente classificados pelo modelo de treinamento. Medida que pode mascarar a eficácia do classificador caso haja prevalência de uma classe.
- **Sensibilidade** é a fração dos que foram classificados como positivos entre aqueles que são realmente positivos. A sensibilidade relata o acerto do modelo em relação à classe de interesse (positiva).

- **Especificidade** é a fração dos que foram classificados pertencentes à classe negativa entre aqueles que são realmente negativos.
- **VPP**, valor preditivo (+), relata a taxa de acerto positiva, mede o desempenho de predição por classe, corresponde à medida de precisão. É a proporção de exemplos verdadeiramente positivos dentre os classificados positivamente.
- **VPN**, valor preditivo (–, idem ao anterior, relatando a taxa de acerto da classe negativa.

Outra maneira de avaliar o desempenho de classificadores binários é pela curva **ROC** (*Receiver Operating Characteristic*) representada pela relação da sensibilidade *versus* (1- especificidade) ou TPR no eixo das abscissas (Y) *versus* FPR no eixo das ordenadas (X), e apresenta uma forma convexa e a área sob esta curva vem, gradativamente, ganhando espaço como medida de avaliação de modelos em MD³⁴.

Como o valor dado por **(1- especificidade)** ou **FPR** pode ser interpretado como a taxa de falso alarme a qual queremos minimizar e o valor da **sensibilidade** ou **TPR** como a taxa de detecção que queremos maximizar, o ponto ideal do classificador encontra-se no canto superior esquerdo do espaço *ROC*, como pode ser identificado no gráfico da Figura 22.

Fonte: <http://stats.stackexchange.com/questions/105760/how-we-can-draw-an-roc-curve-for-decision-trees>

Figura 22 - Espaço ROC definido por TPR *versus* FPR.

Na área médica, a acurácia, sensibilidade e especificidade são as medidas de desempenho mais utilizadas¹⁴.

3.2.2.3 Classificação por Árvore de Decisão

Dentre as técnicas de classificação, a Árvore de Decisão (AD) é uma das mais utilizadas na construção dos modelos de análise de dados pelos seus resultados e por ter um número reduzido de parâmetros de configuração.

Monard¹¹, no capítulo sobre “Indução de Regras e Árvores de Decisão”, avalia as técnicas de Aprendizado de Máquina simbólico e relata que a indução por Árvore de Decisão é um dos métodos de aprendizado mais utilizados na prática, sendo rápido para o aprendizado de conceitos, simples de programar, permite transformar seus resultados em regras interpretáveis, podendo tratar exemplos com ruído, além de ser uma tecnologia madura e utilizada em vários produtos comerciais.

Zupan e col.¹⁴, em seu artigo sobre o uso da mineração de dados na Medicina, ressaltam a importância dos algoritmos que utilizam representações simbólicas na construção de um conceito, pois permitem a sua interpretação.

A AD destaca-se pela sua transparência, pois permite ao usuário examinar e interpretar o modelo resultante e seu funcionamento pode ser representado por regras do tipo SE-ENTÃO. Sua principal característica é o tipo de representação, constituída por uma estrutura hierárquica, invertida e que se desenvolve da raiz para as folhas.

A construção de uma AD emprega a estratégia conhecida como “dividir-para-conquistar” e que tem como objetivo dividir um problema complexo em subproblemas mais simples, até que a solução para cada um deles tenha sido encontrada. Deste modo, os classificadores baseados em AD buscam meios de dividir um problema, representado pelo conjunto de exemplos de entrada, em subproblemas caracterizados pelos nós da AD. Esta divisão ocorre até que cada um destes nós contenha apenas uma classe ou uma das classes se mostrar majoritária, assim, quando não necessitar mais divisões, chega-se ao nó folha ou, simplesmente, à folha^{12,35}.

Figura 23 - Exemplo de uma AD para as variáveis registradas em um ambulatório fictício.

A Figura 23 apresenta a estrutura de uma AD por meio de um exemplo cujos nós estão representados pelos atributos “Dor” (nó raiz), doença arterial coronariana “DAC” e pressão arterial média “PAM”, sendo que a raiz da árvore corresponde ao atributo mais importante e representativo da classe. Os atributos “Dor” e “DAC” são do tipo categórico e são avaliados pelos valores que podem assumir, dando origem aos ramos que são em número de 3 e 2, respectivamente, enquanto que os ramos gerados pelo atributo PAM, por ser numérico, são representados por intervalos de valor, sendo este intervalo obtido por meio de cálculo inerente ao algoritmo. Os círculos ao final dos ramos representam a classe associada aos nós folhas: “IAM” e “NÃO IAM”.

A classificação pela AD resulta do percurso que parte do nó raiz até atingir os nós folhas, considerando o caminho por meio de seus ramos e os nós intermediários. Por meio desses caminhos descritos ao longo da árvore, é possível derivar regras do tipo:

“Se Dor = Instável e se PAM > 110 mmHg, então diagnóstico de IAM”.

A AD também é usada para a indução de regras de classificação e é única na apresentação por gerar resultados num formato com priorização, com o atributo mais importante apresentado no primeiro nó (raiz), seguido pelos menos relevantes por meio dos nós subsequentes. Assim, é possível identificar facilmente os atributos mais influentes pela posição em que eles se apresentam na árvore^{12,35}.

3.2.2.4 Indução de Árvores de Decisão

A construção da AD baseia-se em um processo recursivo apresentado, inicialmente, por Ross Quilan, da Universidade de Sydney³⁶, o ID3 é considerado o algoritmo pioneiro em indução de árvores de decisão, o qual usa o ganho de informação como critério de divisão.

O ID3 apresentava algumas limitações, como lidar somente com atributos categóricos não ordinais, não apresentar uma forma para tratar os valores desconhecidos (*missing values*) e não possuir um método de “pós-poda”.

Posteriormente, Quilan apresenta outro algoritmo conhecido como C4.5, em 1993³⁷, uma evolução do ID3 e que se tornou referência, sendo muito utilizado, atualmente, por ter apresentado ótimos resultados em problemas de classificação³³. Dentre suas principais características, destacam-se:

- Trabalhar com atributos categóricos ordinais e não ordinais, inclusive com atributos numéricos inteiros e contínuos.
- Tratar os valores desconhecidos (*missing values*).
- Utilizar a razão de ganho para selecionar o atributo como critério de divisão, que se mostrou melhor do que o ganho de informação, gerando árvores mais precisas e menos complexas.
- Possuir método de pós-poda para as árvores geradas, o que reduz o seu tamanho final e minimiza o problema de *overfitting*.

Na mesma época e sem ter conhecimento do trabalho de Quilan, um grupo de estatísticos da Universidade da Califórnia (L. Breiman, J. Friedman, R. Olshen e C. Stone) desenvolveu um algoritmo conhecido como *Classification and Regression Trees* (CART), que induzia tanto árvores de classificação quanto árvores de regressão dependendo se o atributo-classe é categórico ou numérico³⁸.

Tanto o C4.5 quanto o CART são considerados os algoritmos precursores na indução da AD, sendo que, posteriormente, outros surgiram como variações destes.

A seguir, algumas das principais características dos algoritmos de indução de árvore de decisão:

- A indução de uma árvore de decisão, na construção do modelo, apresenta uma abordagem não paramétrica, ou seja, não requer que os atributos assumam qualquer tipo de distribuição de probabilidade;
- A maioria dos algoritmos de indução utiliza a abordagem baseada na metodologia heurística para a busca de uma hipótese para o modelo. (Método heurístico são algoritmos exploratórios que buscam resolver problemas). Partem de uma solução viável e baseiam-se em sucessivas aproximações direcionadas a um ponto ótimo;
- As árvores de decisão, especialmente as de pequeno tamanho, são fáceis de interpretar, sua acurácia é comparável a de outras técnicas de classificação, sendo muito pouco afetadas por atributos redundantes e ainda são robustas na presença de ruído;
- Como a maioria dos algoritmos de indução de árvore de decisão emprega uma abordagem de cima para baixo, com a subdivisão dos atributos (nós) de forma recursiva, a quantidade de registros nas folhas que determinam uma classe vai se tornando menor à medida que caminhamos no sentido da raiz para os ramos mais distantes. Este fenômeno é conhecido como fragmentação dos dados e pode representar um problema quando o número de registros nas folhas for tão pequeno que a decisão na classificação tem pouca significância estatística. Para evitar esta situação, utiliza-se um parâmetro para desabilitar a divisão quando atingir um limite mínimo de registros.

Na Figura 24, pode-se ver a interpretação geométrica do treinamento de uma árvore de decisão para duas classes simbolizadas por ($\blacktriangledown$, $\bullet$) e considerando apenas dois atributos numéricos X e Y, o que se caracteriza por uma árvore de tamanho igual a 7 com número de folhas igual a 4 e refere-se a 15 instâncias.

As folhas contêm a definição de uma classe apenas, o que corresponde ao valor de entropia = 0 e fim do processo de ramificação, pois o objetivo foi alcançado.

A linha divisional entre as classes é sempre paralela ao eixo, porque a condição do teste envolve um atributo por vez.

A complexidade de uma árvore depende da profundidade “d” de sua ramificação, podendo chegar até 2^d para as árvores com divisões binárias. A Figura 24 apresenta uma árvore com $d = 2$, que corresponde à cobertura de $2^2 = 4$ regiões.

Figura 24 - Exemplo da representação gráfica da divisão de classes de uma árvore de decisão.

As regras que representam os ramos da árvore são disjuntas, ou seja, apenas uma regra dispara quando formos classificar um novo exemplo, percorrendo apenas um caminho desde a raiz até alcançar uma folha que irá conter a definição da classe.

3.2.2.5 Construção da Árvore de Decisão

A construção de uma Árvore de Decisão segue um processo recursivo.

Primeiro, é selecionado o atributo que ocupará o nó raiz, ou seja, o primeiro nó. Os ramos derivados deste nó correspondem aos valores que

tem este atributo, o que acaba dividindo os exemplos de treinamento em *subsets*, um para cada valor do atributo ou um para cada ramo.

Este processo repete-se recursivamente para cada ramo até que o nó gerado possua todos os exemplos ou tuplas pertencentes à mesma classe, então, o nó se transforma em folha que é etiquetada com a referida classe.

Tomando como exemplo os registros da Tabela da Figura 25, que descreve as condições meteorológicas para que haja um determinado jogo.

A questão, agora, é determinar o atributo que deve ser escolhido para iniciar a divisão da árvore em sua raiz, sendo que o *atributo-classe* é Play que, neste caso, pode ser Yes ou No.

Reg	Outlook	Temperature	Humidity	Windy	Play?
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rain	mild	high	true	No

Figura 25 - Subset exemplo das condições meteorológicas para jogo de tênis, transcrito do livro Data Mining de Ian H. Witten.

A princípio, existem quatro possibilidades para iniciar a divisão da árvore, numa construção de cima para baixo, as quais estão representadas na Figura 26.

Figura 26 - Quatro inícios de árvores possíveis para os exemplos da tabela da Figura 25.

Na Figura 26, são apresentadas para cada atributo as vezes em que houve jogo (Yes) ou não (No), ou seja, uma referência direta ao atributo-classe. Seguindo um critério intuitivo, a seleção seria pelo atributo que produz o nó mais puro^{35,39}, pois todos seriam da mesma classe e não geraria mais divisão, chegando, assim, mais rápido às folhas que representam o *atributo-classe* e, conseqüentemente, ao fim do processo de crescimento do ramo. Neste caso, a escolha seria pelo atributo *Outlook*.

Os critérios de seleção para encontrar o melhor atributo para realizar a divisão são baseados em diferentes medidas, tais como *impureza*, *distância* e *dependência*, sendo que a maioria dos algoritmos de indução de árvores de decisão busca dividir os dados de um nó PAI de forma a minimizar o grau de impureza dos nós FILHOS. Dentre as medidas mais utilizadas para a seleção da melhor divisão, tem-se o *Ganho de informação* (IG), *Razão de ganho* (GR), *índice Gini* (GI), e outras.

Sendo a função Info() que define o grau de pureza de um atributo num determinado nó, *a qual representa a quantidade de informação esperada que seria necessária para especificar se uma nova instância seria classificada como Yes ou No, uma vez chegado até este nó, tem-se*^{35,39}.

$$\text{Info (nó)} = \sum_{i=1}^n \frac{n_i}{N} \text{Entropia (A}_i\text{)}$$

Onde:

$$\text{Entropia (A}_i\text{)} = - \left(\frac{Y_i}{n_i} \log_2 \frac{Y_i}{n_i} + \frac{N_i}{n_i} \log_2 \frac{N_i}{n_i} \right)$$

Sendo:

A_i = uma tabela com n_i tuplas das quais Y_i são classificadas como YES e N_i são classificadas como NO (caso de classe binária).

O sinal negativo da fórmula de entropia é necessário para deixar o seu valor positivo, uma vez que os logaritmos são negativos, pois são calculados sobre números que variam de 0 a 1 (proporções)

Se escolhermos o atributo *Outlook* da Figura 26, e substituírmos os valores nas equações acima, tem-se:

$A_1 \Rightarrow \text{Folha } \textit{sunny}$	$\text{Entropia (A}_1\text{)} = - \left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \Rightarrow 0,971$
$n_1 = 5 \text{ (YES+NO)}$	
$Y_i = 2 \text{ (nº de YES)}$	
$N_i = 3 \text{ (nº de NO)}$	

$A_2 \Rightarrow \text{Folha } \textit{overcast}$	$\text{Entropia (A}_2\text{)} = - \left(\frac{4}{4} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{0}{4} \right) \Rightarrow 0$
$n_1 = 4 \text{ (YES+NO)}$	
$Y_i = 4 \text{ (nº de YES)}$	
$N_i = 0 \text{ (nº de NO)}$	

Nota-se que a entropia da Folha *overcast* é zero, pois todas as tuplas são da mesma classe. Caso tivéssemos metade das tuplas numa classe e outra metade em outra classe, a entropia seria máxima. A Figura 27 mostra a função Entropia que atinge seu máximo para o valor 0,5.

Figura 27 - Medida de entropia para um problema de classificação binária com valores entre 0 e 1.

$A3 \Rightarrow$ Folha <i>rain</i>	Entropia ($A3$) = $-(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}) \Rightarrow 0,971$
$n1 = 5$ (YES+NO)	
$y_i = 3$ (nº de YES)	
$N_i = 2$ (nº de NO)	

Logo o grau de pureza para o nó relativo ao atributo *Outlook* será:

$$\text{Info}(\text{Outlook}) = (\frac{5}{14} \text{Entropia}(A1) + \frac{4}{14} \text{Entropia}(A2) + \frac{5}{14} \text{Entropia}(A3))$$

$$\text{Info}(\text{Outlook}) = (\frac{5}{14} 0,971 + \frac{4}{14} 0 + \frac{5}{14} 0,971)$$

$$\text{Info}(\text{Outlook}) = 0,693$$

- Para o atributo *Humidity*, tem-se:

$$\text{Info}(\text{Humidity}) = (\frac{7}{14} \text{Entropia}(\text{high}) + \frac{7}{14} \text{Entropia}(\text{normal}))$$

$$\text{Info}(\text{Humidity}) = 0,788$$

- Para o atributo *Temperature*, tem-se:

$$\text{Info}(\text{Temperature}) = \left(\frac{4}{14} \text{Entropia}(\text{hot}) + \frac{6}{14} \text{Entropia}(\text{mild}) + \frac{4}{14} \text{Entropia}(\text{cool}) \right)$$

$$\text{Info}(\text{Temperature}) = 0,911$$

- Para o atributo *Windy*, tem-se:

$$\text{Info}(\text{Windy}) = \left(\frac{8}{14} \text{Entropia}(\text{false}) + \frac{6}{14} \text{Entropia}(\text{true}) \right)$$

$$\text{Info}(\text{Windy}) = 0,892$$

Ganho de Informação (IG)

Define-se **IG** de um atributo como sendo a diferença entre a informação associada (*Info-pré*) ao nó antes da divisão e a informação associada (*Info-pós*) ao nó após a divisão.

A entropia do nó PAI antes da divisão é:

$$\text{Info-pré} = \left(\frac{N_{yes}}{N} \log_2 \frac{N_{yes}}{N} + \frac{N_{no}}{N} \log_2 \frac{N_{no}}{N} \right)$$

Onde:

N_{yes} = nº total de tuplas cujo atributo de classificação é YES.

N_{no} = nº total de tuplas cujo atributo de classificação é NO.

N = nº total de tuplas no nó.

Considerando a situação apresentada na Figura 26, temos que a informação associada ao nó antes da divisão é única e igual a:

$$\text{Info-pré} = \left(\frac{9}{14} \log_2 \frac{9}{14} + \frac{5}{14} \log_2 \frac{5}{14} \right)$$

$$\text{Info-pré} = 0,940$$

Logo, os Ganhos de Informação para cada um dos quatro atributos são:

$$IG (Outlook) = 0,940 - Info (Outlook)$$

$$IG (Humidity) = 0,940 - Info (Humidity)$$

$$IG (Temperature) = 0,940 - Info (Temperature)$$

$$IG (Windy) = 0,940 - Info (Windy)$$

Ou:

$$IG (Outlook) = 0,940 - 0,693 = 0,247$$

$$IG (Humidity) = 0,940 - 0,788 = 0,152$$

$$IG (Temperature) = 0,940 - 0,911 = 0,029$$

$$IG (Windy) = 0,940 - 0,892 = 0,048$$

Pelo critério do Ganho de Informação, a escolha é para o atributo *Outlook* que apresenta o maior ganho, concordando com o critério intuitivo, em que uma de suas Folhas, a *Overcast*, apresenta todas as suas tuplas sendo da mesma classe, o que dá a *Outlook* uma vantagem considerável frente a outros atributos. O valor 0,247 pode ser interpretado como sendo o valor informacional ao se criar uma divisão da árvore utilizando o atributo *Outlook*.

O algoritmo ID3 (Quinlan – 1986) pioneiro na indução de árvores de decisão utiliza este critério para escolha dos atributos para divisão da árvore.

No caso em que sejam avaliados certos tipos de atributos como o nº de registro hospitalar, que será único para cada paciente, este apresentará um exemplo por nó, com uma única classe e cujo valor da entropia seria mínimo e apresentaria um IG máximo. A escolha deste atributo como divisor seria totalmente inútil.

Razão de Ganho (GR)

Quinlan, em 1993, propõe a Razão de Ganho como sendo a razão entre o próprio ganho de informação e a entropia do nó conforme definido pela equação abaixo:

$$\text{Razão de Ganho (nó)} = \frac{IG}{\text{Entropia (nó)}}$$

A Razão de Ganho é uma das melhorias propostas ao algoritmo ID3, que culminou na conhecida versão C4.5 e que solucionou algumas situações limitantes como tratar atributos categóricos ordinais, atributos numéricos, dados ruidosos e atributos com valores desconhecidos (*missing*).

Quinlan mostrou que a Razão de Ganho supera o Ganho de Informação tanto em acurácia quanto na complexidade da árvore gerada³⁵.

O Processo de Poda da Árvore de Decisão

Após sua construção, a árvore de decisão pode refletir ruídos ou erros do conjunto de treinamento, causando muitas arestas, significando que o aprendizado foi muito específico. Esta situação reduz a capacidade de generalização do modelo e é conhecida como *overfitting*.

Para melhorar a taxa de acerto do modelo para novos exemplos, são utilizados métodos de *poda* para detectar e eliminar estas arestas, o que torna a árvore mais simples e facilita a sua interpretabilidade.

Como é de se esperar, deve-se ter o cuidado para que a poda não seja excessiva, pois isto acaba limitando o modelo de classificação.

Existem formas de realizar a poda em uma árvore, as quais são classificadas como *pré-poda* e *pós-poda*.

A pré-poda é realizada durante o processo de construção, o qual é interrompido transformando o nó corrente em uma folha da árvore.

A pós-poda é realizada após a construção da árvore removendo ramos completos, em que tudo o que está abaixo de um nó é excluído e este nó transforma-se numa folha representando a classe mais frequente no ramo³⁵.

3.2.2.6 Indução por regras de classificação

Outra maneira de induzir um modelo simbólico é por meio de Regras de Classificação, que são regras apresentadas na forma “Se *isto* Então *aquilo*”.

Como mostra a Figura 28, uma regra pressupõe uma condição composta por atributos-testes que, se satisfeita, resulta numa classe.

Regra: (*Condição*) $\rightarrow y$

A *Condição* é um conjunto de atributos-teste como:

$(A1=v1)$ e $(A2=v2)$... e $(An=vn)$ e y é o rótulo da classe

As regras não são mutuamente exclusivas, ou seja, um exemplo pode disparar mais de uma regra. Neste caso, a escolha da regra se dá pela ordenação ou por uma estratégia de votação.

Regras de Classificação
"Se isto Então aquilo"
Regra: (Condição) -> y
Onde:
Condição é o conjunto de atributos teste ou Antecedente da regra (A1=v1) e (A2=v2) ... e (An=vn)
y é o rótulo da classe ou Conseqüente da regra

Figura 28 - Regras de decisão que pressupõem a existência de uma condição para a classificação.

As regras podem não cobrir todos os registros apresentados (exemplos), neste caso, é utilizada a classe *default* para a classificação.

As construções de regras de classificação podem ocorrer por método indireto ou direto. O método indireto extrai as regras de outro modelo de classificação, como, por exemplo, da Árvore de Decisão. O método direto extrai as regras dos exemplos de treinamento usando algoritmos específicos, como RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*), CN2 classification rule learner, etc.

Pelo método direto, uma regra é gerada por meio de certa condição, os registros cobertos por esta condição são excluídos do conjunto de treinamento, e, posteriormente, é verificada a necessidade de "poda" ou não desta regra para, depois, ela ser adicionada ao conjunto de regras do modelo e diferem da árvore de decisão por não obedecer a uma hierarquia rígida.

Uma característica forte dos modelos gerados por algoritmos baseados em regras é que estes são fáceis de serem entendidos por usuários não especialistas.

3.3 Ferramentas para mineração dos dados

Atualmente, existem diversas ferramentas disponíveis para exploração e mineração de dados que dispõem de métodos e algoritmos que facilitam o manuseio, o processamento e a análise das informações. Estudos comparativos mostram que não existe uma ferramenta que seja a melhor para todas as aplicações em mineração de dados.

Wahbeh⁴⁰ fez um estudo comparativo entre quatro ferramentas de código aberto e gratuitas para mineração de dados (Weka, Tanagra, KMine e Orange), disponíveis na Internet e muito utilizadas na pesquisa e na academia. Para a avaliação, foi utilizado o mesmo conjunto de dados (*dataset*) e foram realizadas tarefas de classificação por meio de diferentes algoritmos (Árvore de decisão, *Naive Bayes*, *OneR*, *Support Vector Machine*, *ZeroR* e *K nearest Neighbor*).

Concluiu que a ferramenta Weka apresentou o melhor desempenho, seguido pelo Orange, e, depois, pelo KMine e Tanagra.

Em outro estudo recente, “*Comparing Decision Tree Method Over Three Data Mining Software*” Moghimipour⁴¹ compara o desempenho de três ferramentas de mineração (*SPSS-Clementine*, *Rapid Miner* e *Weka*) para a tarefa de classificação por meio da técnica da Árvore de Decisão, devido a este ser considerado o método mais utilizado e efetivo de classificação. O algoritmo *Decision Tree* do SPSS-Clementine foi considerado o melhor pelo critério da acurácia, embora não apresentasse diferença estatística em relação aos outros dois. O algoritmo J48 do Weka ficou em segundo lugar.

Para nosso trabalho, foram escolhidas as ferramentas Weka e Orange por serem muito utilizadas no meio acadêmico, serem gratuitas e terem sido bem avaliadas quando comparadas a outras.

3.3.1 WEKA

Weka (*Waikato Environment for Knowledge Analysis*) é uma ferramenta gráfica que agrega diversos algoritmos de mineração de dados para as tarefas de classificação, regressão, agrupamento e associação.

Foi desenvolvida pelo Departamento de Ciências da Computação da Universidade Waikato, na Nova Zelândia³⁵. Escrita na linguagem Java, permite que seu código seja executado em diferentes plataformas, dando a esse *software* boa portabilidade, além de ser distribuído sob a licença *General Public License* (GPL é a designação da licença para *software* livre idealizada por Richard Matthew Stallman, em 1989, no âmbito do projeto GNU da *Free Software Foundation*), o que lhe confere a possibilidade de se alterar o código-fonte.

Figura 29 - Tela de entrada da ferramenta WEKA versão 3.7.11.

A Figura 29 mostra a versão 3.7.11 da ferramenta utilizada para a mineração dos dados e que se encontra disponível no endereço (<http://www.cs.waikato.ac.nz/ml/weka>).

WEKA - Arquivo de entrada

A entrada de dados obedece ao padrão ARFF (*Attribute-relation file format*) composto por duas partes: um cabeçalho, contendo o nome e tipo

dos atributos e os itens, em que cada linha representa os valores destes atributos relativos a cada paciente.

WEKA - Classificação

Na Figura 30, é apresentada a tela de trabalho da ferramenta Weka para a tarefa de classificação.

A região assinalada pelo círculo nº 1 indica por onde é feita a escolha do algoritmo de classificação, bem como, a configuração de seus parâmetros, como o nº mínimo de exemplos por classe, o fator de confiança de pós-poda, se a divisão dos atributos categóricos é binária ou multivariada, etc. No Weka, o algoritmo indutor da árvore de decisão é o C4.5 de Ross Quilan referenciado por J.48 e o algoritmo indutor para as regras de classificação RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) leva o nome de JRip.

O círculo nº 2 refere-se à forma de acesso aos dados tanto para a indução do modelo quanto para os testes e a avaliação. Assim, pode-se definir que, para o aprendizado, seja utilizado 2/3 dos dados e, para o teste e avaliação, o 1/3 restante ou pode-se carregar um arquivo externo. Neste trabalho, optamos pela validação cruzada dividida em 10 partes (*10-fold cross validation*). Como já mencionado anteriormente, a indução do modelo sobre o conjunto de dados, dividido em 10 partes iguais, utiliza os 9/10 e é avaliado no 1/10 restante, e, assim, sucessivamente, até completar os 10 ciclos.

A execução do algoritmo de classificação é iniciada pela tecla “start” (círculo nº 3) e o resultado do processamento apresentado no formato de relatório indicado pelo círculo nº 4.

Figura 30 - Tela de trabalho do WEKA para escolha do algoritmo classificador, mostrando relatório de saída.

Neste relatório, são apresentadas diversas informações relativas ao desempenho do classificador com suas medidas objetivas, os atributos que fizeram parte da indução, as regras associadas ao modelo gerado, incluindo informações sobre o tamanho da árvore e o número de folhas resultantes.

A estatística Kappa também é fornecida como resultado da classificação. Kappa é um índice de concordância que informa o quanto as predições se afastam das verdadeiras classes, indicando-nos o quão legítimas são as interpretações dadas pela matriz de confusão. O valor máximo de kappa é 100% e o valor esperado para um classificador aleatório é zero, sua escala é interpretada pelos intervalos: 0,01 a 0,20 – levemente concordante; 0,21 a 0,40 razoavelmente concordantes; 0,41 a 0,60 moderadamente concordantes; 0,61 a 0,80 substancialmente concordantes e 0,81 a 0,99 quase perfeitamente concordantes⁴¹.

O Weka permite visualizar a árvore gerada por meio de um comando direto e disponibiliza a mesma árvore no formato de texto. A curva ROC do modelo também pode ser visualizada, bem como, o valor da área sob a curva AUC.

3.3.2 Orange

O Orange (versão 2.7) foi criado pelo laboratório de Inteligência Artificial da Faculdade de Computação e Ciência da Informação da Universidade de Ljubljana na Eslovênia⁴², e distribuído como *software* livre sob a licença *General Public License*.

Orange é uma ferramenta baseada em componentes, composta por objetos e rotinas escritas em C++ com uma variedade de algoritmos de aprendizado de máquina e mineração de dados, e incluem, ainda, rotinas de entrada e manipulação de dados. Por meio de sua interface gráfica denominada *Orange Canvas*, é possível interligar os objetos e criar processos para o desenvolvimento de modelos de classificação, incluindo Árvore de Decisão, *Naive Bayes*, Regras de Decisão, *Support Vector Machine*, etc..

A Figura 31 apresenta detalhes da aplicação desenvolvida com a ferramenta *Orange Canvas* (versão 2.7) que descreve claramente cada passo do processo para indução do modelo de classificação, como a entrada de dados, seleção de atributos, classificação e avaliação do modelo resultante gerado pelo algoritmo de árvore de decisão C4.5 e pelo algoritmo baseado em regras de classificação que utiliza o algoritmo CN2. Esta foi a montagem utilizada para a avaliação dos diversos *datasets*.

A leitura do arquivo com extensão ARFF, contendo o *dataset* a ser analisado, é realizada pelo componente *File*, cujos atributos podem ser visualizados por meio do primeiro objeto conectado o *Data Select Attributes*. Este objeto responde pela seleção dos atributos que irão compor o modelo, possibilitando uma filtragem antes que se aplique o algoritmo de

classificação. Uma conexão com o objeto *Rank* permite avaliar os atributos mais significativos pelos scores de ganho de informação (IG), razão de ganho (GR), entre outros.

Figura 31 - Orange, diagrama dos componentes do processo de classificação pela árvore de decisão e por regras.

O componente *Classification Tree C4.5* é o módulo responsável pela classificação e geração da Árvore de Decisão, enquanto que o resultado do modelo gerado pode ser avaliado graficamente.

Pelos módulos *Test Learner*, *Confusion Matrix* e *ROC Analysis*, é possível coletar os resultados da avaliação do modelo gerado tanto pelo C4.5 quanto pelo CN2.

4 RESULTADOS E DISCUSSÃO

4 RESULTADOS E DISCUSSÃO

O dataset original do Desire é composto por um conjunto de dados com 119 variáveis que se referem às informações colhidas de 6.377 pacientes acompanhados pelo Registro, que relatam o histórico, o quadro clínico, a intervenção percutânea e os eventos. Estas informações após uma fase de pré-processamento são armazenadas em um único arquivo do tipo texto dentro do formato ARFF (*Attribute Relation File Format*), que é um padrão de leitura aceito por diversas ferramentas de mineração de dados, incluindo o Weka e o Orange.

Para a tarefa de classificação, utilizou-se as técnicas de Árvore de Decisão (algoritmo: Weka-J.48 ou Orange-C4.5) e as Regras de Classificação (algoritmo: Weka-JRip, que se baseia no aprendizado de regras proposicional RIPPER e Orange-CN2) com o objetivo de extrair regras por outra técnica simbólica.

A tarefa de classificação foi realizada de modo supervisionado pelo atributo-classe que representa a ocorrência do primeiro evento cardíaco adverso maior ao paciente que pode ser o infarto agudo do miocárdio com onda Q e sem onda Q [IAM_NQ, IAM_Q], a revascularização de uma lesão previamente tratada por cirurgia ou intervenção percutânea [TLR] ou de uma nova lesão [Novalesão], incluindo, também, a morte por causa cardíaca [MCARD].

Devido à natureza binária da classificação, a avaliação do desempenho do modelo foi feita pela matriz de confusão (ou matriz de contingência), por meio das medidas escalares da acurácia, da sensibilidade, da especificidade, do valor preditivo positivo (VPP) e do valor preditivo negativo (VPN), como já mencionado na metodologia.

Uma medida muito poderosa para construção e avaliação de modelos e particularmente útil nos casos de classes desbalanceadas, é extraída da Curva ROC (representada pela taxa de verdadeiros positivos dada pela

sensibilidade e pela taxa de falsos positivos equivalente a '1-especificidade') e representada pela área existente sob esta curva (AUC – *Area Under Curve*), indicando que, quanto maior seu valor, melhor o desempenho do modelo³⁴. O valor máximo da sensibilidade quanto dos falsos positivos (1-especificidade) é de 100%, representando a unidade, logo, a medida máxima da área sob a curva ROC terá o valor 1, equivalente à área de um quadrado de lado unitário.

4.1 Seleção dos atributos

A redução das variáveis por meio da seleção dos atributos mais significativos do *dataset* visa colaborar com o melhor desempenho do algoritmo de aprendizado e a simplificação do modelo sem a perda de sua capacidade analítica.

Foi realizada uma redução inicial na dimensionalidade do *dataset* original com a identificação e a exclusão dos atributos irrelevantes que não contêm informação útil (ex.: identificador de registro hospitalar, do nº do procedimento, etc.); dos atributos redundantes que não agregam valor para a construção do modelo (ex.: peso e altura uma vez que o atributo IMC já expressa uma relação entre eles); das variáveis que não eram totalmente independentes como MACE_IAM, MACE_REVASC e outras que faziam referência ao evento cardíaco adverso maior.

Seguindo a orientação do especialista de domínio, ao longo do processo para geração dos modelos, uma série de reclassificações nos valores dos atributos para as variáveis categóricas foram feitas com o objetivo de melhor ajustar a ramificação da árvore gerada, as quais já foram mencionadas na metodologia.

A curva de aprendizado conduziu para a criação de um único arquivo contendo todos os atributos devidamente preparados em que já se utilizavam os recursos disponíveis nas ferramentas Weka e Orange para filtrar as variáveis de interesse antes da fase de mineração dos dados. Desta

forma, praticamente, reduziram-se os longos ciclos de iteração do KDD, devido ao pré-processamento das variáveis com as repetidas reentradas no processo para recriar novamente o arquivo ARFF.

O arquivo ARFF na sua versão final, denominado D400_127_1A.arff, permite gerar dois conjuntos de dados, o primeiro com 70 atributos, priorizando os dados numéricos, e outro, com 65 atributos, priorizando os categóricos, estes *datasets* serão referenciados mais adiante pelo nome de FULL_NUM e FULL_CATEG, respectivamente, e estão detalhados no Anexo D.

Para uma visão geral do tipo e da fase de ocorrência dos eventos cardíacos, a Figura 32 mostra os eventos ocorridos na fase hospitalar, enquanto que a Figura 33 resume a distribuição dos eventos na fase pós-hospitalar, os quais se encontram agrupados separadamente ao longo do primeiro ano, do segundo ano, do terceiro ao quinto ano, do sexto ao décimo ano e acima do décimo ano.

Figura 32 - Eventos ocorridos na fase hospitalar.

Na fase hospitalar, observa-se a prevalência, em 97% dos casos, do infarto agudo do miocárdio sem onda Q (IAM_NQ). Segundo o especialista,

boa parte destes eventos resulta da instrumentação para o tratamento da lesão considerando seu acesso, o posicionamento do cateter e a expansão do stent, a pós-dilatação adicional, quando necessária, com o balão ocasionando a interrupção do fluxo sanguíneo temporária, a embolização distal, oclusão microvascular, interrupção do fluxo colateral, etc..

Figura 33 - Distribuição dos eventos pós-hospitalar ao longo dos 13 anos.

A ocorrência de lesões celulares, associadas à necrose que caracterizam o IAM_NQ, é detectada por meio de exame laboratorial de biomarcadores cardíacos, como CKMB e Troponina, realizado antes e após o procedimento cujos níveis são o principal critério para a determinação do evento IAM_NQ e que, na maioria das vezes, não apresentam consequências clínicas relevantes. Tal fato tem merecido a atenção da comunidade médica mundial em periódicas revisões sobre a definição de infarto do miocárdio em diversas situações, inclusive durante a intervenção percutânea⁴³.

Considerando os fatos acima e orientado pelo especialista, duas ações foram tomadas: a primeira foi verificar se havia pacientes com valores de CKMB pós-procedimento aumentado, em mais de 5 vezes, sem o devido registro de evento IAM_NQ. Como resultado, foram encontrados e ajustados 112 casos, alterando, assim, a quantidade de eventos hospitalares de 400 para 512, o que elevou o total de eventos registrados para 1.641.

A segunda foi alterar os valores do atributo [NroCKMBPos_cat] que, antes, dividiam os pacientes em faixas subsequentes apenas pelo resultado pós-procedimento em 'ENTRE_1E3, ENTRE_3E10, MAIOR_10' para 'IN_0_OUT_0, IN_0_OUT_MAIOR_5 E OUTRO' que separa os pacientes em subgrupos e considera os valores do biomarcador nas seguintes situações: os que tiveram o valor da CKMB pré e pós-procedimento zerados e inalterados, os que entraram com a CKMB zerada e saíram com o valor 5 vezes acima da referência, e o grupo de pacientes que não se encaixaram em nenhum dos dois grupos anteriores. Esta alteração no valor da variável qualifica melhor o paciente, pois relaciona o biomarcador segundo o valor de referência antes e após o procedimento, que é critério para determinação do IAM_NQ.

4.2 Indução e avaliação dos modelos

Como o processo de KDD é complexo e oferece uma diversidade de alternativas a serem seguidas⁴⁴, houve a necessidade de sistematizar as induções realizadas sobre os *datasets*. Portanto, definiu-se um quadro que resume o processamento realizado por parâmetros, como a classe, tarefa/técnica e os filtros aplicados. As Figuras 34 e 35 explicam este quadro, em que a coluna "item" contém os parâmetros e a coluna "função" a sua relação com a ferramenta de mineração Weka.

Por exemplo, o quadro da Figura 34 especifica um arquivo denominado D400_127_1A.arff, TOP20_GR indica que, antes da mineração, serão selecionados os 20 atributos mais significativos, J4.8 M10 indica que o

algoritmo usado para a MD irá gerar uma árvore de decisão em que o número mínimo de registros por classe deve ser 10, FULL_NUM indica que foi aplicado um filtro para remoção de atributos do arquivo original, deixando as variáveis numéricas e, por último, MACE, que indica o atributo-classe do modelo.

item	função
D400_127_1A	nome do arquivo ARFF
TOP20-GR	weka.classifiers.meta.AttributeSelectedClassifier (TOP 20)
J4.8 M10	weka.classifiers.trees.J48 -- -C 0.25 -M 10
FULL NUM	weka.filters.unsupervised.attribute.Remove (categoricas)
MACE	class attribute

Figura 34 - Exemplo da relação dos itens e as funções associadas à mineração dos dados.

item	descrição
D400_127_1A	Meta classificador
TOP20-GR	Seleciona os 20 melhores atributos pelo critério da razão de ganho
J4.8 M10	Induz árvore de decisão pelo J4.8 (C4.5) com no mínimo 10 casos por folha
FULL NUM	Variáveis independentes do tipo numéricas
MACE	Atributo classe : MACE

Figura 35 - Exemplo da relação dos itens e a descrição do processamento.

O primeiro modelo induzido considerou os eventos ocorridos ao longo dos 13 anos de acompanhamento, segundo o atributo-classe MACE.

A Figura 36 resume estes eventos dividindo-os entre as fases hospitalar e pós-hospitalar. Nota-se que, na fase hospitalar, predomina o infarto agudo do miocárdio sem onda Q (IAM_NQ), representando 33,46% de todos os eventos, enquanto que, na fase pós-hospitalar, destaca-se a revascularização de uma nova lesão com 33,58% dos eventos, seguido pela revascularização da lesão previamente tratada (TLR) e morte cardíaca com 13,65% e 16,09%, respectivamente. O infarto com onda Q representou apenas 3,23% do total de eventos.

Figura 36 - Registro do primeiro ECAM do paciente na fase hospitalar e pós-hospitalar (13 anos).

Modelo

Classe: MACE

A Figura 37 apresenta os resultados dos modelos induzidos P-1, P-2, e P-3, classificados pelo atributo MACE após executar o algoritmo de classificação Weka-J4.8.

P-1 é o resultado apresentado pelo modelo de classificação por Árvore de Decisão composto pelas variáveis independentes numéricas sobre um *dataset* composto por 70 atributos.

P-2 é o resultado quando se empregou um metaclassificador que, antes de gerar a árvore, seleciona os 20 atributos mais significativos pelo critério da razão de ganho.

P-3 é o resultado, quando se aplicou o mesmo procedimento de P-2, sendo que, antes, foi corrigido o desbalanceamento entre as classes pelo método direto de amostragem com redução pela menor classe numa relação 1:1.

Segundo Prati e cols.⁴⁶, o ajuste no sentido do equilíbrio da distribuição entre as classes promove a melhora no modelo de classificação.

CLASSE = MACE		
Parâmetros	P-1	Resultados
D400_127_1A		Acurácia: 81,03%
		Sensibilidade: 36,9 %
J4.8 M10		Especificidade: 96,3 %
FULL NUM		Falso Positivo: 3,7%
MACE		AUC: 0,78
70 ATTRIB		Tam. árvore/Nº folhas: 124/70
		Kappa = 0,400
Parâmetros	P-2	Resultados
D400_127_1A		Acurácia: 81,30%
TOP20-GR		Sensibilidade: 34,8 %
J4.8 M10		Especificidade: 97,4 %
FULL NUM		Falso Positivo: 2,6 %
MACE		AUC: 0,77
70 ATTRIB		Tam. árvore/Nº folhas: 41/24
		Kappa = 0,397
Parâmetros	P-3	Resultados
D400_127_1A		Acurácia: 70,01%
TOP20-GR		Sensibilidade: 70,0 %
J4.8 M10		Especificidade: 70,0%
FULL NUM		Falso Positivo: 30,0%
MACE		AUC: 0,77
70 ATTRIB		Tam. árvore/Nº folhas: 80/50
UnderSample 1:1		Kappa = 0,400
N= 3.282		

Figura 37 - Resultado da classe MACE no período 13 anos, para os modelos P-1, P-2 e P-3.

Analisando objetivamente os dois primeiros modelos P-1 e P-2, nota-se uma acurácia elevada com uma sensibilidade relativamente baixa. A baixa sensibilidade sugere que, quanto mais longo for o período de ocorrência do 1º evento, menor a influência das variáveis independentes do modelo. O índice kappa para P-1 foi de 0,400 e para P-2, de 0,397, indicando que o

modelo induzido apresenta moderada concordância entre a classe e as variáveis independentes.

Em relação ao tamanho do modelo, pode-se observar que P-1 possui uma ramificação composta por 124 variáveis, das quais 70 folhas representam as classes, enquanto P-2 apresenta uma árvore com 41 ramificações e 24 folhas relativas à classe, isto sem perder o poder do modelo, uma vez que a sensibilidade, acurácia e AUC permaneceram equivalentes. A redução no tamanho da árvore deve-se à redução dos atributos, isto foi possível por meio do metaclassificador que selecionou os atributos mais preditivos antes de submeter o *dataset* ao algoritmo de classificação (o Anexo E apresenta o resultado dos 20 atributos selecionados pelo critério da razão de ganho para o *dataset* numérico).

O resultado é uma árvore mais concisa evitando-se o *overfitting*, que é o efeito causado quando o modelo induzido se ajusta demasiadamente ao conjunto de treinamento, desfavorecendo a sua generalização. Por outro lado, modelos muito concisos levam ao conhecimento do óbvio impossibilitando a análise de trechos da árvore que podem apresentar relações subjacentes interessantes entre as variáveis, sugerindo novas hipóteses para serem investigadas.

Em relação ao modelo P-3, optou-se por corrigir o desbalanceamento entre as classes por meio da técnica direta pelo método da subamostragem equalizando-as pela de menor valor (25,7% para classe positiva). O resultado foi um modelo com maior sensibilidade, mantendo o valor da AUC e com aumento do tamanho da árvore em relação ao modelo P-2 em duas vezes, sendo que o índice kappa permaneceu inalterado. (Vide Figura 37)

O modelo P-3 apresentou uma sensibilidade aumentada em relação aos anteriores e uma taxa de falsos positivos de 30%, o que caracteriza que, devido ao custo do erro de classificação, mais pacientes do que o normal serão classificados com a possibilidade de ocorrência de evento futuro, o que é protetor, pois irá demandar um maior acompanhamento assistencial destes pacientes, embora desnecessários. Para o nosso caso, o inverso seria indesejável.

4.3 Criação de duas novas classes

A interação com o especialista permitiu investigar o conjunto de dados, variando o período de observação do evento e a fase de ocorrência.

4.3.1 Classe MFP

Classe que registra os eventos hospitalares e os pós-hospitalares ocorridos até o 1º ano.

A percepção de que os eventos acima do 1º ano passam a ter menos influência das variáveis independentes colhidas no momento do tratamento como o histórico do paciente, o quadro clínico e os dados do procedimento, conduziu para que a análise dos eventos fosse reduzida ao 1º ano, o que acarretou numa diminuição do nº total de eventos de 1.641 para 814. A Figura 38 mostra a nova distribuição destes eventos para a classe MFP.

Figura 38 - Distribuição dos primeiros eventos ocorridos até 1 ANO após o procedimento index.

4.3.2 Classe MACE_REFERENCIA

Classe dos eventos ocorridos na fase pós-hospitalar até o 1º ano e, parcialmente, na fase hospitalar.

Uma segunda hipótese de avaliação dos eventos foi verificar o comportamento do modelo caso fossem retirados todos os pacientes que tiveram eventos hospitalares do tipo IAM_NQ, e fossem registrados somente os segundos eventos destes pacientes na fase pós-hospitalar caso existissem e ainda dentro do período do 1º ano. Esta situação resultou na desclassificação de 494 pacientes com IAM_NQ na fase hospitalar, dos quais somente 21 pacientes deste grupo tiveram um segundo evento incluído dentro do 1º ano da fase pós-hospitalar. Esta alteração reduziu o nº total de eventos para 341 e sua distribuição pode ser verificada na Figura 39.

Figura 39 - Distribuição dos eventos quando se retirou os pacientes com IAM_NQ da fase hospitalar e inseriu um segundo evento destes pacientes, caso houvesse ocorrido dentro do 1º ANO.

Deste modo, nosso estudo passa a avaliar três conjuntos de dados, todos com 6.377 pacientes e que se diferenciam pelas suas classes positivas associadas à ocorrência do evento cardíaco, identificadas por MACE, MFP e MACE_REFERENCIA com 1.614, 814 e 341 pacientes, respectivamente. Os três *datasets* caracterizam-se por apresentar desbalanceamento entre as classes positiva e negativa com percentuais que variam de 25,7%, 12,7% e 5,3% (Vide Figuras 36, 38 e 39).

Aplicando a mesma sequência para minerar a classe positiva MFP, que corresponde à ocorrência de todos os eventos hospitalares e dos pós-hospitalares somente até o 1º ano, nota-se uma melhora no resultado do modelo confirmando a hipótese de que as variáveis independentes passam a ter maior influência.

Modelo

Classe: MFP

CLASSE = MFP		
Parâmetros	P-4	Resultados
D400_127_1A		Acurácia: 94,29 %
		Sensibilidade: 58,8 %
J4.8 M10		Especificidade: 99,5 %
FULL NUM		Falso Positivo: 0,5 %
MFP		AUC: 0,804
70 ATTRIB		Tam. árvore/Nº folhas: 20/13
		Kappa = 0,695
Parâmetros	P-5	Resultados
D400_127_1A		Acurácia: 93,57 %
TOP20-GR		Sensibilidade: 56,0 %
J4.8 M10		Especificidade: 99,1 %
FULL NUM		Falso Positivo: 0,9 %
MFP		AUC: 0,810
70 ATTRIB		Tam. árvore/Nº folhas: 27/16
		Kappa = 0,656
Parâmetros	P-6	Resultados
D400_127_1A		Acurácia: 79,36 %
TOP20-GR		Sensibilidade: 69,3 %
J4.8 M10		Especificidade: 89,4 %
FULL NUM		Falso Positivo: 10,6 %
MFP		AUC: 0,833
70 ATTRIB		Tam. árvore/Nº folhas: 12/7
UnderSample 1:1		Kappa = 0,587
N= 1.628		

Figura 40 - Modelos resultantes da classe MFP com período até o 1º ANO, para os modelos P-4, P-5 e P-6.

Pela Figura 40, observa-se, nos modelos P-4 e P-5, um aumento de 60% na sensibilidade e de 15% na acurácia em relação às mesmas medidas dos modelos P-1 e P-2. A melhora também foi observada nos valores da AUC dos dois modelos, e nos valores de kappa que, para P-4, foi de 0,695 e, para P-5, igual a 0,656. O índice kappa entre 0,65 e 0,69 indica que o modelo induzido apresenta uma substancial concordância entre a classe e as variáveis independentes.

Em relação ao seu tamanho, a árvore gerada apresentou uma redução de seis vezes do modelo P-1 para o P-4 que é explicada pela redução do número de observações dos eventos para o período de um ano. Os modelos que foram construídos pelos metaclassificadores que selecionam previamente os 20 atributos mais significativos, no caso de P-2 para P-5, a redução foi de 1,5 vezes (Vide Anexo E seleção dos atributos mais significativos).

Em P-6, ao aplicar previamente a correção para o desbalanceamento, nota-se o aumento da sensibilidade do novo modelo em, aproximadamente, 10% e, também, uma melhora no valor da AUC. O modelo P-6, embora tenha a sensibilidade aumentada, apresenta uma perda na precisão, com aumento dos falsos positivos, de 0,9% para 10,6%. Avaliando o resultado deste modelo no contexto da aplicação, que é identificar os pacientes que possam vir a ter um evento cardíaco indesejável, conclui-se que ele é mais protetor, pois o erro na precisão traduz-se nos falsos positivos. O índice kappa observado para o modelo P-6 foi de 0,587, aproximadamente, 10% menor do que P-4 e P-5.

Modelo
Classe: MACE_REFERENCIA

CLASSE = MACE_REFERENCIA		
Parâmetros	P-7	Resultados
D400_127_1A		Acurácia: 94,65 %
		Sensibilidade: 0 %
J4.8 M10		Especificidade: 100 %
FULL CATEG		Falso Positivo: 0 %
MACE_REFERENCIA		AUC: 0,498
65 ATTRIB		Tam. árvore/Nº folhas: 1/1
Parâmetros	P-8	Resultados
D400_127_1A		Acurácia: 57,03 %
		Sensibilidade: 55,1 %
J4.8 M10		Especificidade: 58,9 %
FULL CATEG		Falso Positivo: 41,1 %
MACE_REFERENCIA		AUC: 0,563
65 ATTRIB		Tam. árvore/Nº folhas: 31/19
UnderSample 1:1		Kappa = 0,140
N= 682		
Parâmetros	P-9	Resultados
D400_127_1A		Acurácia: 57,62 %
TOP20-GR		Sensibilidade: 54,8 %
J4.8 M10		Especificidade: 60,4 %
FULL CATEG		Falso Positivo: 39,6 %
MACE_REFERENCIA		AUC: 0,563
65 ATTRIB		Tam. árvore/Nº folhas: 20/12
UnderSample 1:1		Kappa = 0,153
N= 682		

Figura 41 - Modelos resultantes da classe MACE_REFERENCIA dentro do 1º ANO, para os modelos P-7 , P-8 e P-9.

Quando foi aplicado o algoritmo de indução Weka-J4.8 sobre o conjunto de dados com a classe positiva MACE_REFERENCIA, que se refere aos eventos pós-hospitalares ocorridos no 1º ano e os hospitalares excluídos os IAM_NQ, o algoritmo não conseguiu gerar o modelo e optou

pela escolha da classe predominante, conforme pode ser visto pelo modelo P-7 na Figura 41.

Isto se deve ao acentuado desbalanceamento entre as classes, na razão de 5,3% (positiva) e 94,7% (negativa), logo, se o algoritmo simplesmente apostar na classe (negativa), ele terá uma acurácia de 94,7%, o que não é ruim, mas a sensibilidade será zero, o que é ruim, pois, apesar de ter o valor da acurácia elevado, o modelo não terá sensibilidade de detectar a classe de interesse (positiva).

A correção do desbalanceamento, neste caso, possibilitou a geração de uma nova árvore de decisão cujos resultados da avaliação estão relatados em P-8 e P-9 e apresentados na Figura 41, com o índice kappa do modelo P-8 = 0,140 e do modelo P-9 = 0,153. O índice kappa entre 0,14 e 0,15 indica que o modelo induzido apresenta uma fraca relação entre a classe e as variáveis independentes.

Diferentemente dos casos anteriores, a correção do desbalanceamento gerou um resultado melhor com o *dataset* categórico (acurácia = 57,03% e AUC = 0,563) comparativamente ao *dataset* numérico (acurácia = 55,86% e AUC = 0,549), inclusive para o modelo composto pelos 20 atributos mais significativos.

4.4 Detalhe do tipo de evento dentro da classe

4.4.1 Classe: MFP_FIRST e Classe: MACE_FIRST

Com o objetivo de avaliar o tipo de evento, induziu-se uma nova árvore de decisão para o atributo-classe **MFP_FIRST** cujos parâmetros e seus resultados podem ser vistos na Figura 42.

Este novo modelo P10, apresentado no modo texto pela Figura 43, classifica os pacientes segundo o tipo de evento ocorrido dentro do primeiro ano e que tem os seguintes valores categóricos IAM_NQ, IAM_Q, MCARD, TLR e NovaLesão.

item		Resultados CLASSE = IAM_NQ		
D400_127_1A		Acurácia: 93,31 %		Sensibilidade: 0 %
TOP20-GR		Sensibilidade: 84,3 %		
J4.8 M10		Especificidade: 99,2 %		Resultados CLASSE = MCARD
FULL NUM		Falso Positivo: 1,0 %		Resultados CLASSE = IAM_Q
MFP_FIRST		AUC: 0,970		Resultados CLASSE = TLR
70 ATTRIB		Tam. árvore/Nº folhas: 19/11		Resultados CLASSE = NEW

Figura 42 - Resultado da classe MFP_FIRST com período até o 1º ano, para o modelo P-10.

O mesmo foi feito para a classe **MACE_FIRST**, gerando o modelo P-11, apresentado no modo texto pela Figura 44 e que classifica os pacientes segundo o tipo de evento ao longo dos 13 anos.

Modelo no modo texto

Classe: MFP_FIRST

Classifier Model P - 10	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0: NAO (4662.0/224.0)	1
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: IAM_NQ (307.0/18.0)	2
NroCKMBPos_cat = OUTRO	3
NroCKMBPos <= 2.98: NAO (937.89/59.33)	4
NroCKMBPos > 2.98	5
Infarto_Recente = S	6
ext_NroVasos_Index <= 2: NAO (230.39/66.34)	7
ext_NroVasos_Index > 2: IAM_NQ (18.67/7.67)	8
Infarto_Recente = N	9
NROSTENTS_Index <= 1	10
STENT_GERACAO = NOVA	11
NroTropoPos <= 37.05: NAO (12.38/3.0)	12
NroTropoPos > 37.05: IAM_NQ (17.96/5.96)	13
STENT_GERACAO = MIX: NAO (0.0)	14
STENT_GERACAO = VELHA	15
IDADE <= 62: NAO (13.34/4.0)	16
IDADE > 62: IAM_NQ (17.01/7.01)	17
NROSTENTS_Index > 1: IAM_NQ (160.36/39.36)	18
Number of Leaves : 11	
Size of the tree : 19	

Figura 43 - Modelo P-10, apresentação da árvore de decisão no modo texto. Eventos ocorridos até 1 ANO.

Modelo no modo texto**Classe: MACE_FIRST**

Classifier Model P - 11	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0: NAO (4662.0/929.0)	1
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: IAM_NQ (307.0/18.0)	2
NroCKMBPos_cat = OUTRO	3
NroCKMBPos <= 2.98: NAO (937.89/143.65)	4
NroCKMBPos > 2.98	5
Infarto_Recente = S	6
TipoInfartoRecente = N: NAO (0.0)	7
TipoInfartoRecente = MENOS_24H: NAO (120.34/40.0)	8
TipoInfartoRecente = ENTRE_24_E_72H: NAO (76.69/26.67)	9
TipoInfartoRecente = MAIS_72H: IAM_NQ (52.02/25.36)	10
Infarto_Recente = N	11
NROSTENTS_Index <= 1	12
STENT_GERACAO = NOVA	13
NroTropoPos <= 37.05: NAO (12.38/3.47)	14
NroTropoPos > 37.05: IAM_NQ (17.96/5.96)	15
STENT_GERACAO = MIX: NAO (0.0)	16
STENT_GERACAO = VELHA	17
ext_lesao_pos_dilat = S: IAM_NQ (19.67/9.67)	18
ext_lesao_pos_dilat = N: NAO (10.68/3.0)	19
NROSTENTS_Index > 1: IAM_NQ (160.36/39.36)	20
Number of Leaves : 13	
Size of the tree : 21	

Figura 44 - Modelo P-11, apresentação da árvore de decisão no modo texto. Eventos ocorridos 13 ANOS.

Os modelos P-10 e P-11 são muito semelhantes e observa-se o predomínio do evento IAM_NQ.

Uma análise mais detalhada entre os modelos mostra que, ao variar o período de observação, eles diferem em dois pontos: a) Para o período de 13 anos, os pacientes que foram tratados com apenas 1 stent da velha geração e que foram pós-dilatados tiveram evento (P-11, linha 18), enquanto que, para o período de 1 ano, nos pacientes na mesma situação, tratados com 1 stent da velha geração, a variável idade maior do que 62 anos foi que determinou a ocorrência do evento IAM_NQ (P-10, linha 17); b) Para os

pacientes que chegaram com infarto recente, o período do infarto foi determinante para o evento quando o período observado cobria 13 anos (P-11, linha 10). Enquanto que, para o 1º ano, ter mais do que 2 vasos tratados é que determinou o evento (P-10, linha 8).

4.5 Extração das regras pelas árvores de decisão

A principal força do modelo induzido pela árvore de decisão reside na sua capacidade de interpretação. Ao contrário dos modelos considerados “caixa-preta”, que são baseados na acurácia, o modelo interpretável traz a compreensão do próprio conjunto de dados que o gerou e, assim, pode fornecer novos *insights* ao especialista de domínio que o analisa, por meio das relações entre suas variáveis, isto é, identificando os atributos mais fortes e sua inter-relação com o atributo-classe. Este tipo de modelo aplica-se muito bem na área da Medicina por possibilitar a proposição de novas hipóteses a respeito do problema em estudo, além do seu poder de retenção do conhecimento com vistas à educação e ao treinamento.

A compreensibilidade de uma árvore de decisão é facilitada por alguns fatores, começando pela visão de uma estrutura gráfica representada por algumas variáveis, não todas, permitindo ao observador focar sua análise nas mais relevantes. A hierarquia de sua estrutura fornece a informação da importância de seus atributos, ou seja, quanto mais próximo da raiz, mais relevante o atributo para a classificação.

Contudo, neste ponto, cabem algumas observações, pois, às vezes, o atributo pode aparecer mais de uma vez numa ramificação partindo da raiz para as folhas, neste caso, a importância está associada ao seu primeiro aparecimento na estrutura da árvore⁴⁷.

Um atributo **A** pode estar mais próximo da raiz do que o atributo **B** e, mesmo assim, **B** pode ser mais relevante devido a ter classificado um número maior de registros.

Outro detalhe a ser observado ao interpretar uma árvore é que, no processo de indução, algum ramo (subárvore) eventualmente pode conter valor irrelevante, mesmo sem a presença de ruído. Isto se deve à natureza do algoritmo de indução, pois, uma vez que um atributo é selecionado para ser o nó da árvore, cada valor deste atributo deve ser incluído. Assim, pode haver algum ramo da árvore que tenha sido adicionado apenas para preservar sua estrutura, associando, assim, ramos irrelevantes que não agregam valor real na interpretação⁴⁷. Este fato, normalmente, está associado ao problema da fragmentação, pois acaba por consumir alguns registros do *dataset* na construção destes ramos, os quais acabam ficando de fora da próxima ramificação.

Na indução de uma árvore, há a fase de crescimento da árvore seguido de uma fase de poda que contribui para melhorar o modelo resultante. O método de poda não consegue resolver o problema de fragmentação apresentado, uma vez que o único caminho para remover os valores irrelevantes (ramos) da árvore remove, também, o atributo relevante (nó).

Para exemplificar a leitura de uma árvore de decisão, a Figura 45 apresenta trecho da AD de um modelo gerado. A análise, partindo da raiz para as folhas, resulta em regras como:

- a- “Do conjunto total, 39 pacientes apresentaram trombo na lesão tratada no procedimento índice e, deste grupo, 72% tiveram um evento cardíaco.”

A folha que classifica os pacientes com atributo **ext_trombo** igual a SIM tem a informação S(39.0/11.0), significando que o modelo identificou 39 pacientes com evento e errou em 11 deles. Este dado permite o cálculo da estimativa de probabilidade da classe positiva = $\{(39-11)/39\}$ para o subgrupo coberto pela regra;

- b- “Dos 89 pacientes, cuja lesão tratada não apresentou trombo no procedimento índice, e que tiveram infarto recente e chegaram ao hospital com um quadro de SCA sem supradesnívelamento de ST, 66% tiveram um evento cardíaco;

- c- “Dos 80 pacientes cuja lesão tratada não apresentou trombo no procedimento índice e que não tiveram infarto recente e que todas as lesões tratadas foram novas e que tiveram a média das estenoses residuais acima de 6,1%, 68% deles tiveram um evento cardíaco.”

Figura 45 - Trecho da árvore de decisão extraída do dataset Desire (modelo P-9).

No Anexo C, é possível verificar as árvores de decisão relativas aos modelos P-1, P-2, P-3, P-4, P-5, P-6, e P-8 no modo texto.

4.5.1 Regras para a classe MACE

A análise da árvore do modelo P-2 relativo à classe MACE, que foi induzida pelo metaclassificador, permitiu extrair regras que relatam a probabilidade de ocorrência ou não de eventos adversos nos pacientes na fase hospitalar e pós-hospitalar ao longo dos 13 anos de seguimento e que são apresentadas na Figura 46.

Classe: MACE - Modelo P-2 (6.377 pacientes sendo 1.614 com eventos ao longo de 13 anos)				
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde
2	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de segunda geração	N	94%	1765
3	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira e segunda geração	N	87%	110
7	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com histórico cardiovascular e com estenose residual média das lesões pós procedimento $\leq 8,6\%$	N	69%	1248
10	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com quadro diferente de ISQUEMIA SILENCIOSA, com histórico cardiovascular e a estenose residual média das lesões pós procedimento $> 8,6\%$	S	60%	65
11	Idem a linha 7, em pacientes sem histórico cardiovascular	N	78%	1283
22	Se o valor da CKMB = 0 no exame pré e se for maior do 5 vezes no exame pós procedimento	S	100%	307
24	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MENOR do que 2,98	N	85%	937
26	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 mas SEM infarto recente	S	73%	221
31	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade > 83 anos	S	83%	12
30	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 32,78 em pacientes COM infarto recente MENOS_24H e idade MENOR do que 83 anos	S	65%	20
29	Idem a regra anterior exceto pelo valor do CKMB pós MENOR do que 32,78	N	82%	88

Figura 46 - Regras extraídas do modelo P-2 (13 ANOS).

A primeira regra cobre um grupo de 1.765 pacientes classificados como não tendo evento com uma probabilidade de 94%. A sexta regra cobre um grupo de 307 pacientes em que 100% tiveram um evento cardíaco.

4.5.2 Regras para a classe MFP

A Figura 47 apresentam as regras extraídas da árvore para o modelo P-5 relativas à classe MFP, que classifica as ocorrências do primeiro evento cardíaco adverso na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Modelo P-5 (6.377 pacientes sendo 814 com eventos até o 1º ano)				
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO	N	95%	4662
2	Se o valor da CKMB = 0 no exame pré e for maior do 5 vezes a no exame pós procedimento	S	98%	307
4	Se o valor do CKMB no exame pré for diferente de ZERO e no pós procedimento estiver entre ZERO e 2,98	N	94%	937
9	Se o paciente apresentou-se com infarto recente MENOS_24H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver no intervalo entre 2,98 E 35,48	N	83%	98
10	Idem a regra anterior, mas com valor da CKMB pós procedimento MAIOR do que 35,48	S	73%	22
11	Se o paciente apresentou-se com infarto recente ENTRE_24_E_72H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver MAIOR do que 35,48	N	80%	76
16	Se o paciente não teve infarto e o valor do CKMB no exame pré for DIFERENTE de ZERO e no exame pós procedimento estiver MAIOR do que 2,98 e tiver overlap de stent	S	89%	46
20	Se o paciente apresentou-se com quadro de SCA_SSST e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU APENAS 1 VASO sem overlap de stent	S	69%	97
26	Se o paciente apresentou-se sem infarto recente e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU MAIS DE 1 VASO sem overlap de stent	S	75%	100

Figura 47 - Regras extraídas do modelo P-5 (1 ANO).

4.5.3 Regras para a classe MACE_REFERENCIA

O modelo de classificação P-8 foi reproduzido na Figura 48 é uma cópia do Anexo C, para que se possa observar a ligação entre as regras extraídas e o modelo textual a que são referenciadas pelo número da linha. O quadro da Figura 49 apresenta as regras extraídas do modelo P-8.

Estas regras classificam os pacientes pela classe MACE_REFERENCIA, similar à classe MFP com a diferença de que foram excluídos os eventos relativos aos infartos sem onda Q da fase hospitalar.

Árvore de Decisão: MACE_REFERENCIA

Classifier Model - P - 8	
J48 pruned tree	

ext_trombo = S: S (39.0/11.0)	1
ext_trombo = N	2
DoencaVascPeriferica = S: S (33.0/11.0)	3
DoencaVascPeriferica = N	4
ext_residual_Media <= 6.13	5
Infarto_Recente = S	6
FAIXA_ETARIA = IDOSO: S (31.2/8.68)	7
FAIXA_ETARIA = ADULTO	8
DIABETES = N	9
NroCKMBPos_cat = IN_0_OUT_0	10
ext_calcio = NAO: S (11.84/4.84)	11
ext_calcio = MODERADO: N (10.0/3.0)	12
ext_calcio = ALTO: S (1.0)	13
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: N (4.0)	14
NroCKMBPos_cat = OUTRO: S (19.84/5.0)	15
DIABETES = SIM_NIDep: N (12.0/2.0)	16
DIABETES = SIM_IDep: S (1.84)	17
FAIXA_ETARIA = JOVEM: S (2.84)	18
Infarto_Recente = N	19
LESOES_NOVAS = S	20
ext_calcio = NAO: N (212.24/82.2)	21
ext_calcio = MODERADO	22
FUNC_RENAL = NORMAL_DISCRETA: N (119.03/33.91)	23
FUNC_RENAL = MODERADA_GRAVE: S (35.85/14.93)	24
ext_calcio = ALTO: S (27.84/10.0)	25
LESOES_NOVAS = N	26
ext_calcio = NAO: S (12.0/2.0)	27
ext_calcio = MODERADO: N (11.0/4.0)	28
ext_calcio = ALTO: S (0.0)	29
ext_residual_Media > 6.13: S (97.47/30.4)	30
Number of Leaves : 19	
Size of the tree : 31	

Figura 48 - Modelo P-8: Árvore de decisão no modo texto.

Neste modelo (Figura 48), pode-se observar que outras variáveis passam a ter maior significância, como a presença de trombo nas lesões tratadas representa aos pacientes cobertos por esta regra 72% de probabilidade de ter um evento adverso no primeiro ano (P-8, linha 1), a presença da doença vascular periférica representa ao paciente uma probabilidade de 67% de vir a ter um evento dentro do primeiro ano (P-8, linha 3), incluem, também, os atributos estenose residual média (P-8, linha 30), a presença de cálcio nas lesões tratadas (P-8, linha 25) e o estado da função renal do paciente (P-8, linha 24).

Classe: MACE_REFERENCIA - Modelo P-8 (6.377 pacientes sendo 341 com eventos)				
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde
1	Se houver presença de trombo nas lesões tratadas	S	72%	39
3	Se não houver trombo nas lesões mas o paciente tiver doença vascular periférica	S	67%	33
7	Paciente idoso COM iam recente SEM doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$	S	72%	31
21	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM valor da estenose residual média das lesões $\leq 6,13\%$. Se TODAS lesões novas e SEM a presença de cálcio	N	61%	212
24	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$ com TODAS lesões novas, COM presença MODERADA de cálcio e função renal MODERADA_GRAVE	S	60%	35
23	Idem à anterior mas com função renal NORMAL_DISCRETA	N	72%	119
25	Idem à 21 diferindo da presença ALTA do cálcio nas lesões.	S	63%	27
30	Paciente sem doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $> 6,13\%$.	S	69%	97

Figura 49 - Principais regras extraídas do modelo P-8.

4.6 Regras de classificação

A regra de classificação é outra maneira de indução de modelo interpretável e apresenta-se na forma: **Se** (condição), **então** (classe), assim como as regras oriundas da árvore de decisão, embora não possuam uma representação gráfica.

A sua apresentação não dá pistas da importância de cada atributo, diferentemente da posição hierárquica em que se encontram os atributos na árvore de decisão.

Cada regra fornece ao analista uma visão de uma parte do conhecimento, assim como as peças de um quebra-cabeça, o que, de certa forma, restringe a visão do todo.

Uma diferença importante entre as regras de classificação e a árvore de decisão é que, na AD, cada registro classificado termina numa folha muito bem definida, ou seja, as regras derivadas da AD partem da raiz para as folhas e são mutuamente exclusivas. O mesmo não ocorre com as regras de classificação em que um registro pode ser atendido por uma ou mais regras. Para solucionar este tipo de conflito, as regras são descritas na forma de listagem ordenada considerando índices como a cobertura e a qualidade da regra.

As regras de classificação tendem a não apresentar condições irrelevantes, como as mencionadas na árvore de decisão devido à própria natureza da construção da AD. A explicação para isto é que o algoritmo da AD tipicamente seleciona um atributo por vez, enquanto que o algoritmo da regra de decisão seleciona um valor do atributo por vez quando está se expandindo⁴⁷.

A seguir, é apresentado o modelo minerado pelo algoritmo de regra de classificação da ferramenta Orange-CN2, que possui uma formatação de saída melhor do que o Weka-RIPPER. As regras encontram-se ordenadas segundo sua qualidade e cobertura para o atributo-classe MFP = SIM e MFP = NÃO, conforme mostram as Figuras 50 e 51.

A primeira regra da Figura 50 é:

“IF NroCKMBPos_cat=IN_0_OUT_MAIOR_5 AND NroTropo Pos>29.00
AND IDADE>35.00 THEN MFP =S”.

Cuja interpretação é: se o paciente com idade maior do que 35 anos, tiver o valor da CKMB pré-procedimento zerada e do pós-procedimento aumentado em mais de 5 vezes em relação ao valor de referência e a Troponina estiver aumentada em mais de 29 vezes, ele será classificado como sujeito a um evento no período de até um ano.

Esta regra apresenta um comprimento igual a 3 que corresponde ao número de atributos envolvidos e com uma relevância traduzida pela qualidade da regra que é igual a 0,995 (intervalo varia de 0 a 1) e sua cobertura abrange 184 dos 841 pacientes com eventos, num total de 6.377 pacientes.

A terceira regra indica que pacientes com idade maior do que 69 anos com valores de CKMB pré-procedimento normal e a CKMB pós-procedimento duas vezes acima da referência, e com a função renal moderada a grave, terá um evento em até um ano, com uma cobertura para 39 pacientes:

“IF Nro CKMBPos>2.00 AND CKMB_PRE=NORMAL AND
FUNC_RENAL=MODERADA_GRAVE AND IDADE>69.00 THEN MFP=S”.

Freitas⁴⁷ avaliou a interpretabilidade dos modelos de classificação por especialistas médicos e identificou que modelos muito simples não foram bem aceitos. Há uma preferência por modelos maiores, pois estes contêm mais variáveis associadas, fazendo mais sentido ao usuário embora a ideia do que venha a ser um modelo grande tenha variado muito segundo o autor.

Uma característica própria dos modelos interpretáveis é que sua leitura permite, também, observar as exceções, pesquisar as classificações que contradizem a lógica e, até mesmo, avaliar a classe negativa invertendo,

assim, a lógica da análise. Tal possibilidade amplia ainda mais o poder deste tipo de modelo, pois permite chegar a novas hipóteses no domínio de aplicação.

Entre as regras que analisam a não ocorrência de eventos, pode-se citar a regra abaixo, uma das primeiras com a classe MFP = N, extraída da posição central da Figura 51, que tem uma cobertura de 318 pacientes e qualidade = 0,997:

```
"IF NroCKMBPos_cat = IN0_OUT_0 AND NroTropoPos > 0 AND ext_calcio  
= MODERADO AND CLEARANCE > 64 and CLEARANCE < 130 THEN  
MFP = N".
```

Ou seja,

Houve um grupo de 318 pacientes que tiveram a CKMB = 0 antes e após o procedimento, com valor de troponina pós-procedimento aumentada, com clearance entre 64 e 130, e lesões com cálcio moderado e não tiveram registro de evento cardíaco no primeiro ano após o procedimento.

Outra regra relativa a não ocorrência de evento, com uma cobertura de 108 pacientes e qualidade = 0,993, refere-se à bifurcação da lesão, é:

```
"IF NroCKMBPos_cat = IN0_OUT0 AND ext_bifurcacao = S AND IMC>=34  
AND ext_lesoes_compr_tot <= 40 THEN MFP = N".
```

Ou

Existem 108 pacientes que tiveram a CKMB = 0 antes e após o procedimento, com IMC ≥ 34 cuja soma das lesões tratadas tiveram o comprimento máximo de 40 mm e havia, ao menos, uma bifurcação que não registraram evento cardíaco no primeiro ano após o procedimento.

CN2 Rules Viewer				Wed Mar 09 16, 20:37:51	
Rule length	Rule quality	Coverage	Predicted class	Rule	
3	0.995	184.000	S	IF NoCKMBPos_cat=IN_0_OUT_MAIOR_5 AND NoTroPos>29.00 AND IDADE>35.00 THEN MFP=S	
3	0.989	90.000	S	IF NoCKMBPos_cat=IN_0_OUT_MAIOR_5 AND ext_vaso_diam_menor>2.00 AND ext_trombo=N THEN MFP=S	
4	0.976	39.000	S	IF NoCKMBPos>2.00 AND CKMB_PRE=NORMAL AND FUNC_RENAL=MODERADA_GRAVE AND IDADE>69.00 THEN MFP=S	
4	0.967	28.000	S	IF NoCKMBPos>2.00 AND NROSTENTS_Index>3.00 AND NoCKMBPos>3.00 AND NoCKMBPos<=13.00 THEN MFP=S	
2	0.929	12.000	S	IF TROPONINA_VAR=DELTA>1000 AND NoCKMBPre<=2.00 THEN MFP=S	
5	0.962	24.000	S	IF NoCKMBPos>2.00 AND CKMB_PRE=NORMAL AND Quadro_Chegada=ANG_INST AND Hiperlipidemia=S AND ext_NroLesions_Index<=3.00 THEN MFP=S	
5	0.962	24.000	S	IF NoCKMBPos>2.00 AND Quadro_Chegada=ISQUEMIA AND ext_NroVasos_Index_categoria=2 AND ext_residual_Media<=3.00 AND IMC>23.00 THEN MFP=S	
4	0.929	12.000	S	IF NoCKMBPos>2.00 AND Quadro_Chegada=ANG_EST AND CLEARANCE<=100.00 AND Stent_Compr_Total_Index>16.00 THEN MFP=S	
3	0.909	9.000	S	IF NoCKMBPos_cat=IN_0_OUT_MAIOR_5 AND QUADRO_SCA=SCA_CSST AND ext_territ_CX=N THEN MFP=S	
3	0.900	8.000	S	IF TROPONINA_VAR=30012.00 THEN MFP=S	
3	0.929	12.000	S	IF NoCKMBPos>2.00 AND TipoInfartoRecente=MAIS_72H AND CKMB_PRE=NORMAL THEN MFP=S	
5	0.952	19.000	S	IF NoCKMBPos>2.00 AND Quadro_Chegada=ISQUEMIA AND ext_NroTerritorios_Index_categoria=HUM AND ext_vaso_diam_menor<=2.00 AND Stent_Compr_Total_Index<=48.00 THEN MFP=S	
3	0.857	5.000	S	IF NoCKMBPos>2.00 AND ext_Ponte_Idade>13.00 AND Isquemia_Silenciosa=N THEN MFP=S	
2	0.800	3.000	S	IF NoTroPre<=1.00 AND NoCKMBPos>19.00 THEN MFP=S	
4	0.889	7.000	S	IF NoCKMBPos>2.00 AND Quadro_Chegada=ANG_INST AND Stent_Compr_Total_Index>58.00 AND FVE=NORMAL_DISCRETO THEN MFP=S	
3	0.750	2.000	S	IF NoCKMBPos_cat=IN_0_OUT_MAIOR_5 AND lesao_longa=MIX AND CLEARANCE<=87.00 THEN MFP=S	
4	0.833	4.000	S	IF NoTroPre<=1.00 AND IMC>28.00 AND ext_lesao_classif=C AND Stent_Compr_Total_Index>15.00 THEN MFP=S	
2	0.750	2.000	S	IF NoCKMBPre>54.00 AND NoCKMBPre<=60.00 THEN MFP=S	
4	0.917	10.000	S	IF NoCKMBPos>2.00 AND NoCKMBPre<=2.00 AND NoCKMBPos>4.00 AND HistonaFamiliar_DCA=N THEN MFP=S	
1	0.667	1.000	S	IF Stent_Compr_Total_Index>165.00 THEN MFP=S	
4	0.833	4.000	S	IF NoCKMBPos>2.00 AND OneTypeStent_Index=PROMUS AND NoTroPos<=52.00 AND IMC<=29.00 THEN MFP=S	
2	0.750	2.000	S	IF NoCKMBPos>2.00 AND TipoAngEstavel=CCS1 THEN MFP=S	
3	0.800	3.000	S	IF NoCKMBPos>2.00 AND TipoAngEstavel=CCS2 AND IDADE<=54.00 THEN MFP=S	
2	0.750	2.000	S	IF NoCKMBPos>2.00 AND TipoAngInstavel=IB THEN MFP=S	
3	0.800	3.000	S	IF ext_trombo=S AND NoTroPre>289.00 AND CLEARANCE>84.00 THEN MFP=S	
3	0.800	3.000	S	IF IMC>42.00 AND IMC<=43.00 AND CLEARANCE<=126.00 THEN MFP=S	
3	0.800	3.000	S	IF DoencaVascPeriferica=S AND IMC<=20.00 AND IDADE<=62.00 THEN MFP=S	
3	0.800	3.000	S	IF NoCKMBPos>2.00 AND Stent_Compr_Total_Index<=13.00 AND IMC<=25.00 THEN MFP=S	
3	0.800	3.000	S	IF CLEARANCE<=35.00 AND TipoInfartoRecente=ENTRE_24_E_72H AND ext_residual_Media>6.00 THEN MFP=S	
3	0.800	3.000	S	IF CLEARANCE<=28.00 AND ext_residual_Media>10.00 AND Stent_Compr_Total_Index<=33.00 THEN MFP=S	
4	0.800	3.000	S	IF ext_trombo=S AND CVA=S AND CLEARANCE<=70.00 AND IAM_ANT=NO THEN MFP=S	
5	0.833	4.000	S	IF NoCKMBPos>2.00 AND TipoInfartoRecente=MAIS_72H AND ext_territ_CX=S AND IDADE<=64.00 AND QUADRO_SCA=SCA_SSST THEN MFP=S	
3	0.750	2.000	S	IF IDADE<=39.00 AND NoTroPre>20.00 AND Stent_Compr_Total_Index<=23.00 THEN MFP=S	
3	0.833	4.000	S	IF ext_calcio=ALTO AND TipoAngInstavel=IB AND Stent_Compr_Total_Index<=24.00 THEN MFP=S	
4	0.889	7.000	S	IF NoCKMBPos>3.00 AND NoCKMBPos<=4.00 AND ext_residual_maxima>5.00 AND IMC<=29.00 THEN MFP=S	
4	0.833	4.000	S	IF ext_Ponte_Idade_maior_10anos=N AND ext_residual_Media>6.00 AND IDADE>73.00 AND Estatina=N THEN MFP=S	
3	0.750	2.000	S	IF DIABETES=SIM_IDep AND ext_NroVasos_Index_categoria=4 AND CLEARANCE<=105.00 THEN MFP=S	
4	0.800	3.000	S	IF IDADE>86.00 AND Quadro_Chegada=ISQUEMIA AND vaso_fino=MIX AND Hiperlipidemia=N THEN MFP=S	
2	0.750	2.000	S	IF TipoAngInstavel=IIA AND IMC>30.00 THEN MFP=S	
3	0.750	2.000	S	IF DIABETES=SIM_IDep AND ext_Ponte_Idade<=7.00 AND Stent_Compr_Total_Index>31.00 THEN MFP=S	
4	0.800	3.000	S	IF CLEARANCE<=35.00 AND NoCKMBPos<=2.00 AND NoTroPre>2.00 AND NoCKMBPre<=2.00 THEN MFP=S	
4	0.800	3.000	S	IF ext_residual_Media>5.00 AND AgenteTrombolitico=S AND ext_lesao_classif=C AND CLEARANCE>51.00 THEN MFP=S	
3	0.750	2.000	S	IF DoencaVascPeriferica=S AND NoTroPre<=2.00 AND ext_residual_maxima>3.00 THEN MFP=S	
3	0.750	2.000	S	IF ext_lesoes_compr_tot<=5.00 AND IMC>26.00 AND Stent_Compr_Total_Index<=12.00 THEN MFP=S	
4	0.833	4.000	S	IF ext_residual_Media>6.00 AND Clopidogrel=N AND IDADE<=59.00 AND Stent_Compr_Total_Index>18.00 THEN MFP=S	
5	0.833	4.000	S	IF ext_residual_Media>6.00 AND LESOES_NOVAS=N AND IAM_ANT=QWAVE AND ext_vaso_diam_medio<=2.00 AND ext_calcio=NAO THEN MFP=S	
4	0.800	3.000	S	IF ext_NroTerritorios_Index_categoria=TRES AND IMC>36.00 AND Stent_Compr_Total_Index>78.00 AND Stent_Compr_Total_Index<=98.00 THEN MFP=S	
1	0.667	1.000	S	IF NoCKMBPos>168.00 THEN MFP=S	
2	0.667	1.000	S	IF KILLIP=III AND ext_NroVasos_Index_categoria=1 THEN MFP=S	
2	0.667	1.000	S	IF DIABETES=SIM_IDep AND NoCKMBPos>17.00 THEN MFP=S	
2	0.667	1.000	S	IF DoencaVascPeriferica=S AND TipoAngInstavel=IIIC THEN MFP=S	
3	0.750	2.000	S	IF ext_Ponte_Idade>8.00 AND IDADE<=51.00 AND CLEARANCE<=80.00 THEN MFP=S	
4	0.800	3.000	S	IF DIABETES=SIM_IDep AND ext_residual_maxima>9.00 AND IMC>29.00 AND Stent_Compr_Total_Index>23.00 THEN MFP=S	
4	0.800	3.000	S	IF DoencaVascPeriferica=S AND TipoInfartoRecente=MAIS_72H AND ext_residual_maxima>3.00 AND ext_residual_maxima<=5.00 THEN MFP=S	
4	0.800	3.000	S	IF ext_NroVasos_Index_categoria=3 AND CLEARANCE>151.00 AND ext_NroTerritorios_Index_categoria=DOIS AND CLEARANCE<=176.00 THEN MFP=S	
2	0.667	1.000	S	IF NoCKMBPre>2.00 AND IMC<=18.00 THEN MFP=S	
2	0.667	1.000	S	IF TipoAngInstavel=IIIB AND ext_residual_maxima>31.00 THEN MFP=S	
3	0.667	1.000	S	IF NoCKMBPos<=1.00 AND AgenteTrombolitico=S AND Stent_Compr_Total_Index<=13.00 THEN MFP=S	
2	0.667	1.000	S	IF IDADE<=49.00 AND TipoAngInstavel=IIC THEN MFP=S	
2	0.667	1.000	S	IF ext_NroTerritorios_Index_categoria=TRES AND ext_vaso_diam_medio<=2.00 THEN MFP=S	
3	0.750	2.000	S	IF OneTypeStent_Index=TAXUS AND CLEARANCE<=30.00 AND Stent_Compr_Total_Index>28.00 THEN MFP=S	
3	0.750	2.000	S	IF NoCKMBPos<=1.00 AND CLEARANCE<=34.00 AND Stent_Compr_Total_Index<=37.00 THEN MFP=S	
3	0.750	2.000	S	IF ext_calcio=ALTO AND ext_residual_maxima>32.00 AND ext_residual_maxima<=35.00 THEN MFP=S	
3	0.750	2.000	S	IF IDADE<=49.00 AND ext_calcio=ALTO AND Quadro_Chegada=ANG_INST THEN MFP=S	
2	0.667	1.000	S	IF REVASC_ANT=S AND NoCKMBPos>39.00 THEN MFP=S	
3	0.667	1.000	S	IF ext_residual_maxima>2.00 AND ext_NroLesions_Index>5.00 AND Stent_Compr_Total_Index>132.00 THEN MFP=S	
3	0.750	2.000	S	IF TipoInfartoRecente=MAIS_72H AND NoCKMBPos>3.00 AND NoCKMBPre<=2.00 THEN MFP=S	
4	0.800	3.000	S	IF TipoAngInstavel=IIIB AND NoCKMBPos<=1.00 AND ext_calcio=NAO AND IDADE<=54.00 THEN MFP=S	
4	0.750	2.000	S	IF ext_residual_Media>2.00 AND TipoAngInstavel=IIA AND ext_lesao_pos_dilat=N AND IMC>25.00 THEN MFP=S	
4	0.800	3.000	S	IF Fumo=S AND ext_residual_Media>7.00 AND CLEARANCE>102.00 AND CLEARANCE<=124.00 THEN MFP=S	
4	0.800	3.000	S	IF Fumo=PAROU AND TERRITORIO_TCE=S AND NoCKMBPos<=1.00 AND ext_lesoes_compr_tot>36.00 THEN MFP=S	
4	0.750	2.000	S	IF ext_territ_CD=S AND NoCKMBPre>27.00 AND CLEARANCE>100.00 AND Stent_Compr_Total_Index>18.00 THEN MFP=S	
4	0.800	3.000	S	IF ext_territ_DA=N AND CLEARANCE>179.00 AND IMC>37.00 AND ext_vaso_diam_menor<=2.00 THEN MFP=S	
4	0.750	2.000	S	IF STENT_GERACAO=VELHA AND IDADE>87.00 AND ext_lesao_classif=C AND ext_residual_maxima<=3.00 THEN MFP=S	
5	0.833	4.000	S	IF REVASC_ANT=S AND IDADE<=48.00 AND Fumo=S AND Estatina=N AND TERRITORIO_TCE=N THEN MFP=S	
4	0.800	3.000	S	IF ext_SomenteNativos_Index=N AND NoTroPre<=3.00 AND NoCKMBPos>1.00 AND TERRITORIO_TCE=N THEN MFP=S	
4	0.750	2.000	S	IF SEXO=F AND ext_NroVasos_Index_categoria=3 AND NoCKMBPos<=1.00 AND CLEARANCE>35.00 THEN MFP=S	
4	0.750	2.000	S	IF CKMB_POS=N AND NoTroPre>14.00 AND TipoInfartoRecente=MAIS_72H AND Fumo=N THEN MFP=S	

Figura 50 - Primeira parte do modelo de classificação pelo atributo MFP, composto por 6 partes. (algoritmo CN2).

3	0.667	1.000	S	IF lesao_longa=S AND IMC<=20.00 AND OneTypeStent_Index=XCIENCE THEN MFP=S
4	0.667	1.000	S	IF Hiperlipidemia=N AND Stent_Compr_Total_Index<=12.00 AND TipoAngEstavel=CCS2 AND CLEARANCE>32.00 THEN MFP=S
4	0.750	2.000	S	IF ext_residual_Media>3.00 AND ext_residual_maxima<=3.00 AND TipoAngEstavel=CCS2 AND SEXO=M THEN MFP=S
5	0.800	3.000	S	IF IMC>26.00 AND IMC<=27.00 AND ext_residual_maxima>7.00 AND ext_territ_CD=S AND QUADRO_DC_ESTAVEL=S THEN MFP=S
4	0.750	2.000	S	IF Fumo=PAROU AND ext_NroTerritorios_Index_cat=DOIS AND ext_residual_maxima<=1.00 AND IAM_ANT=QWAVE THEN MFP=S
4	0.667	1.000	S	IF NoCKMBPos>1.00 AND ext_residual_Media>4.00 AND ext_residual_maxima<=4.00 AND HistoriaFamiliar_DCA=S THEN MFP=S
4	0.750	2.000	S	IF ext_lesao_classif=A AND ext_vaso_diam_medio<=2.00 AND ext_vaso_diam_menor>2.00 AND STENT_GERACAO=NOVA THEN MFP=S
4	0.750	2.000	S	IF ext_lesao_classif=B1 AND FVE=MODERADO_GRAVE AND IMC<=23.00 AND Quadro_Chegada=ISQUEMIA THEN MFP=S
3	0.667	1.000	S	IF OneTypeStent_Index=XCIENCE AND ext_residual_Media>11.00 AND ext_residual_maxima<=16.00 THEN MFP=S
4	0.667	1.000	S	IF STENT_OVERLAP=S AND ext_lesao_classif=B1 AND ext_vaso_diam_medio>3.00 AND Stent_Compr_Total_Index<=49.00 THEN MFP=S
4	0.667	1.000	S	IF OneTypeStent_Index=CYPHER AND DIABETES=SIM_IDep AND ST_elevation=N AND IAM_ANT=NOQWAVE THEN MFP=S
5	0.750	2.000	S	IF Fumo=PAROU AND ext_calcio=ALTO AND ext_vaso_diam_medio<=2.00 AND IDADE>67.00 AND Quadro_Chegada=ISQUEMIA THEN MFP=S
4	0.667	1.000	S	IF OneTypeStent_Index=TAXUS AND HAS=N AND IMC>28.00 AND IMC<=28.00 THEN MFP=S
3	0.667	1.000	S	IF ext_NroLesions_Index>1.00 AND CLEARANCE<=44.00 AND CLEARANCE>44.00 THEN MFP=S
3	0.667	1.000	S	IF OneTypeStent_Index=CYPHER AND CLEARANCE>79.00 AND CLEARANCE<=80.00 THEN MFP=S
4	0.667	1.000	S	IF ext_NroTerritorios_Index>1.00 AND KILLIP=I AND STENT_GERACAO=MIX AND Stent_Compr_Total_Index>57.00 THEN MFP=S
4	0.667	1.000	S	IF OneTypeStent_Index=CYPHER AND ext_lesao_classif=B1 AND Clopidogrel=N AND DIABETES=SIM_IDep THEN MFP=S
5	0.667	1.000	S	IF Fumo=PAROU AND ext_vaso_diam_medio<=2.00 AND TipoAngInstavel=IIB AND ext_vaso_diam_medio>2.00 AND Stent_Compr_Total_Index<=36.00 THEN MFP=S
4	0.667	1.000	S	IF ext_NroVasos_Index_cat=1 AND IDADE<=57.00 AND IAM_ANT=NOQWAVE AND lesao_longa=MIX THEN MFP=S
4	0.667	1.000	S	IF ext_residual_Media>2.00 AND ext_vaso_diam_medio>2.00 AND STENT_GERACAO=MIX AND ext_bifurcacao=S THEN MFP=S
5	0.667	1.000	S	IF Fumo=PAROU AND Estatina=N AND FVE=MODERADO_GRAVE AND CLEARANCE>88.00 AND Stent_Compr_Total_Index>18.00 THEN MFP=S
4	0.667	1.000	S	IF IDADE<=53.00 AND ext_vaso_diam_medio>3.00 AND ext_vaso_diam_medio<=3.00 AND Stent_Compr_Total_Index<=16.00 THEN MFP=S
5	0.750	2.000	S	IF ext_lesoes_compr_tot<=26.00 AND Fumo=PAROU AND ext_lesoes_compr_tot>24.00 AND IDADE>74.00 AND Quadro_Chegada=ISQUEMIA THEN MFP=S
4	0.667	1.000	S	IF OneTypeStent_Index=CYPHER AND ext_lesoes_compr_tot<=20.00 AND ext_lesoes_compr_tot>20.00 AND Quadro_Chegada=ANG_EST THEN MFP=S
4	0.667	1.000	S	IF OneTypeStent_Index=CYPHER AND ext_vaso_diam_menor>3.00 AND CLEARANCE>70.00 AND CLEARANCE<=71.00 THEN MFP=S
6	0.750	2.000	S	IF ext_lesao_classif=B2 AND ext_residual_Media>4.00 AND ext_vaso_diam_menor>3.00 AND CLEARANCE>84.00 AND Tipo_angina=N AND Stent_Compr_Total_Index>12.00 THEN MFP=S
4	0.667	1.000	S	IF TipoAngInstavel=IIB AND OneTypeStent_Index=TAXUS AND CKMB_POS=N AND Stent_Compr_Total_Index>24.00 THEN MFP=S
4	0.667	1.000	S	IF ext_residual_maxima<=1.00 AND ext_residual_Media>1.00 AND TipoAngInstavel=IB AND ext_lesao_classif=A THEN MFP=S
5	0.750	2.000	S	IF ext_lesoes_compr_tot<=20.00 AND IMC>30.00 AND IMC<=30.00 AND Stent_Compr_Total_Index>15.00 AND CLEARANCE<=71.00 THEN MFP=S
4	0.667	1.000	S	IF ext_residual_maxima>8.00 AND Stent_Compr_Total_Index<=13.00 AND DIABETES=SIM_NIDep AND Stent_Compr_Total_Index>12.00 THEN MFP=S
3	0.667	1.000	S	IF Fumo=S AND CKMB_PRE=N AND IAM_ANT=NOQWAVE THEN MFP=S
3	0.667	1.000	S	IF CLEARANCE>106.00 AND CLEARANCE<=106.00 AND Fumo=S THEN MFP=S
3	0.667	1.000	S	IF ext_residual_maxima<=1.00 AND CLEARANCE>111.00 AND CLEARANCE<=111.00 THEN MFP=S
4	0.667	1.000	S	IF Quadro_Chegada=ISQUEMIA AND IMC>32.00 AND CLEARANCE<=48.00 AND CLEARANCE>42.00 THEN MFP=S
4	0.667	1.000	S	IF ext_residual_maxima<=1.00 AND IDADE<=49.00 AND IMC>33.00 AND IMC<=33.00 THEN MFP=S
6	0.750	2.000	S	IF ext_residual_Media>4.00 AND CLEARANCE>80.00 AND IDADE>65.00 AND IMC<=26.00 AND OneTypeStent_Index=CYPHER AND CLEARANCE<=87.00 THEN MFP=S
4	0.667	1.000	S	IF TipoAngInstavel=IIB AND ext_vaso_diam_medio<=2.00 AND Estatina=N AND ext_lesao_classif=B1 THEN MFP=S
3	0.996	240.000	N	IF CKMB_POS=NORMAL AND ext_residual_Media<=0.00 AND CLEARANCE<=111.00 THEN MFP=N
5	0.997	318.000	N	IF NoCKMBPos_cat=IN_0_OUT_0 AND NoTropoPos>0.00 AND ext_calcio=MODERADO AND CLEARANCE>64.00 AND CLEARANCE<=130.00 THEN MFP=N
4	0.993	147.000	N	IF NoCKMBPos_cat=IN_0_OUT_0 AND NoTropoPos>4.00 AND ext_residual_maxima>1.00 AND NoTropoPos<=19.00 THEN MFP=N
3	0.989	85.000	N	IF CKMB_POS=N AND OneTypeStent_Index=TAXUS AND IMC<=30.00 THEN MFP=N
3	0.991	115.000	N	IF NoCKMBPos<=2.00 AND ext_lesao_classif=B1 AND CLEARANCE>34.00 THEN MFP=N
3	0.988	80.000	N	IF CKMB_POS=NORMAL AND ext_lesao_classif=A AND ext_residual_Media>1.00 THEN MFP=N
3	0.987	77.000	N	IF NoTropoPos<=1.00 AND Tipo_angina=N AND ext_trombo=N THEN MFP=N
4	0.993	148.000	N	IF NoCKMBPos<=2.00 AND Quadro_Chegada=ISQUEMIA AND ext_residual_Media<=3.00 AND IMC>25.00 THEN MFP=N
5	0.993	145.000	N	IF CKMB_POS=NORMAL AND STENT_OVERLAP=S AND FVE=NORMAL_DISCRETO AND HistoriaFamiliar_DCA=N AND IMC>20.00 THEN MFP=N
3	0.986	69.000	N	IF CKMB_POS=NORMAL AND OneTypeStent_Index=PROMUS AND IDADE<=68.00 THEN MFP=N
4	0.991	108.000	N	IF CKMB_POS=NORMAL AND OneTypeStent_Index=XCIENCE AND IDADE<=56.00 AND ext_lesoes_compr_tot<=56.00 THEN MFP=N
5	0.994	177.000	N	IF CKMB_POS=NORMAL AND ext_lesoes_compr_tot<=11.00 AND Fumo=N AND OneTypeStent_Index=CYPHER AND CLEARANCE<=121.00 THEN MFP=N
3	0.987	77.000	N	IF CKMB_POS=N AND TipoAngEstavel=CCS2 AND ext_vaso_diam_menor<=2.00 THEN MFP=N
3	0.988	82.000	N	IF CKMB_POS=N AND Quadro_Chegada=ISQUEMIA AND ext_vaso_diam_menor>2.00 THEN MFP=N
3	0.982	55.000	N	IF NoCKMBPos_cat=IN_0_OUT_0 AND TERRITORIO_TCE=S AND ext_residual_Media>1.00 THEN MFP=N
5	0.990	94.000	N	IF NoCKMBPos<=2.00 AND ext_calcio=MODERADO AND ext_NroLesions_Index<=2.00 AND SEXO=M AND ext_territ_CD=N THEN MFP=N
6	0.995	206.000	N	IF CKMB_POS=NORMAL AND Estatina=N AND Fumo=N AND IDADE<=75.00 AND ext_residual_Media<=4.00 AND ext_vaso_diam_menor>2.00 THEN MFP=N
5	0.993	149.000	N	IF CKMB_POS=NORMAL AND ext_vaso_diam_medio>2.00 AND IMC>28.00 AND CLEARANCE>89.00 AND IDADE>39.00 THEN MFP=N
4	0.991	108.000	N	IF NoCKMBPos_cat=IN_0_OUT_0 AND ext_bifurcacao=S AND IMC<=34.00 AND ext_lesoes_compr_tot<=40.00 THEN MFP=N
5	0.993	132.000	N	IF CKMB_POS=NORMAL AND Isquemia_Silenciosa=S AND ext_territ_CD=N AND HAS=N AND NROSTENTS_Index<=3.00 THEN MFP=N
4	0.981	51.000	N	IF CKMB_POS=N AND SEXO=F AND ext_vaso_diam_medio>2.00 AND IMC<=34.00 THEN MFP=N
3	0.980	49.000	N	IF CKMB_POS=NORMAL AND NoCKMBPre<=6.00 AND ext_lesao_classif=C THEN MFP=N
4	0.983	58.000	N	IF NoCKMBPos<=2.00 AND IMC<=24.00 AND STENT_GERACAO=NOVA AND Angina_Instavel=N THEN MFP=N
6	0.992	130.000	N	IF CKMB_POS=NORMAL AND FVE=NORMAL_DISCRETO AND ext_NroTerritorios_Index_cat=DOIS AND CLEARANCE<=89.00 AND ext_vaso_diam_medio>2.00 AND IDADE<=86.00 THEN MFP=N
4	0.978	44.000	N	IF CKMB_POS=N AND ext_NroTerritorios_Index_cat=DOIS AND ext_lesoes_compr_tot<=48.00 AND ext_vaso_diam_menor<=2.00 THEN MFP=N
3	0.972	34.000	N	IF NoCKMBPos<=2.00 AND ext_lesoes_compr_tot>63.00 AND Stent_Compr_Total_Index<=86.00 THEN MFP=N
3	0.977	41.000	N	IF NoCKMBPre>3.00 AND ext_residual_Media<=1.00 AND ext_lesoes_compr_tot>9.00 THEN MFP=N
6	0.990	100.000	N	IF CKMB_POS=NORMAL AND ext_calcio=NAO AND HistoriaFamiliar_DCA=S AND ext_vaso_diam_menor>2.00 AND ext_vaso_diam_medio<=3.00 AND IDADE>36.00 THEN MFP=N
5	0.977	42.000	N	IF CKMB_POS=NORMAL AND ext_territ_DA=S AND ext_territ_CD=S AND Quadro_Chegada=ANG_INST AND IDADE>46.00 THEN MFP=N
7	0.987	74.000	N	IF CKMB_POS=NORMAL AND FVE=NORMAL_DISCRETO AND CLEARANCE>89.00 AND STENT_GERACAO=VELHA AND ext_territ_DA=S AND ext_vaso_diam_medio>2.00 AND ext_residual_Media<=5.00 THEN MFP=N
4	0.977	42.000	N	IF NoCKMBPos<=2.00 AND ext_vaso_diam_medio_fino=N AND lesao_longa=N AND Fumo=N THEN MFP=N
4	0.981	52.000	N	IF TipoAngEstavel=CCS3 AND ext_lesao_pos_dilat=N AND ext_residual_Media>1.00 AND IDADE>36.00 THEN MFP=N
7	0.993	136.000	N	IF CKMB_POS=NORMAL AND ext_territ_CD=N AND IMC<=25.00 AND IDADE<=78.00 AND ext_residual_maxima>2.00 AND ext_residual_maxima<=7.00 AND ext_vaso_diam_medio>2.00 THEN MFP=N
5	0.988	80.000	N	IF NoCKMBPos_cat=IN_0_OUT_0 AND ext_vaso_diam_medio<=2.00 AND IDADE>62.00 AND IDADE<=74.00 AND DoencaVascPeriferica=N THEN MFP=N
3	0.963	25.000	N	IF NoCKMBPos<=2.00 AND Fumo=S AND Stent_Compr_Total_Index>13.00 THEN MFP=N
6	0.982	55.000	N	IF CKMB_POS=NORMAL AND FVE=NORMAL_DISCRETO AND ext_calcio=NAO AND CLEARANCE<=87.00 AND ext_lesao_pos_dilat=N AND ext_vaso_diam_medio>2.00 THEN MFP=N
5	0.981	51.000	N	IF NoCKMBPos<=2.00 AND ext_NroTerritorios_Index_cat=DOIS AND ext_vaso_diam_medio>2.00 AND

Figura 51 - Terceira parte do modelo de classificação pelo atributo MFP, composto por 6 partes. (algoritmo CN2).

Em resumo, foram apresentados os resultados do processo de KDD aplicado ao banco de dados do Registro Desire, incluindo a fase de pré-processamento com o tratamento das variáveis, a preparação do arquivo ARFF, a mineração dos dados, a formatação e apresentação dos modelos de saída.

Utilizando o algoritmo J4.8 da ferramenta Weka, foi possível gerar, pela técnica da árvore de decisão, modelos de classificação interpretáveis e supervisionado pelo atributo-classe, que representa a ocorrência de um evento cardíaco adverso a estes pacientes.

Tres *datasets* foram criados onde o tempo e a fase da ocorrência do evento foram os elementos variantes.

Diferentes recursos da ferramenta foram utilizados na mineração destes *datasets* para gerar os modelos, com características que permitissem uma melhor análise, conduzida por métricas objetivas e apropriada para a avaliação de classes binárias. Dentre estes recursos, fez-se o uso de metaclassificadores, filtros e correção do desbalanceamento pelo método de redução pela menor classe.

Dos modelos de árvore de decisão apresentados, extraíram-se as principais regras de classificação para a classe MACE (eventos ao longo de 13 anos de acompanhamento) e MFP (eventos até o primeiro ano), selecionando o modelo P-2 e P-5, respectivamente, e, para a classe, MACE_REFERENCIA (eventos até o primeiro ano excluindo o IAM_NQ hospitalar) selecionou-se o modelo P-8.

Com o objetivo de realizar uma investigação junto aos usuários, no sentido de avaliar os resultados alcançados, estas regras foram apresentadas a 5 especialistas em hemodinâmica, dos quais dois clínicos e três intervencionistas, para que fosse indicado subjetivamente o grau de novidade e de concordância para cada uma delas.

Para o grau de novidade, seguiu-se o critério utilizado por Carvalho⁴⁸ em sua tese sobre generalização de regras de associação, que foi definido como: (0) A regra não faz sentido; (1) Regra esperada; (2) Regra não tão esperada; (3) Regra nada esperada; e a concordância leva em conta o quanto ao conteúdo da regra não conflita com sua experiência, cuja escala varia entre 100%, 75%, 50%, 25% e 0%, em que 100% representa a concordância plena.

A cada especialista foi apresentado um relatório de avaliação composto por quatro páginas, em que as três primeiras se referem às regras obtidas das árvores de decisão e, na última, foram apresentadas quatro regras de classificação geradas pelo algoritmo Orange-CN2. No Anexo D, encontram-se as avaliações dos especialistas às regras.

As Figuras 52, 53 e 54 apresentam o resumo das avaliações para as regras extraídas dos modelos gerados pela técnica da árvore de decisão. Na última linha de cada tabela, é apresentada a Moda, indicando a classificação mais frequente dentre os especialistas para cada regra apresentada, tanto para o grau de novidade quanto para a concordância.

Modelo P-2																							
Regra nº	1		2		3		4		5		6		7		8		9		10		11		
Especialista	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	
1	1	100	2	75	2	50	2	50	2	75	1	100	1	100	3	0	1	75	2	50	2	50	
2	1	75	1	75	1	50	1	50	1	75	1	100	1	75	2	50	1	75	1	75	1	50	
3	2	100	2	75	1	100	2	50	2	50	1	100	1	100	1	100	1	100	1	100	2	0	
4	1	100	1	100	1	100	0	0	1	100	1	100	1	100	1	75	1	100	1	100	0	0	
5	2	75	2	75	3	50	3	25	2	75	1	100	1	100	1	50	2	50	3	0	2	75	
Moda	1	100	2	75	1	50	2	50	2	75	1	100	1	100	1	50	1	75*	1	100	2	0*	
GN=Grau Novidade																							
C%=Concordância com a regra																							
* Empate na moda, selecionou-se o menor valor																							

Figura 52 - Quadro-resumo das avaliações para as regras do modelo P-2, que considera a ocorrência de eventos ao longo dos 13 anos de acompanhamento.

Modelo P-5														
Regra nº	1		2		3		4		5		6		7	
Especialista	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %
1	1	100	1	75	1	100	2	50	2	50	3	25	2	25
2	1	100	1	75	1	75	2	50	1	50	2	50	2	25
3	1	100	1	100	2	75	2	50	1	100	3	0	2	100
4	1	100	1	100	1	100	0	0	1	100	0	0	1	100
5	1	100	1	0**	1	100	1	75	2	50	2	50	3	50
Moda	1	100	1	75	1	100	2	50	1	50	2*	0	2	50
** Especialista questionou valor da probabilidade da regra														

Figura 53 - Quadro-resumo das avaliações para as regras do modelo P-5, que considera a ocorrência de eventos no 1º ano de acompanhamento.

Modelo P-8												
Regra nº	1		2		3		4		5		6	
Especialista	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %	GN	C %
1	1	100	2	50	0	0	2	50	1	75	1	100
2	1	50	1	75	1	50	2	25	1	75	1	75
3	1	100	2	75	3	25	1	100	2	75	3	0
4	1	50	1	75	0	0	1	75	1	100	1	100
5	1	100	2	50	3	25	3	25	2	50	2	75
Moda	1	100	2	75	0	0	1	25	1	75	1	75

Figura 54 - Quadro-resumo das avaliações para as regras do modelo P-8, que considera a ocorrência de eventos no 1º ano de acompanhamento e exclui os infartos não Q da fase hospitalar.

A Figura 55 apresenta o resultado da avaliação para as quatro regras extraídas do modelo gerado pela Regra de Classificação.

Orange CN2								
Regra nº	1		2		3		4	
Especialista	GN	C %	GN	C %	GN	C %	GN	C %
1	1	100	1	100	1	100	-	-
2	1	-	1	-	1	-	1	-
3	1	100	2	75	1	100	3	50
4	1	100	1	100	1	100	1	100
5	1	100	1	75	2	50	1	75
Moda	1	100	1	75	1	100	1	50

Figura 55 - Quadro-resumo das avaliações para as regras de classificação, que considera a ocorrência de eventos no 1º ano de acompanhamento.

Vinte foram as regras geradas sem novidade $GN=1$, representando 64% para o modelo P-2, 44% para o modelo P-5 e 56% para o modelo P-8 e 100% para o modelo CN2.

Três regras foram desconsideradas por não fazer sentido (regra 11 modelo P-2, regra 6 do modelo P-5 e regra 3 do modelo P-8).

Nove regras foram classificadas como não tão esperadas $GN=2$ e tiveram grau de concordância igual ou maior do que 50%, representando 27% do modelo P-2, 44% para o modelo P-5 e 22% para o modelo P-8. Estas são regras candidatas para serem analisadas, em que uma avaliação mais detalhada do grupo de pacientes envolvidos e dos atributos relacionados é sugestiva de uma investigação mais profunda.

5 CONCLUSÃO

5 CONCLUSÃO

O objetivo deste trabalho foi a indução e análise de modelos de classificação supervisionados para ocorrência de um evento cardíaco adverso nos pacientes tratados unicamente com stent farmacológico, cuja população de estudo foram os pacientes do Registro Desire.

Assim como nos resultados descritos pela revisão bibliográfica, este trabalho forneceu regras que expressam as relações entre as variáveis mais significativas do Desire segundo o algoritmo de classificação selecionado.

O atributo-classe para estes modelos foi a ocorrência ou não de um evento cardíaco maior, o que caracteriza uma classificação binária em que as métricas de avaliação se baseiam no resultado apresentado pela matriz de confusão ou matriz de contingência. Medidas como acurácia, sensibilidade, especificidade, índice kappa e área sob a curva ROC foram utilizadas para uma avaliação objetiva destes modelos.

Ao longo do trabalho, a evolução no conhecimento das ferramentas de mineração de dados permitiu que o processo iterativo do KDD fosse reduzido, pois, ao final, com apenas um arquivo (extensão: ARFF) e aplicando dentro da ferramenta diversos recursos disponíveis, foi possível gerar os modelos.

A interação com os especialistas médicos facilitou a compreensão dos detalhes da rotina e as situações diretamente associadas ao procedimento, o que direcionou a definição dos modelos, a definição dos valores de alguns atributos e a escolha do atributo-classe em função do período de observação dos eventos.

Os modelos estudados classificaram os pacientes segundo a ocorrência do primeiro evento cardíaco, e pelo tempo decorrido entre a data do procedimento índice e a data da ocorrência do evento.

- O primeiro modelo considerou os eventos ocorridos na fase hospitalar e ao longo dos 13 anos de acompanhamento dos pacientes.
- O segundo considerou a fase hospitalar e reduziu o período pós-hospitalar para um ano.
- O terceiro foi igual ao segundo, mas retirou os infartos periprocedimento do modelo, ou seja, excluiu os pacientes com evento do tipo IAM_NQ da fase hospitalar e adicionou, na fase pós-hospitalar, um segundo evento destes pacientes, desde que ocorrido dentro do primeiro ano.

Assim, os modelos cobriram diferentes situações e forneceram os dados para uma análise objetiva de desempenho destes e, também, para uma avaliação subjetiva pelos especialistas de domínio, por meio das regras de classificação extraídas indiretamente das árvores de decisão e diretamente do algoritmo CN2.

Entre o primeiro e segundo modelo, foi possível perceber que a influência das variáveis sobre o evento cardíaco adverso, coletadas por ocasião do tratamento, diminuía com o passar do tempo. O segundo modelo que considera os eventos até o primeiro ano, embora mais desbalanceado com uma redução na classe positiva de 25,7% para 12,7%, apresentou uma melhora significativa em todos os índices, a acurácia passou de 81% para 94%, a sensibilidade de 36 para 58% a AUC de 0,78 para 0,84 e o índice kappa de 0,40 para 0,69, o que indica uma substancial correlação entre a classe e as variáveis independentes.

Ficou evidente a interferência da instrumentação no local da lesão para a recanalização do vaso, que está intimamente associada ao registro de eventos do tipo IAM_NQ.

A criação do terceiro modelo foi para eliminar o efeito deste tipo de evento, que é intrínseco ao procedimento da intervenção e cujo critério de detecção é dado pela CKMB pós-procedimento estar elevada mais de 5 vezes em relação ao pré-procedimento. A maioria destes infartos não tem

repercussão clínica e está diretamente relacionada à manipulação durante o procedimento, como o acesso dos cateteres ao local da lesão para sua desobstrução que ocasionam interrupções intermitentes do fluxo coronariano e são captadas pelos biomarcadores.

Embora este novo modelo, identificado como Modelo P-8 (vide Anexo C), tenha apresentado um desbalanceamento ainda maior da classe positiva, com apenas 5,3 % dos pacientes, ter sido considerado mais fraco pela análise objetiva, foi possível verificar o aparecimento de outras variáveis com maior significância ao se eliminar a ocorrência de IAM_NQ da fase hospitalar.

Posteriormente, induziu-se um quarto e quinto modelo de classificação, agora segundo o tipo de evento ocorrido, com base no atributo-classe [MFP_FIRST], modelo P-10 ao longo do primeiro ano, e [MACE_FIRST], modelo P-11 ao longo dos treze anos, em que ficou clara a polarização dos eventos IAM_NQ em ambos os modelos.

O desbalanceamento entre as classes positiva e negativa nos percentuais de 25%, 12,7% e 5,3% foi um fator restritivo na indução dos modelos apresentados e utilizou-se a técnica de correção do desbalanceamento pelo método de redução pela menor classe.

Constatou-se que a redução das variáveis pelos metaclassificadores, selecionando os 20 atributos mais significativos para a mineração de dados, não alterou significativamente o poder dos modelos.

Os pacientes podem ter recebido influência de variáveis que não foram contempladas pelo sistema do Registro Desire, como outros resultados laboratoriais e outras informações hospitalares.

Finalmente, foram apresentadas as regras extraídas dos modelos P-2, P-5, P-8 e do modelo gerado pelo algoritmo CN2, aos especialistas do domínio buscando identificar por meio de uma avaliação subjetiva o grau de novidade.

Conclui-se que os modelos foram capazes de gerar 63% das regras sem novidade aos especialistas atestando coerência com seus conhecimentos, 9% delas foram rejeitadas e 28% foram avaliadas como não

tão esperadas, denotando serem candidatas a uma análise mais profunda por parte dos especialistas.

Explorar a propriedade de interpretabilidade da árvore de decisão, com o objetivo de investigar uma eventual importância deste grupo de regras, é uma das propostas desta metodologia, a de sugerir novas hipóteses para serem investigadas.

Carvalho e cols.⁴⁹, em uma revisão sistemática de 18 publicações com o objetivo de avaliar a efetividade do uso da Mineração de Dados na área da Saúde, mostram a tarefa de classificação, pela técnica da árvore de decisão, como a mais utilizada pela clareza da apresentação dos atributos discriminadores e relata, ainda, um trabalho em que as regras que denotaram aleatoriedade foram consideradas as mais relevantes pelos especialistas, devido à combinação dos atributos apresentados.

A proposta apresentada neste trabalho foi oferecer um modelo que apresente as relações existentes entre suas variáveis mais preditivas, extraídas de modo automático pelo algoritmo de classificação selecionado, que seja descritivo e auxilie na determinação da ocorrência de um evento entre os pacientes do Registro Desire.

A experiência de aplicar a metodologia do KDD ao Registro Desire permitiu a passagem de etapa por etapa do processo, e nas soluções de seus detalhes e desafios. Ficou patente que, para se chegar a um novo conhecimento, é fundamental a participação do especialista em todas as etapas, principalmente na fase de pós-processamento, que é a fase de validação da descoberta e, sem esta colaboração, fica impossível a implantação desta metodologia.

Devido às ferramentas disponíveis que auxiliam o processo de mineração, estes modelos podem ser refeitos sempre que necessário, uma vez que eles se baseiam nas informações armazenadas em banco de dados e que são continuamente atualizadas.

O potencial dos modelos interpretáveis é grande dentro do processo de tomada de decisão na área da saúde quando aliado à experiência do profissional e vem contribuir à Medicina baseada em evidência⁵⁰.

Lembrando que o objetivo principal deste trabalho foi a aplicação da metodologia de KDD ao banco de dados do Registro Desire e obter um modelo descritivo que classifique os pacientes quanto ao risco de um evento cardíaco, considera-se que se chegou, plenamente, a este objetivo devido aos resultados obtidos.

O objetivo secundário, que foi extrair as principais regras e avaliar junto aos especialistas do domínio foi atendido, pois, conforme comentado neste capítulo, os modelos geraram regras conhecidas e outras com certo grau de novidade sugestiva de investigação.

6 ANEXOS

6.1 ANEXO A - Lista dos atributos do Registro Desire extraídos da base de dados com suas definições e descrições

ANTECEDENTES		
CAMPO_ARFF	VALORES DOS ATRIBUTOS	DESCRIÇÃO
idPatients	NUMERIC	
IDADE	NUMERIC	
FAIXA ETARIA	JOVEM,QUARENT,CINQUENT,SEXAG,SEPTAG,OCTOG	
EuroScore	NUMERIC	
ext_EuroScoreLogistic	NUMERIC	
EuroScoreLogistic_cat	BAIXO,MODERADO,ALTO	Baixo (índice de 0 a 5) / Moderado (de 6 a 10) / Alto (> 10)
SEXO	M,F	
IMC	NUMERIC	
IMC_CAT	ABAIXO,NORMAL,ACIMA_PESO,OBESO,OBES_SEV,OBES_MORB	
CLEARANCE	NUMERIC	Homens: $(\text{Peso} \times (140 - \text{idade})) / 72 \times \text{Creatinina}$ e Mulheres: $0,85 \times (\text{Peso} \times (140 - \text{idade})) / 72 \times \text{Creatinina}$
FUNC RENAL	NORMAL,DISCRETA,MODERADA,GRAVE	Normal(clear ≥ 90), Discreta($60 \leq \text{clear} < 90$), Moderada ($30 \leq \text{clear} < 60$), Grave (clear < 30)
HISTORIA_CV	S,N	Histórico Cardiovascular (IAM,Revascularização por CABG ou PTCA, doença vascular periférica, AVC)
IAM_ANT	NO,QWAVE,NOQWAVE	IAM anterior
REVASC_ANT	S,N	IAM
CVA	S,N	AVC anterior
DoencaVasoPeriferica	S,N	Doença vascular periférica anterior
FATOR_RISCO	S,N	Fator Risco (Hist familiar de DCA, HAS, Hipertensão, fumo, diabetes)
HistoriaFamiliar_DCA	S,N	Histórico familiar de doença coronariana
HAS	S,N	Hipertensão
Hipertensão	S,N	Hipertensão
Fumo	N,PAROU,S	Fumo
DIABETES	N,SIM_NIDep,SIM_MED,SIM_IDep	Diabetes

Figura 56 - Desire, atributos relativos aos antecedentes.

PROCEDIMENTO		
CAMPO_ARFF	VALORES DOS ATRIBUTOS	DESCRIÇÃO
idProcedure	NUMERIC	Sem efeito
DisplayProcedureNumber	NUMERIC	Sem efeito
Quadro Chegada	ISQUEMIA,ANG_EST,ANG_INST,RECENT_MI	Classificação do paciente segundo quadro clínico
QUADRO_DC_ESTAVEL	S,N	Isquemia silenciosa ou angina estável
QUADRO_SCA	S,N	Infarto recente ou angina instável
Isquemia Silenciosa	S,N	Quadro clínico de Isquemia silenciosa
Angina Estável	S,N	Quadro clínico de Angina Estável
Angina Instável	S,N	Quadro clínico de Angina Instável
Infarto Recente	S,N	Quadro clínico de IAM
Tipo_angina	N,ESTAVEL,INSTAVEL	Quadro clínico de Angina
TipoAngEstavel	CCS1,CCS2,CCS3,CCS4	Quadro clínico de Angina Estável
TipoAngInstavel	IA,IIA,IIIA,IB,IIIB,IC,IIIC,IIIC	Quadro clínico de Angina Instável
TipoinfartoRecente	N,MENOS_24H,ENTRE_24_E_72H,MAIS_72H	Quadro clínico de Infarto recente
ST_elevation	S,N	Grupo: Infarto recente -> subgrupo com e sem elevação do ST
AgenteTrombolitico	S,N	Grupo: Infarto recente -> subgrupo que utilizou trombolítico (revascularização química)
FVE	NORMAL,DISCRETO,MODERADO,GRAVE,SEM_INFO	Importante, informa a capacidade de bombeamento
KILLIP	I,II,III,IV	para os pacientes com Infarto recente
Clopidogrel	S,N	Medicação na chegada
Estatina	S,N	Medicação na chegada (indica paciente com tratamento de dislipidemia)
CKMB_PRE	N,NORMAL,ANORMAL	Exame da enzima cardíaca na chegada
NroCKMBPre	NUMERIC	Dos resultados anormais, quantas vezes acima do valor normal
CKMB_POS	N,NORMAL,ANORMAL	Exame da enzima cardíaca no final do procedimento
NroCKMBPos	NUMERIC	Dos resultados anormais, quantas vezes acima do valor normal
SucessoProc	S,N	
SyntaxScore_cat	BAIXO,MODERADO,ALTO	Aplica-se somente para vasos nativos e multiarterial (baseia-se na angiografia)

Figura 57 - Desire, atributos relativos ao procedimento.

ANGIOGRÁFICOS		
CAMPO_ARFF	VALORES DOS ATRIBUTOS	DESCRIÇÃO
FUP_ANGIO	S,N	
ext_SomenteNativos_Index	S,N	Somente foram tratados vasos nativos no INDEX
ext_Ponte_Idade	NUMERIC	Identifica o tratamento de ao menos uma ponte e a que for a mais antiga. No INDEX
ext_Ponte_Idade_maior_10anos	S,N	Maior igual 10anos / Menor 10anos
ext_Ponte	NUMERIC	Informa o nº de pontes tratadas no INDEX
ext_NroLesions_Index	NUMERIC	Quantidade de lesões tratadas no INDEX
NroLesions_Index_cat	1,2,3,4,5,6,7	Quantidade de lesões tratadas no INDEX cat
ext_NroVasos_Index	NUMERIC	Quantidade de vasos tratados no INDEX
ext_NroVasos_Index_cat	1,2,3,4,5,6,7	Quantidade de vasos tratados no INDEX cat
ext_NroTerritorios_Index	NUMERIC	Quantidade de territórios tratados no INDEX
ext_NroTerritorios_Index_cat	HUM,DOIS,TRES,QUATRO	Quantidade de territórios tratados no INDEX cat
TERRITORIO_TCE	S,N	No procedimento INDEX a lesão tratada está no território TCE
ext_territ_DA	S,N	No procedimento INDEX a lesão tratada está no território DA
ext_territ_CX	S,N	No procedimento INDEX a lesão tratada está no território CX
ext_territ_CD	S,N	No procedimento INDEX a lesão tratada está no território CD
vaso_fino	S,N,MIX	Todas as lesões do INDEX com: (S) VesselDiameter <=2,5 mm; (N)-VesselDiameter >2,5 mm; (MIX)-Misturados
ext_vaso_diam_menor	NUMERIC	Vaso com diametro de menor valor no procedimento INDEX
ext_vaso_diam_medio	NUMERIC	Valor médio dos diâmetros. INDEX
ext_vaso_diam_medio_fino	S,N	Valor médio dos diâmetros. INDEX é considerado fino (x <= 2,75 mm) (Diam max <= 6,0 mm)
ext_lesao_classif	A,B1,B2,C	Das lesões tratadas pega a pior classificada
ext_trombo	S,N	Das lesões tratadas se alguma tiver trombo
ext_calcio	NAO,MODERADO,ALTO	Das lesões tratadas se alguma tiver cálcio
ext_residual_maxima	NUMERIC	(%) Valor da maior estenose residual. INDEX
ext_residual_Media	NUMERIC	(%) Valor da média das estenoses residuais. INDEX
ext_lesao_pos_dilat	S,N	Indica se ao menos uma lesão tratada no INDEX foi pós dilatada
ext_lesao_maior_30	S,N	Caso 1 vaso tratado com mais de 30 mm ou a soma dos lesões (multiarterial) maior do que 30 mm. INDEX
ext_lesoes_compr_tot	NUMERIC	Soma dos comprimentos das lesões tratadas. INDEX
LESÕES_NOVAS	S,N	No procedimento index todas as lesões são NOVAS
lesao_longa	S,N,MIX	Todas as lesões do procedimento index com: (N- Lesion lenght < 20 mm; S-Lesion lenght >=20 mm; MIX- Misturados)
ext_bifurcacao	S,N	

Figura 58 - Desire, atributos relativos aos dados angiográficos.

STENTS		
CAMPO_ARFF	VALORES DOS ATRIBUTOS	DESCRIÇÃO
NROSTENTS_Index	NUMERIC	Nº de stents implantados no procedimento INDEX
NROSTENTS_Total	NUMERIC	Nº de stents implantados no paciente considerando todos procedimentos
Stent_Compr_Total_Index	NUMERIC	Comprimento total dos stents implantados no procedimento INDEX
STENT_OVERLAP	S,N	Houve overlap de algum stent no procedimento INDEX
STENT_GERACAO	NOVA,MIX,VELHA	Há duas gerações de stents
OneTypeStent_Index	CYPHER,TAXUS,BALAO,ENDEAVOR,XCIENCE,PROMUS,BIOMATRIX,RESOLUT	Pacientes que no INDEX implantou o mesmo modelo de stent

Figura 59 - Desire, atributos relativos aos stents implantados.

EVENTOS ADVERSOS		
CAMPO_ARFF	VALORES DOS ATRIBUTOS	DESCRIÇÃO
TROMBOSE	S,N	
TROMBOSE_CLASS	NAO_HA,PROVAVEL,DEFINITIVA,POSSIVEL	
OBITO	N,CARDIACO,NAO_CARDIACO	
MACE_MORTE_CARDIACA	S,N	Evento de óbito cardíaco
MACE_IAM	S,N	Evento de IAM qualquer (Q ou NQ)
MACE_IAM_FASE	N,FASE_HOSP,FASE_FUP	Evento de IAM na fase hospitalar ou Fup
MACE_IAM_Q	S,N	Evento de IAM com onda Q [sp_Cyp_Corrige_EXT_Mace_IAM_Q]
MACE_IAM_NQ	S,N	Evento de IAM sem onda Q [sp_Cyp_Corrige_EXT_Mace_IAM_NQ]
MACE	S,N	EVENTO DE MACE QQ (há ao menos 1 evento)
TEMPO_DE_MACE	NUMERIC	Tempo decorrido para ocorrência do primeiro evento de MACE (>=0 ou ?)
TEMPO_DE_MACE_cat	6M,1A,2A,3A,4A,5A,6A,7A,8A,9A,10A	
MACE_REVASC	S,N	Evento de PTCA ou CABG de vaso novo ou já tratado
MACE_NOVALESAO	S,N	Evento de PTCA ou CABG de vaso novo
MACE_TLR	S,N	Evento de PTCA ou CABG de vaso previamente tratado
TEMPO_DE_MACE_TLR	NUMERIC	Tempo decorrido para ocorrência do primeiro evento de MACE TLR (>=0 ou ?)
TEMPO_DE_MACE_NOVALESAO	NUMERIC	Tempo decorrido entre o procedimento INDEX e o primeiro evento de revascularização de uma nova lesão após o INDEX
TEMPO_DE_MACE_NOVALESAO_cat	6M,1A,2A,3A,4A,5A,6A,7A,8A,9A,10A	
ext_tempo_mace_tlr	NUMERIC	Tempo decorrido entre o procedimento INDEX e o primeiro evento de revascularização de qualquer das lesões alvo tratadas
TEMPO_DE_MACE_TLR_cat	6M,1A,2A,3A,4A,5A,6A,7A,8A,9A,10A	
ext_tempo_morte_card	NUMERIC	Tempo em dias desde o INDEX até a morte de causa cardíaca
TEMPO_DE_MACE_MORTE_CARD_cat	6M,1A,2A,3A,4A,5A,6A,7A,8A,9A,10A	
ext_tempo_mace_IAM	NUMERIC	Tempo em dias desde o INDEX até o primeiro MACE IAM
TEMPO_DE_MACE_IAM_cat	6M,1A,2A,3A,4A,5A,6A,7A,8A,9A,10A	
ext_tempo_ini_fim_mace	NUMERIC	Intervalo de tempo entre o primeiro e último MACE
ext_tot_mace	NUMERIC	Nº total de MACES registrados do paciente
TOTAL_DE_MACE_cat	1,2,3,4,5,6,7,8,9	Nº total de MACES registrados do paciente
ext_progressao	S,N	Caso haja um segundo procedimento após 45 dias do index, considera-se progressão
ext_tot_proc	NUMERIC	Total de procedimentos do paciente [sp_Cyp_Corrige_ext_tot_proc]
TOTAL_DE_PROC_cat	1,2,3,4,5,6,7,8,9	

Figura 60 - Desire, atributos relativos aos eventos MACE.

6.2 ANEXO B - Dataset Desire: análise dos atributos categóricos

ANTECEDENTES			
FAIXA_ETARIA	%	DoencaVascPeriferica	%
JOVEM	1,32%	N	96,57%
ADULTO	67,27%	S	3,43%
IDOSO	31,41%		100,00%
	100,00%		
IMC_CAT	%	HistoriaFamiliar_DCA	%
ABAIXO	2,48%	N	66,93%
NORMAL_ACIMA	70,02%	S	33,07%
OBESO	27,51%		100,00%
	100,00%		
EuroScoreLogistic_cat	%	HAS	%
?	0,03%	N	23,55%
ALTO	9,55%	S	76,45%
BAIXO	79,65%		100,00%
MODERADO	10,77%		
	100,00%		
SEXO	%	Hiperlipidemia	%
F	22,05%	N	36,30%
M	77,95%	S	63,70%
	100,00%		100,00%
HISTORIA_CV	%	Fumo	%
N	51,04%	N	71,77%
S	48,96%	PAROU	19,63%
	100,00%	S	8,59%
IAM_ANT	%		100,00%
?	0,02%	DIABETES	%
NO	79,58%	N	71,51%
NOQWAVE	5,43%	SIM_NIDep	23,33%
QWAVE	14,98%	SIM_IDep	5,16%
	100,00%		100,00%
REVASC_ANT	%		
N	59,90%		
S	40,10%		
	100,00%		
CVA	%		
N	97,87%		
S	2,13%		
	100,00%		

Figura 61 - Desire, variáveis categóricas referentes aos antecedentes.

PROCEDIMENTO					
FUNC_RENAL	%			AgenteTrombolitico	%
?	8,04%			N	98,60%
MODERADA_GRAVE	24,10%			S	1,40%
NORMAL_DISCRETA	67,85%				100,00%
	100,00%				
Quadro_Chegada	%			FVE	%
?	0,78%			?	28,32%
ANG_EST	19,41%			MODERADO_GRAVE	5,74%
ANG_INST	25,00%			NORMAL_DISCRETO	65,94%
ISQUEMIA	37,15%				100,00%
RECENT_MI	17,66%				
	100,00%			KILLIP	%
				?	96,57%
				I	3,07%
				II	0,24%
QUADRO_DC_ESTAVEL	%			III	0,09%
N	42,65%			IV	0,03%
S	57,35%				100,00%
	100,00%				
QUADRO_SCA	%			Clopidogrel	%
N	57,35%			N	2,70%
SCA_CSST	4,69%			S	97,30%
SCA_SSST	37,96%				100,00%
	100,00%				
				Estatina	%
TipoAngEstavel	%	AE %		N	24,84%
?	80,59%	-		S	75,16%
CCS1	2,20%	11,31%			100,00%
CCS2	12,62%	65,02%			
CCS3	3,94%	20,27%		CKMB_PRE	%
CCS4	0,66%	3,39%		?	1,29%
	100,00%	100,00%		ANORMAL	9,60%
				N	6,15%
				NORMAL	82,97%
					100,00%
TipoAngInstavel	%	AI %			
?	75,00%	-		CKMB_POS	%
IA	0,39%	1,57%		?	0,20%
IB	3,14%	12,55%		ANORMAL	23,65%
IC	0,06%	0,25%		N	9,47%
IIA	0,27%	1,07%		NORMAL	66,68%
IIB	13,58%	54,33%			100,00%
IIC	0,22%	0,88%			
IIIA	0,08%	0,31%		NroCKMBPos_cat	%
IIIB	6,82%	27,29%		IN_0_OUT_0	73,11%
IIIC	0,44%	1,76%		IN_0_OUT_MAIOR_5	4,81%
	100,00%	100,00%		OUTRO	22,08%
					100,00%
TipoInfartoRecente	%	Infarto Recente %		SyntaxScore_cat	%
ENTRE_24_E_72H	4,01%	22,74%		?	86,48%
MAIS_72H	9,16%	51,87%		ALTO	6,12%
MENOS_24H	4,48%	25,40%		BAIXO	2,84%
N	82,34%	-		MODERADO	4,56%
	100,00%	100,00%			100,00%
ST_elevation	%	Infarto Recente %			
?	82,34%	-			
N	8,92%	50,53%			
S	4,69%	26,55%			
sem info	4,05%	22,91%			
	100,00%	100,00%			

Figura 62 - Desire, variáveis categóricas referentes ao procedimento.

ANGIOGRÁFICOS				STENTS	
FUP_ANGIO	%	ext_vaso_diam_medio_fino	%	NROSTENTS_Index	%
N	76,53%	?	2,78%	1	48,52%
S	23,47%	N	39,99%	2	31,74%
	100,00%	S	57,24%	3	13,45%
			100,00%	4	4,85%
ext_SomenteNativos_Index	%			5	1,13%
N	6,37%	ext_lesao_classif	%	6	0,28%
S	93,63%	?	12,20%	7	0,03%
	100,00%	A	3,07%		100,00%
		B1	19,63%	STENT_OVERLAP	%
ext_Ponte_Idade_maior_10anos	%	B2	21,30%	N	88,30%
?	94,78%	C	43,80%	S	11,70%
N	1,82%		100,00%		100,00%
S	3,40%	ext_trombo	%	STENT_GERACAO	%
	100,00%	N	95,95%	?	0,16%
ext_Ponte	%	S	4,05%	MIX	3,00%
?	93,63%		100,00%	NOVA	43,37%
1.00	4,97%	ext_calcio	%	VELHA	53,47%
2.00	1,18%	ALTO	4,74%		100,00%
3.00	0,19%	MODERADO	36,74%	OneTypeStent_Index	%
4.00	0,03%	NAO	58,52%	?	12,86%
	100,00%		100,00%	BIOMATRIX	4,44%
ext_NroLesions_Index	%	ext_lesao_pos_dilat	%	CYPHER	45,76%
1	55,04%	N	28,30%	ENDEAVOR	0,89%
2	30,47%	S	71,70%	PROMUS	3,94%
3	11,26%		100,00%	RESOLUT	10,22%
4	2,67%	ext_lesao_maior_30	%	TAXUS	4,31%
5	0,45%	N	63,42%	XCIENCE	17,58%
6	0,11%	S	36,58%		100,00%
	100,00%		100,00%		
ext_NroVasos_Index	%	LESOES_NOVAS	%		
1	60,28%	N	5,27%		
2	30,75%	S	94,73%		
3	7,65%		100,00%		
4	1,19%	lesao_longa	%		
5	0,11%	MIX	22,80%		
6	0,02%	N	53,72%		
	100,00%	S	23,47%		
ext_NroTerritorios_Index	%		100,00%		
1	68,87%	ext_bifurcacao	%		
2	26,92%	N	92,83%		
3	4,16%	S	7,17%		
4	0,05%		100,00%		
	100,00%				

Figura 63 - Desire, variáveis categóricas referentes aos dados angiográficos e stents.

EVENTOS ADVERSOS	
TROMBOSE	%
N	97,88%
S	2,12%
	100,00%
TROMBOSE_CLASS	%
NAO_HA	97,88%
DEFINITIVA	1,00%
POSSIVEL	1,02%
PROVAVEL	0,09%
	100,00%
MACE_MORTE_CARDIACA	%
N	94,68%
S	5,32%
	100,00%
MACE_IAM	%
N	91,77%
S	8,23%
	100,00%
MACE_IAM_NQ	%
N	94,57%
S	5,43%
	100,00%
MACE_IAM_Q	%
N	99,04%
S	0,96%
	100,00%
MACE_TLR	%
N	95,55%
S	4,45%
	100,00%
MACE_NOVALESAO	%
N	90,36%
S	9,64%
	100,00%
MACE	%
N	74,27%
S	25,73%
	100,00%

Figura 64 - Desire, variáveis categóricas referentes aos eventos.

Dataset Desire: análise dos atributos numéricos

IMC <> 0	Max	min	med	desvio	C.V
6250	52,21	15,8	28,08	4,31	0,15
CLEARANCE <> 0	Max	min	med	desvio	C.V
6055	287,9	1	81,45	31,41	0,39
NroCKMBPre <> 0	Max	min	med	desvio	C.V
616	337	0,4	9,20	26,17	2,85
NroCKMBPos <> 0	Max	min	med	desvio	C.V
1508	391,1	0,59	8,56	21,22	2,48
ext_Ponte_Idade <> 0	Max	min	med	desvio	C.V
333	35	0,2	11,49	5,79	0,50
ext_vaso_diam_menor <> 0	Max	min	med	desvio	C.V
6200	5,3	0,9	2,61	0,47	0,18
ext_vaso_diam_medio <> 0	Max	min	med	desvio	C.V
6200	5,3	1	2,74	0,43	0,16
ext_residual_maxima <> 0	Max	min	med	desvio	C.V
6097	100	0,1	4,07	5,75	1,41
ext_residual_Media <> 0	Max	min	med	desvio	C.V
6097	100	0,05	3,33	3,58	1,08
ext_lesoes_compr_tot <> 0	Max	min	med	desvio	C.V
6109	313	1	29,82	19,77	0,66
Stent_Compr_Total_Index <> 0	Max	min	med	desvio	C.V
6377	196	8	36,48	22,46	0,62

Figura 65 - Desire, variáveis numéricas onde o diâmetro e o comprimento estão em milímetros e o valor residual de estenose em porcentagem.

(CV: coeficiente de variação, calculado pela divisão entre o desvio padrão e a média. CV é uma medida de dispersão que estima a variabilidade dos dados em relação à média.)

6.3 ANEXO C - (Modelos P-1, P-2 e P-3) Classe: MACE

Classifier Model - P - 1	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0	1
STENT_GERACAO = NOVA: N (1765.65/110.0)	2
STENT_GERACAO = MIX: N (110.17/14.0)	3
STENT_GERACAO = VELHA	4
ext_SomenteNativos_Index = S	5
ext_residual_Media <= 3.6: N (1268.91/260.62)	6
ext_residual_Media > 3.6	7
HISTORIA_CV = S	8
TipolInfartoRecente = N	9
ext_NroVasos_Index <= 1	10
ext_residual_maxima <= 11.2	11
vaso_fino = S	12
Isquemia_Silenciosa = S: N (40.95/10.49)	13
Isquemia_Silenciosa = N	14
DoencaVascPeriferica = S: S (11.0/2.0)	15
DoencaVascPeriferica = N	16
NROSTENTS_Index <= 1	17
ext_lesao_pos_dilat = S: N (27.98/7.0)	18
ext_lesao_pos_dilat = N	19
ext_vaso_diam_menor <= 2.3: S (15.29/3.15)	20
ext_vaso_diam_menor > 2.3: N (36.18/12.84)	21
NROSTENTS_Index > 1: S (22.0/7.0)	22
vaso_fino = N	23
CVA = S: N (12.0/2.0)	24
CVA = N	25
lesao_longa = S	26
SEXO = M	27
REVASC_ANT = S	28
Stent_Compr_Total_Index <= 27: S (18.0/2.0)	29
Stent_Compr_Total_Index > 27	30
CLEARANCE <= 80.51: S (24.61/7.54)	31
CLEARANCE > 80.51: N (20.87/3.92)	32
REVASC_ANT = N: N (15.49/5.0)	33
SEXO = F: N (23.49/6.49)	34
lesao_longa = N	35
LESOES_NOVAS = S: N (172.93/51.97)	36
LESOES_NOVAS = N	37
Estatina = S: N (23.0/9.0)	38
Estatina = N: S (11.0/3.0)	39
lesao_longa = MIX: N (5.49/1.0)	40

Figura 66 - Modelo P-1: apresentação parcial (1/3) da árvore de decisão no modo texto.

Estatina = N: S (11.0/3.0)	39
lesao_longa = MIX: N (5.49/1.0)	40
vaso_fino = MIX: N (10.0/1.0)	41
ext_residual_maxima > 11.2: S (25.64/5.39)	42
ext_NroVasos_Index > 1: N (169.17/48.59)	43
TipolInfartoRecente = MENOS_24H: S (4.0)	44
TipolInfartoRecente = ENTRE_24_E_72H: S (14.0/6.0)	45
TipolInfartoRecente = MAIS_72H	46
IMC <= 25.14: S (17.64/4.32)	47
IMC > 25.14	48
ext_vaso_diam_menor <= 2.43: S (12.0/5.0)	49
ext_vaso_diam_menor > 2.43: N (25.36/3.68)	50
HISTORIA_CV = N	51
CKMB_PRE = N: N (29.71/4.0)	52
CKMB_PRE = NORMAL	53
Hiperlipidemia = S	54
ext_territ_CD = S	55
STENT_OVERLAP = S: N (10.0/2.0)	56
STENT_OVERLAP = N	57
ext_territ_CX = S: N (20.03/5.52)	58
ext_territ_CX = N	59
ext_calcio = NAO	60
FUNC_RENAL = NORMAL_DISCRETA	61
IDADE <= 63	62
ext_lesao_pos_dilat = S: S (14.5/6.18)	63
ext_lesao_pos_dilat = N: N (13.65/4.33)	64
IDADE > 63: S (10.0/1.0)	65
FUNC_RENAL = MODERADA_GRAVE: N (19.37/4.36)	66
ext_calcio = MODERADO: S (22.03/6.52)	67
ext_calcio = ALTO: N (1.0)	68
ext_territ_CD = N	69
TipolInfartoRecente = N	70
QUADRO_DC_ESTAVEL = S: N (131.07/30.52)	71
QUADRO_DC_ESTAVEL = N	72
IMC <= 28.53	73
ext_lesoes_compr_tot <= 14.85: N (12.0/3.0)	74
ext_lesoes_compr_tot > 14.85: S (21.61/6.61)	75
IMC > 28.53: N (21.42/3.0)	76
TipolInfartoRecente = MENOS_24H: S (4.0/1.0)	77
TipolInfartoRecente = ENTRE_24_E_72H: N (5.52/1.0)	78
TipolInfartoRecente = MAIS_72H: N (12.0/5.0)	79
Hiperlipidemia = N: N (254.24/56.07)	80
CKMB_PRE = ANORMAL: N (0.0)	81
ext_SomenteNativos_Index = N	82

Figura 67 - Modelo P-1: apresentação parcial (2/3) da árvore de decisão no modo texto.

Hiperlipidemia = N: N (254.24/56.07)	80
CKMB_PRE = ANORMAL: N (0.0)	81
ext_SomenteNativos_Index = N	82
ext_trombo = S: S (12.0/2.0)	83
ext_trombo = N	84
ext_lesao_pos_dilat = S: N (67.0/23.0)	85
ext_lesao_pos_dilat = N	86
ext_Ponte <= 1	87
ext_vaso_diam_medio_fino = S: N (19.0/7.0)	88
ext_vaso_diam_medio_fino = N: S (37.0/12.0)	89
ext_Ponte > 1: S (22.0/5.0)	90
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: S (307.0)	91
NroCKMBPos_cat = OUTRO	92
NroCKMBPos <= 2.98: N (937.89/143.65)	93
NroCKMBPos > 2.98	94
CKMB_PRE = N: S (15.1/3.03)	95
CKMB_PRE = NORMAL	96
ext_trombo = S: N (12.0/4.0)	97
ext_trombo = N: S (155.74/15.69)	98
CKMB_PRE = ANORMAL	99
NroCKMBPre <= 3	100
NroTropoPos <= 127.5	101
SEXO = M	102
DIABETES = N	103
ext_vaso_diam_medio_fino = S: S (30.29/11.22)	104
ext_vaso_diam_medio_fino = N: N (20.2/6.89)	105
DIABETES = SIM_NIDep: N (18.97/5.26)	106
DIABETES = SIM_IDep: N (1.04)	107
SEXO = F: N (15.32/2.79)	108
NroTropoPos > 127.5	109
Infarto_Recente = S: S (33.19/4.71)	110
Infarto_Recente = N: N (13.92/5.42)	111
NroCKMBPre > 3	112
STENT_GERACAO = NOVA: N (75.33/6.99)	113
STENT_GERACAO = MIX: N (2.67)	114
STENT_GERACAO = VELHA	115
ext_vaso_diam_medio_fino = S	116
ext_lesao_maior_30 = S: S (12.33/3.33)	117
ext_lesao_maior_30 = N	118
TipolInfartoRecente = N: N (4.0/1.0)	119
TipolInfartoRecente = MENOS_24H: N (13.67/4.33)	120
TipolInfartoRecente = ENTRE_24_E_72H: S (12.0/5.0)	121
TipolInfartoRecente = MAIS_72H: S (3.0/0.33)	122
ext_vaso_diam_medio_fino = N: N (31.34/6.34)	123
Number of Leaves : 70	
Size of the tree : 124	

Figura 68 - Modelo P-1: apresentação parcial (3/3) da árvore de decisão no modo texto.

Classifier Model - P - 2	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0	1
STENT_GERACAO = NOVA: N (1765.65/110.0)	2
STENT_GERACAO = MIX: N (110.17/14.0)	3
STENT_GERACAO = VELHA	4
ext_SomenteNativos_Index = S	5
HISTORIA_CV = S	6
ext_residual_Media <= 8.6: N (1248.68/384.92)	7
ext_residual_Media > 8.6	8
Isquemia_Silenciosa = S: N (31.0/14.36)	9
Isquemia_Silenciosa = N: S (65.51/26.79)	10
HISTORIA_CV = N: N (1283.99/285.0)	11
ext_SomenteNativos_Index = N	12
ext_trombo = S: S (12.0/2.0)	13
ext_trombo = N	14
FVE = NORMAL_DISCRETO	15
Isquemia_Silenciosa = S: N (42.43/13.46)	16
Isquemia_Silenciosa = N	17
QUADRO_SCA = SCA_CSST: N (1.0)	18
QUADRO_SCA = SCA_SSST: N (34.17/14.72)	19
QUADRO_SCA = N: S (30.15/11.46)	20
FVE = MODERADO_GRAVE: S (37.25/12.11)	21
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: S (307.0)	22
NroCKMBPos_cat = OUTRO	23
NroCKMBPos <= 2.98: N (937.89/143.65)	24
NroCKMBPos > 2.98	25
TipoInfartoRecente = N: S (221.06/59.72)	26
TipoInfartoRecente = MENOS_24H	27
IDADE <= 83	28
NroCKMBPos <= 32.78: N (88.16/16.82)	29
NroCKMBPos > 32.78: S (20.18/7.0)	30
IDADE > 83: S (12.0/2.0)	31
TipoInfartoRecente = ENTRE_24_E_72H	32
STENT_GERACAO = NOVA: N (39.34/5.67)	33
STENT_GERACAO = MIX: N (2.33/1.0)	34
STENT_GERACAO = VELHA	35
NroCKMBPos <= 4.47: S (18.16/5.6)	36
NroCKMBPos > 4.47: N (16.86/7.45)	37
TipoInfartoRecente = MAIS_72H	38
NroTropoPos <= 29.7: N (17.35/7.34)	39
NroTropoPos > 29.7: S (34.67/8.67)	40
Number of Leaves : 24	
Size of the tree : 41	

Figura 69 - Modelo P-2: Árvore de decisão no modo texto.

Classifier Model - P - 3	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0	1
STENT_GERACAO = NOVA: N (642.88/110.0)	2
STENT_GERACAO = MIX: N (50.07/14.0)	3
STENT_GERACAO = VELHA	4
REVASC_ANT = S: S (693.0/255.0)	5
REVASC_ANT = N	6
DoencaVascPeriferica = S: S (29.0/10.0)	7
DoencaVascPeriferica = N	8
ext_residual_Media <= 3.45	9
TipolInfartoRecente = N: N (310.41/104.68)	10
TipolInfartoRecente = MENOS_24H: S (3.0/1.0)	11
TipolInfartoRecente = ENTRE_24_E_72H: N (7.0/2.0)	12
TipolInfartoRecente = MAIS_72H	13
ext_residual_Media <= 2.13: N (16.46/3.23)	14
ext_residual_Media > 2.13: S (15.43/5.22)	15
ext_residual_Media > 3.45	16
ext_trombo = S: N (15.55/5.55)	17
ext_trombo = N	18
OneTypeStent_Index = CYPHER	19
Isquemia_Silenciosa = S	20
HISTORIA_CV = S: S (17.0/7.0)	21
HISTORIA_CV = N: N (94.11/40.28)	22
Isquemia_Silenciosa = N	23
Quadro_Chegada = ISQUEMIA: S (0.0)	24
Quadro_Chegada = ANG_EST: S (108.09/50.81)	25
Quadro_Chegada = ANG_INST	26
ext_residual_maxima <= 4.7: S (30.78/8.05)	27
ext_residual_maxima > 4.7: N (61.01/26.41)	28
Quadro_Chegada = RECENT_MI: S (52.32/20.62)	29
OneTypeStent_Index = TAXUS	30
CLEARANCE <= 75.56: N (31.68/14.15)	31
CLEARANCE > 75.56: S (21.99/5.36)	32
OneTypeStent_Index = BALAO: S (0.0)	33
OneTypeStent_Index = ENDEAVOR: N (4.21/1.14)	34
OneTypeStent_Index = XCIENCE: S (0.0)	35
OneTypeStent_Index = PROMUS: S (0.0)	36
OneTypeStent_Index = BIOMATRIX: S (0.0)	37
OneTypeStent_Index = RESOLUT: S (0.0)	38

Figura 70 - Modelo P-3: apresentação parcial (1/2) da árvore de decisão no modo texto.

OneTypeStent_Index = RESOLUT: S (0.0)	38
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: S (307.0)	39
NroCKMBPos_cat = OUTRO	40
NroCKMBPos <= 2.92	41
ext_trombo = S: S (19.58/5.74)	42
ext_trombo = N	43
STENT_GERACAO = NOVA: N (215.61/40.73)	44
STENT_GERACAO = MIX	45
REVASC_ANT = S: N (11.61/1.07)	46
REVASC_ANT = N: S (14.61/6.61)	47
STENT_GERACAO = VELHA	48
HISTORIA_CV = S	49
TipolInfartoRecente = N	50
ext_residual_maxima <= 2.5: N (22.64/6.55)	51
ext_residual_maxima > 2.5	52
ext_residual_Media <= 4.05: S (18.1/2.1)	53
ext_residual_Media > 4.05	54
CLEARANCE <= 68.74: S (13.29/4.1)	55
CLEARANCE > 68.74: N (10.69/2.5)	56
TipolInfartoRecente = MENOS_24H: S (1.0)	57
TipolInfartoRecente = ENTRE_24_E_72H: S (5.29/2.1)	58
TipolInfartoRecente = MAIS_72H: S (11.19/2.0)	59
HISTORIA_CV = N	60
QUADRO_SCA = SCA_CSST: S (3.19/0.55)	61
QUADRO_SCA = SCA_SSST	62
CLEARANCE <= 73.6: S (17.93/7.29)	63
CLEARANCE > 73.6: N (14.84/1.1)	64
QUADRO_SCA = N: N (43.04/14.0)	65
NroCKMBPos > 2.92	66
Infarto_Recente = S	67
TipolInfartoRecente = N: S (0.0)	68
TipolInfartoRecente = MENOS_24H	69
NroCKMBPos <= 6.2: N (17.06/3.32)	70
NroCKMBPos > 6.2	71
CLEARANCE <= 92.3: S (38.79/9.05)	72
CLEARANCE > 92.3: N (16.67/7.29)	73
TipolInfartoRecente = ENTRE_24_E_72H	74
STENT_GERACAO = NOVA: N (16.81/5.9)	75
STENT_GERACAO = MIX: S (1.0)	76
STENT_GERACAO = VELHA: S (28.23/7.16)	77
TipolInfartoRecente = MAIS_72H: S (40.04/5.52)	78
Infarto_Recente = N: S (189.79/27.27)	79
Number of Leaves : 50	
Size of the tree : 80	

Figura 71 - Modelo P-3: apresentação parcial (2/2) da árvore de decisão no modo texto.

(Modelos P-4, P-5 e P-6) Classe: MFP

Classifier Model - P-4	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0: N (4662.0/224.0)	1
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: S (307.0/6.0)	2
NroCKMBPos_cat = OUTRO	3
NroCKMBPos <= 2.98: N (937.89/59.33)	4
NroCKMBPos > 2.98	5
CKMB_PRE = N: S (15.1/3.03)	6
CKMB_PRE = NORMAL	7
TipolInfartoRecente = N: S (129.4/9.69)	8
TipolInfartoRecente = MENOS_24H: N (13.0/1.0)	9
TipolInfartoRecente = ENTRE_24_E_72H: N (8.33/4.0)	10
TipolInfartoRecente = MAIS_72H: S (17.0)	11
CKMB_PRE = ANORMAL	12
NroCKMBPre <= 3	13
NroTropoPos <= 65.29: N (71.33/19.02)	14
NroTropoPos > 65.29	15
STENT_GERACAO = NOVA: S (39.75/7.75)	16
STENT_GERACAO = MIX: N (0.15)	17
STENT_GERACAO = VELHA: N (21.7/5.55)	18
NroCKMBPre > 3: N (154.35/19.33)	19
Number of Leaves : 13	
Size of the tree : 20	

Figura 72 - Modelo P-4: Árvore de decisão no modo texto para classe MFP de eventos ocorridos até 1 ANO.

Classifier Model - P - 5	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0: N (4662.0/224.0)	1
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: S (307.0/6.0)	2
NroCKMBPos_cat = OUTRO	3
NroCKMBPos <= 2.98: N (937.89/59.33)	4
NroCKMBPos > 2.98	5
Infarto_Recente = S	6
TipolInfartoRecente = N: N (0.0)	7
TipolInfartoRecente = MENOS_24H	8
NroCKMBPos <= 35.48: N (98.18/17.82)	9
NroCKMBPos > 35.48: S (22.17/6.98)	10
TipolInfartoRecente = ENTRE_24_E_72H: N (76.69/15.67)	11
TipolInfartoRecente = MAIS_72H	12
NroTropoPos <= 31.71: N (17.35/5.11)	13
NroTropoPos > 31.71: S (34.67/11.12)	14
Infarto_Recente = N	15
STENT_OVERLAP = S: S (46.34/5.34)	16
STENT_OVERLAP = N	17
ext_NroVasos_Index <= 1	18
QUADRO_SCA = SCA_CSST: N (0.0)	19
QUADRO_SCA = SCA_SSST: S (28.34/11.34)	20
QUADRO_SCA = N	21
Isquemia_Silenciosa = S	22
NroCKMBPos <= 3.34: N (15.01/5.0)	23
NroCKMBPos > 3.34: S (15.01/6.01)	24
Isquemia_Silenciosa = N: N (16.01/5.33)	25
ext_NroVasos_Index > 1: S (100.35/25.35)	26
Number of Leaves : 16	
Size of the tree : 27	

Figura 73 - Modelo P-5: Árvore de decisão no modo texto para classe MFP de eventos ocorridos até 1 ANO.

Classifier Model - P-6	
J48 pruned tree	

NroCKMBPos_cat = IN_0_OUT_0	1
DoencaVascPeriferica = S	2
ext_NroLesions_Index <= 1: S (26.0/8.0)	3
ext_NroLesions_Index > 1: N (13.0/4.0)	4
DoencaVascPeriferica = N: N (829.0/202.0)	5
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: S (301.0)	6
NroCKMBPos_cat = OUTRO	7
NroCKMBPos <= 2.92	8
ext_SomenteNativos_Index = S: N (172.57/49.05)	9
ext_SomenteNativos_Index = N: S (13.21/5.81)	10
NroCKMBPos > 2.92: S (273.21/40.67)	11

Figura 74 - Modelo P-6: Árvore de decisão no modo texto.

(Modelo P- 8) Classe: MACE_REFERENCIA

Classifier Model - P - 8	
J48 pruned tree	

ext_trombo = S: S (39.0/11.0)	1
ext_trombo = N	2
DoencaVascPeriferica = S: S (33.0/11.0)	3
DoencaVascPeriferica = N	4
ext_residual_Media <= 6.13	5
Infarto_Recente = S	6
FAIXA_ETARIA = IDOSO: S (31.2/8.68)	7
FAIXA_ETARIA = ADULTO	8
DIABETES = N	9
NroCKMBPos_cat = IN_0_OUT_0	10
ext_calcio = NAO: S (11.84/4.84)	11
ext_calcio = MODERADO: N (10.0/3.0)	12
ext_calcio = ALTO: S (1.0)	13
NroCKMBPos_cat = IN_0_OUT_MAIOR_5: N (4.0)	14
NroCKMBPos_cat = OUTRO: S (19.84/5.0)	15
DIABETES = SIM_NIDep: N (12.0/2.0)	16
DIABETES = SIM_IDep: S (1.84)	17
FAIXA_ETARIA = JOVEM: S (2.84)	18
Infarto_Recente = N	19
LESOES_NOVAS = S	20
ext_calcio = NAO: N (212.24/82.2)	21
ext_calcio = MODERADO	22
FUNC_RENAL = NORMAL_DISCRETA: N (119.03/33.91)	23
FUNC_RENAL = MODERADA_GRAVE: S (35.85/14.93)	24
ext_calcio = ALTO: S (27.84/10.0)	25
LESOES_NOVAS = N	26
ext_calcio = NAO: S (12.0/2.0)	27
ext_calcio = MODERADO: N (11.0/4.0)	28
ext_calcio = ALTO: S (0.0)	29
ext_residual_Media > 6.13: S (97.47/30.4)	30
Number of Leaves : 19	
Size of the tree : 31	

Figura 75 - Modelo P-8: Árvore de decisão no modo texto.

6.4 ANEXO D - *Dataset* com as variáveis numéricas e categóricas.

#	FULL_NUM : dataset com variáveis numéricas (N=65)
1	@attribute IDADE NUMERIC
2	@attribute SEXO {M,F}
3	@attribute IMC NUMERIC
4	@attribute CLEARANCE NUMERIC
5	@attribute FUNC_RENAL {NORMAL_DISCRETA,MODERADA_GRAVE}
6	@attribute HISTORIA_CV {S,N}
7	@attribute IAM_ANT {NO,QWAVE,NOQWAVE}
8	@attribute REVASC_ANT {S,N}
9	@attribute CVA {S,N}
10	@attribute DoencaVascPeriferica {S,N}
11	@attribute HistoriaFamiliar_DCA {S,N}
12	@attribute HAS {S,N}
13	@attribute Hiperlipidemia {S,N}
14	@attribute Fumo {N,PAROU,S}
15	@attribute DIABETES {N,SIM_NIDep,SIM_IDep}
16	@attribute Quadro_Chegada {ISQUEMIA,ANG_EST,ANG_INST,RECENT_MI}
17	@attribute QUADRO_DC_ESTAVEL {S,N}
18	@attribute QUADRO_SCA {SCA_CSST,SCA_SSST,N}
19	@attribute Isquemia_Silenciosa {S,N}
20	@attribute Angina_Estavel {S,N}
21	@attribute Angina_Instavel {S,N}
22	@attribute Infarto_Recente {S,N}
23	@attribute Tipo_angina {N,ESTAVEL,INSTAVEL}
24	@attribute TipoAngEstavel {CCS1,CCS2,CCS3,CCS4}
25	@attribute TipoAngInstavel {IA,IIA,IIIA,IB,IIB,IIC,IIC,IIC}
26	@attribute TipoinfartoRecente {N,MENOS_24H,ENTRE_24_E_72H,MAIS_72H}
27	@attribute ST_elevation {S,N}
28	@attribute AgenteTrombolitico {S,N}
29	@attribute FVE {NORMAL_DISCRETO,MODERADO_GRAVE}
30	@attribute KILLIP {I,II,III,IV}
31	@attribute Clopidogrel {S,N}
32	@attribute Estatina {S,N}
33	@attribute CKMB_PRE {N,NORMAL,ANORMAL}
34	@attribute NroCKMBPre NUMERIC
35	@attribute NroCKMBPos NUMERIC
36	@attribute NroCKMBPos_cat {ENTRE_1E3,ENTRE_3E10,MAIOR_10}
37	@attribute NroTropoPre NUMERIC
38	@attribute NroTropoPos NUMERIC
39	@attribute ext_SomenteNativos_Index {S,N}
40	@attribute ext_Ponte_Idade_maior_10anos {S,N}
41	@attribute ext_Ponte NUMERIC
42	@attribute ext_NroLesions_Index NUMERIC
43	@attribute ext_NroVasos_Index NUMERIC
44	@attribute ext_NroTerritorios_Index NUMERIC
45	@attribute ext_NroTerritorios_Index_categ {HUM,DOIS,TRES,QUATRO}
46	@attribute TERRITORIO_TCE {S,N}
47	@attribute ext_territ_DA {S,N}
48	@attribute ext_territ_CX {S,N}
49	@attribute ext_territ_CD {S,N}
50	@attribute vaso_fino {S,N,MIX}
51	@attribute ext_vaso_diam_menor NUMERIC
52	@attribute ext_vaso_diam_medio NUMERIC
53	@attribute ext_vaso_diam_medio_fino {S,N}
54	@attribute ext_lesao_classif {A,B1,B2,C}
55	@attribute ext_trombo {S,N}
56	@attribute ext_calcio {NAO,MODERADO,ALTO}
57	@attribute ext_residual_maxima NUMERIC
58	@attribute ext_residual_Media NUMERIC
59	@attribute ext_lesao_pos_dilat {S,N}
60	@attribute ext_lesao_maior_30 {S,N}
61	@attribute ext_lesoes_compr_tot NUMERIC
62	@attribute LESOES_NOVAS {S,N}
63	@attribute lesao_longa {S,N,MIX}
64	@attribute ext_bifurcacao {S,N}
65	@attribute NROSTENTS_Index NUMERIC
66	@attribute Stent_Compr_Total_Index NUMERIC
67	@attribute STENT_OVERLAP {S,N}
68	@attribute STENT_GERACAO {NOVA,MIX,VELHA}
69	@attribute OneTypeStent_Index {CYPHER,TAXUS,BALAO,ENDEAVOR,XCIENCE,PROMUS,BIOMATRIX,RESOLUT}
70	@attribute MACE {S,N}

Figura 76 - Desire, dataset numérico composto por 70 atributos, denominado FULL_NUM.

#	FULL_CATEG : dataset com variáveis categóricas (N=65)
1	@attribute FAIXA_ETARIA {IDOSO,ADULTO,JOVEM}
2	@attribute SEXO {M,F}
3	@attribute IMC_CAT {ABAIXO,NORMAL_ACIMA,OBESO}
4	@attribute CLEARANCE_NUMERIC
5	@attribute FUNC_RENAL {NORMAL_DISCRETA,MODERADA_GRAVE}
6	@attribute HISTORIA_CV {S,N}
7	@attribute IAM_ANT {NO,QWAVE,NOQWAVE}
8	@attribute REVASC_ANT {S,N}
9	@attribute CVA {S,N}
10	@attribute DoencaVascPeriferica {S,N}
11	@attribute HistoriaFamiliar_DCA {S,N}
12	@attribute HAS {S,N}
13	@attribute Hiperlipidemia {S,N}
14	@attribute Fumo {N,PAROU,S}
15	@attribute DIABETES {N,SIM_NIDep,SIM_IDep}
16	@attribute Quadro_Chegada {ISQUEMIA,ANG_EST,ANG_INST,RECENT_MI}
17	@attribute QUADRO_DC_ESTAVEL {S,N}
18	@attribute QUADRO_SCA {SCA_CSST,SCA_SSST,N}
19	@attribute Isquemia_Silenciosa {S,N}
20	@attribute Angina_Estavel {S,N}
21	@attribute Angina_Instavel {S,N}
22	@attribute Infarto_Recente {S,N}
23	@attribute Tipo_angina {N,ESTAVEL,INSTAVEL}
24	@attribute TipoAngEstavel {CCS1,CCS2,CCS3,CCS4}
25	@attribute TipoAngInstavel {IA,IIA,IIIA,IB,IIB,IIC,IIIC}
26	@attribute TipoInfartoRecente {N,MENOS_24H,ENTRE_24_E_72H,MAIS_72H}
27	@attribute ST_elevation {S,N}
28	@attribute AgenteTrombolitico {S,N}
29	@attribute FVE {NORMAL_DISCRETO,MODERADO_GRAVE}
30	@attribute KILLIP {I,II,III,IV}
31	@attribute Clopidogrel {S,N}
32	@attribute Estatina {S,N}
33	@attribute CKMB_PRE {N,NORMAL,ANORMAL}
34	@attribute NroCKMBPre_NUMERIC
35	@attribute NroCKMBPos_cat {ENTRE_1E3,ENTRE_3E10,MAIOR_10}
36	@attribute TROPONINA_VAR {300<DELTA<1000,DELTA>1000}
37	@attribute SyntaxScore_cat {BAIXO,MODERADO,ALTO}
38	@attribute ext_SomenteNativos_Index {S,N}
39	@attribute ext_Ponte_Idade_maior_10anos {S,N}
40	@attribute ext_Ponte_NUMERIC
41	@attribute NroLesions_Index_categ {1,2,3,4,5,6,7}
42	@attribute ext_NroVasos_Index_categ {1,2,3,4,5,6,7}
43	@attribute ext_NroTerritorios_Index_categ {HUM,DOIS,TRES,QUATRO}
44	@attribute TERRITORIO_TCE {S,N}
45	@attribute ext_territ_DA {S,N}
46	@attribute ext_territ_CX {S,N}
47	@attribute ext_territ_CD {S,N}
48	@attribute vaso_fino {S,N,MIX}
49	@attribute ext_vaso_diam_medio_fino {S,N}
50	@attribute ext_lesao_classif {A,B1,B2,C}
51	@attribute ext_trombo {S,N}
52	@attribute ext_calcio {NAO,MODERADO,ALTO}
53	@attribute ext_residual_maxima_NUMERIC
54	@attribute ext_residual_Media_NUMERIC
55	@attribute ext_lesao_pos_dilat {S,N}
56	@attribute ext_lesao_maior_30 {S,N}
57	@attribute LESOES_NOVAS {S,N}
58	@attribute lesao_longa {S,N,MIX}
59	@attribute ext_bifurcacao {S,N}
60	@attribute NROSTENTS_Index_categ {1,2,3,4,5,6,7}
61	@attribute Stent_Compr_Total_Index_NUMERIC
62	@attribute STENT_OVERLAP {S,N}
63	@attribute STENT_GERACAO {NOVA,MIX,VELHA}
64	@attribute OneTypeStent_Index {CYPHER,TAXUS,BALAO,ENDEAVOR,XCIENCE,PROMUS,BIOMATRIX,RESOLUT}
65	@attribute MACE {S,N}

Figura 77 - Desire, dataset categórico composto por 65 atributos, denominado FULL_CATEG.

6.5 ANEXO E - Seleção dos 20 atributos mais significativos pelo critério da razão de ganho (GR) para as classes MACE e MFP.

Attribute Evaluator (supervised, Class (nominal): 70 MACE):		
Gain Ratio feature evaluator		
Ranked attributes:		
0.10257	36	NroCKMBPos_cat
0.0292	1	IDADE
0.02867	55	ext_trombo
0.02162	58	ext_residual_Media
0.02047	39	ext_SomenteNativos_Index
0.01831	68	STENT_GERACAO
0.0161	35	NroCKMBPos
0.01541	57	ext_residual_maxima
0.01262	38	NroTropoPos
0.00952	4	CLEARANCE
0.00885	19	Isquemia_Silenciosa
0.00882	26	TipoInfartoRecente
0.00836	69	OneTypeStent_Index
0.00771	22	Infarto_Recente
0.00757	6	HISTORIA_CV
0.00749	8	REVASC_ANT
0.00731	10	DoencaVascPeriferica
0.00707	18	QUADRO_SCA
0.00558	29	FVE
0.00534	17	QUADRO_DC_ESTAVEL

Figura 78 - Para a classe MACE os 20 atributos mais significativos do dataset numérico pelo critério GR.

Attribute Evaluator (supervised, Class (nominal): 70 MFP):		
Gain Ratio feature evaluator		
Ranked attributes:		
0.1754	36	NroCKMBPos_cat
0.04348	35	NroCKMBPos
0.03801	55	ext_trombo
0.02196	38	NroTropoPos
0.01837	26	TipoInfartoRecente
0.0148	22	Infarto_Recente
0.01383	18	QUADRO_SCA
0.01359	4	CLEARANCE
0.0093	66	Stent_Compr_Total_Index
0.00877	17	QUADRO_DC_ESTAVEL
0.00837	39	ext_SomenteNativos_Index
0.00774	61	ext_lesoes_compr_tot
0.00764	67	STENT_OVERLAP
0.00702	1	IDADE
0.00658	65	NROSTENTS_Index
0.00616	16	Quadro_Chegada
0.00564	10	DoencaVascPeriferica
0.00477	60	ext_lesao_maior_30
0.004	19	Isquemia_Silenciosa
0.00386	43	ext_NroVasos_Index

Figura 79 - Para a classe MFP os 20 atributos mais significativos do dataset numérico pelo critério GR.

6.6 ANEXO F - Avaliação das regras extraídas pelos especialistas de domínio.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE (Modelo P – 2)

Estas regras classificam os pacientes pela classe MACE e que relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos nos pacientes na fase hospitalar e pós-hospitalar ao longo dos 13 anos de seguimento.

Classe: MACE - Modelo P-2 (6.377 pacientes sendo 1.614 com eventos ao longo de 13 anos)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de segunda geração	N	94%	1765	1	100
2	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira e segunda geração	N	87%	110	2	75
3	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com histórico cardiovascular e com estenose residual média das lesões pós procedimento $\leq 8,6\%$	N	69%	1248	2	50
4	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com quadro diferente de ISQUEMIA SILENCIOSA, com histórico cardiovascular e a estenose residual média das lesões pós procedimento $> 8,6\%$	S	60%	65	2	50
5	Idem a regra 3, só que em pacientes sem histórico cardiovascular	N	78%	1283	2	75
6	Se o valor da CKMB = 0 no exame pré e se for maior do 5 vezes no exame pós procedimento	S	100%	307	1	100
7	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MENOR do que 2,98	N	85%	937	1	100
8	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 mas SEM infarto recente	S	73%	221	3	0
9	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade > 83 anos	S	83%	12	1	75
10	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 32,78 em pacientes COM infarto recente MENOS_24H e idade MENOR do que 83 anos	S	65%	20	2	50
11	Idem a regra anterior exceto pelo valor do CKMB pós MENOR do que 32,78	N	82%	88	2	50

ESPECIALISTA #1

Figura 80 - Avaliação das regras extraídas do modelo P-2 pelo especialista nº 1.

Avaliação das regras geradas pela árvore de decisão

Classe: MFP (Modelo P – 5)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Modelo P-5 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO	N	95%	4662	1	100
2	Se o valor da CKMB = 0 no exame pré e for maior do 5 vezes a no exame pós procedimento	S	98%	307	2	75
3	Se o valor da CKMB no exame pré for diferente de ZERO e no pós procedimento estiver entre ZERO e 2,98	N	94%	937	1	100
4	Se o paciente apresentou-se com infarto recente MENOS_24H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver no intervalo entre 2,98 E 35,48	N	83%	98	2	50
5	Idem a regra anterior, mas com valor da CKMB pós procedimento MAIOR do que 35,48	S	73%	22	2	50
6	Se o paciente apresentou-se com infarto recente ENTRE_24_E_72H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver MAIOR do que 35,48	N	80%	76	3	25
7	Se o paciente não teve infarto e o valor do CKMB no exame pré for DIFERENTE de ZERO e no exame pós procedimento estiver MAIOR do que 2,98 e tiver overlap de stent	S	89%	46	2	75
8	Se o paciente apresentou-se com quadro de SCA_SSST e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU APENAS 1 VASO sem overlap de stent	S	69%	97	2	50
9	Se o paciente apresentou-se sem infarto recente e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU MAIS DE 1 VASO sem overlap de stent	S	75%	100	2	75

ESPECIALISTA #1

Figura 81 - Avaliação das regras extraídas do modelo P-5 pelo especialista nº 1.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE_REFERÊNCIA (Modelo P – 8)

Estas regras classificam os pacientes pela classe MACE_REFERENCIA e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase pós-hospitalar ao longo do 1º ano, sendo que estão excluídos os IAM_NQ da fase hospitalar.

Classe: MACE_REFERENCIA - Modelo P-8 (6.377 pacientes sendo 341 com eventos)					AVALIAÇÃO	
Regra	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concordância %
1	Se houver presença de trombo nas lesões tratadas	S	72%	39	1	100
2	Se não houver trombo nas lesões mas o paciente tiver doença vascular periférica	S	67%	33	2	50
3	Paciente idoso COM iam recente SEM doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$	S	72%	31	0	0
4	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM valor da estenose residual média das lesões $\leq 6,13\%$. Se TODAS lesões novas e SEM a presença de cálcio	N	61%	212	2	50
5	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$ com TODAS lesões novas, COM presença MODERADA de cálcio e função renal MODERADA_GRAVE	S	60%	35	1	75
6	Idem à anterior mas com função renal NORMAL_DISCRETA	N	72%	119	1	100
7	Idem à regra 4 diferindo pela presença ALTA do cálcio nas lesões.	S	63%	27	2	50
8	Paciente sem doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $> 6,13\%$.	S	69%	97	3	50

ESPECIALISTA #1

Figura 82 - Avaliação das regras extraídas do modelo P-8 pelo especialista nº 1.

Avaliação das regras geradas pelo algoritmo CN2

Classe: MFP (Orange CN2)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Orange CN2 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Pacientes com mais de 35 anos, com o valor da CKMB = 0 no exame pré e maior do que 5 vezes no exame pós procedimento E troponina pós > 29	S		184	↓	100
2	Pacientes com mais de 69 anos, com o valor da CKMB = 0 no exame pré e maior do que 2 vezes no exame pós procedimento E função renal moderada a grave	S		39	↓	100
3	Pacientes com CLEARANCE entre 64 e 130 E troponina pós > 0 E presença moderada de cálcio na lesão	N		318	↓	100
4	Pacientes com IMC maior do que 34, com o valor da CKMB = 0 pré e pós procedimento CUJA soma do comprimento das lesões tratadas não ultrapassam 40 mm E com bifurcação	N		108	?	

ESPECIALISTA #1

Figura 83 - Avaliação das regras extraídas do modelo CN2 pelo especialista nº 1.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE (Modelo P – 2)

Estas regras classificam os pacientes pela classe MACE e que relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos nos pacientes na fase hospitalar e pós-hospitalar ao longo dos 13 anos de seguimento.

lucio → prof. ATC / Hand
total 100

Classe: MACE - Modelo P-2 (6.377 pacientes sendo 1.614 com eventos ao longo de 13 anos)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de segunda geração	N	94%	1765	1	75%
2	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira e segunda geração	N	87%	110	1	75%
3	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com histórico cardiovascular e com estenose residual média das lesões pós procedimento $\leq 8,6\%$	N	69%	1248	1	50%
4	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com quadro diferente de ISQUEMIA SILENCIOSA, com histórico cardiovascular e a estenose residual média das lesões pós procedimento $> 8,6\%$	S	60%	65	1	50%
5	Idem a regra 3, só que em pacientes sem histórico cardiovascular	N	78%	1283	1	75%
6	Se o valor da CKMB = 0 no exame pré e se for maior do 5 vezes no exame pós procedimento	S	100%	307	1	100%
7	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MENOR do que 2,98	N	85%	937	1	75%
8	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 mas SEM infarto recente	S	73%	221	2	50%
9	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade > 83 anos	S	83%	12	1	75%
10	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 32,78 em pacientes COM infarto recente MENOS_24H e idade MENOR do que 83 anos	S	65%	20	1	75%
11	Idem a regra anterior exceto pelo valor do CKMB pós MENOR do que 32,78	N	82%	66	1	50%

Especialista #2

Figura 84 - Avaliação das regras extraídas do modelo P-2 pelo especialista nº 2.

Avaliação das regras geradas pela árvore de decisão

Classe: MFP (Modelo P – 5)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

(MS) também se fala e kidej

Classe: MFP - Modelo P-5 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO	N	95%	4662	1	100%
2	Se o valor da CKMB = 0 no exame pré e for maior do 5 vezes a no exame pós procedimento	S	98%	307	1	75%
3	Se o valor do CKMB no exame pré for diferente de ZERO e no pós procedimento estiver entre ZERO e 2,98	N	94%	937	1	75%
4	Se o paciente apresentou-se com infarto recente MENOS_24H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver no intervalo entre 2,98 E 35,48	N	83%	98	2	50%
5	Idem a regra anterior, mas com valor da CKMB pós procedimento MAIOR do que 35,48	S	73%	22	1	50%
6	Se o paciente apresentou-se com infarto recente ENTRE_24_E_72H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver MAIOR do que 35,48	N	80%	76	2	50%
7	Se o paciente não teve infarto e o valor do CKMB no exame pré for DIFERENTE de ZERO e no exame pós procedimento estiver MAIOR do que 2,98 e tiver overlap de stent	S	89%	46	2	50%
8	Se o paciente apresentou-se com quadro de SCA_SSST e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU APENAS 1 VASO sem overlap de stent	S	69%	97	2	25%
9	Se o paciente apresentou-se sem infarto recente e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU MAIS DE 1 VASO sem overlap de stent	S	75%	100	2	25%

Especialista #2

Figura 85 - Avaliação das regras extraídas do modelo P-5 pelo especialista nº 2.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE_REFERÊNCIA (Modelo P – 8)

Estas regras classificam os pacientes pela classe MACE_REFERENCIA e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase pós-hospitalar ao longo do 1º ano, sendo que estão excluídos os IAM_NQ da fase hospitalar.

Classe: MACE_REFERENCIA - Modelo P-8 (6.377 pacientes sendo 341 com eventos)					AVALIAÇÃO	
Regra	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se houver presença de trombo nas lesões tratadas	S	72%	39	1	50%
2	Se não houver trombo nas lesões mas o paciente tiver doença vascular periférica	S	67%	33	1	75%
3	Paciente idoso COM iam recente SEM doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$	S	72%	31	1	50%
4	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM valor da estenose residual média das lesões $\leq 6,13\%$. Se TODAS lesões novas e SEM a presença de cálcio	N	61%	212	2	25%
5	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$ com TODAS lesões novas, COM presença MODERADA de cálcio e função renal MODERADA_GRAVE	S	60%	35	1	75%
6	Idem à anterior mas com função renal NORMAL_DISCRETA	N	72%	119	1	75%
7	Idem à regra 4 diferindo pela presença ALTA do cálcio nas lesões.	S	63%	27	1	50%
8	Paciente sem doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $> 6,13\%$.	S	69%	97	1	50%

Especialista #2

Figura 86 - Avaliação das regras extraídas do modelo P-8 pelo especialista nº 2.

Avaliação das regras geradas pelo algoritmo CN2

Classe: MFP (Orange CN2)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Orange CN2 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Pacientes com mais de 35 anos, com o valor da CKMB = 0 no exame pré e maior do que 5 vezes no exame pós procedimento E troponina pós > 29	S		184	1	
2	Pacientes com mais de 69 anos, com o valor da CKMB = 0 no exame pré e maior do que 2 vezes no exame pós procedimento E função renal moderada a grave	S		39	1	
3	Pacientes com CLEARANCE entre 64 e 130 E troponina pós > 0 E presença moderada de cálcio na lesão	N		318	1	
4	Pacientes com IMC maior do que 34, com o valor da CKMB = 0 pré e pós procedimento CUJA soma do comprimento das lesões tratadas não ultrapassam 40 mm E com bifurcação	N		108	1	

?

Especialista #2

Figura 87 - Avaliação das regras extraídas do modelo CN2 pelo especialista nº 2.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE (Modelo P – 2)

Estas regras classificam os pacientes pela classe MACE e que relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos nos pacientes na fase hospitalar e pós-hospitalar ao longo dos 13 anos de seguimento.

Classe: MACE - Modelo P-2 (6.377 pacientes sendo 1.614 com eventos ao longo de 13 anos)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de segunda geração	N	94%	1765	2	100%
2	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira e segunda geração	N	87%	110	2	75%
3	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com histórico cardiovascular e com estenose residual média das lesões pós procedimento $\leq 8,6\%$	N	69%	1248	1	100%
4	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com quadro diferente de ISQUEMIA SILENCIOSA, com histórico cardiovascular e a estenose residual média das lesões pós procedimento $> 8,6\%$	S	60%	65	2	50%
5	Idem a regra 3, só que em pacientes sem histórico cardiovascular	N	78%	1283	2	50%
6	Se o valor da CKMB = 0 no exame pré e se for maior do 5 vezes no exame pós procedimento	S	100%	307	1	100%
7	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MENOR do que 2,98	N	85%	937	1	100%
8	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 mas SEM infarto recente	S	73%	221	1	100%
9	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade > 83 anos	S	83%	12	1	100%
10	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 32,78 em pacientes COM infarto recente MENOS_24H e idade MENOR do que 83 anos	S	65%	20	1	100%
11	Idem a regra anterior exceto pelo valor do CKMB pós MENOR do que 32,78	N	82%	88	2	0%

ESPECIALISTA #3

Figura 88 - Avaliação das regras extraídas do modelo P-2 pelo especialista nº 3.

Avaliação das regras geradas pela árvore de decisão

Classe: MFP (Modelo P – 5)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Modelo P-5 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO	N	95%	4662	1	100
2	Se o valor da CKMB = 0 no exame pré e for maior do 5 vezes a no exame pós procedimento	S	98%	307	1	100
3	Se o valor do CKMB no exame pré for diferente de ZERO e no pós procedimento estiver entre ZERO e 2,98	N	94%	937	2	75%
4	Se o paciente apresentou-se com infarto recente MENOS_24H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver no intervalo entre 2,98 E 35,48	N	83%	98	2	50%
5	Idem a regra anterior, mas com valor da CKMB pós procedimento MAIOR do que 35,48	S	73%	22	1	100%
6	Se o paciente apresentou-se com infarto recente ENTRE_24_E_72H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver MAIOR do que 35,48	N	80%	76	3	0%
7	Se o paciente não teve infarto e o valor do CKMB no exame pré for DIFERENTE de ZERO e no exame pós procedimento estiver MAIOR do que 2,98 e tiver overlap de stent	S	89%	46	2	100%
8	Se o paciente apresentou-se com quadro de SCA_SSST e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU APENAS 1 VASO sem overlap de stent	S	69%	97	2	100%
9	Se o paciente apresentou-se sem infarto recente e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU MAIS DE 1 VASO sem overlap de stent	S	75%	100	1	100%

ESPECIALISTA #3

Figura 89 - Avaliação das regras extraídas do modelo P-5 pelo especialista nº 3.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE_REFERÊNCIA (Modelo P – 8)

Estas regras classificam os pacientes pela classe MACE_REFERENCIA e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase pós-hospitalar ao longo do 1º ano, sendo que estão excluídos os IAM_NQ da fase hospitalar.

Classe: MACE_REFERENCIA - Modelo P-8 (6.377 pacientes sendo 341 com eventos)					AVALIAÇÃO	
Regra	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se houver presença de trombo nas lesões tratadas	S	72%	39	1	100%
2	Se não houver trombo nas lesões mas o paciente tiver doença vascular periférica	S	67%	33	2	75%
3	Paciente idoso COM iam recente SEM doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$	S	72%	31	3	25%
4	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM valor da estenose residual média das lesões $\leq 6,13\%$. Se TODAS lesões novas e SEM a presença de cálcio	N	61%	212	1	100%
5	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$ com TODAS lesões novas, COM presença MODERADA de cálcio e função renal MODERADA_GRAVE	S	60%	35	2	75%
6	Idem à anterior mas com função renal NORMAL_DISCRETA	N	72%	119	3	0%
7	Idem à regra 4 diferindo pela presença ALTA do cálcio nas lesões.	S	63%	27	2	75%
8	Paciente sem doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $> 6,13\%$.	S	69%	97	1	100%

ESPECIALISTA #3

Figura 90 - Avaliação das regras extraídas do modelo P-8 pelo especialista nº 3.

Avaliação das regras geradas pelo algoritmo CN2

Classe: MFP (Orange CN2)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Orange CN2 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Pacientes com mais de 35 anos, com o valor da CKMB = 0 no exame pré e maior do que 5 vezes no exame pós procedimento E troponina pós > 29	S		184	1	100%
2	Pacientes com mais de 69 anos, com o valor da CKMB = 0 no exame pré e maior do que 2 vezes no exame pós procedimento E função renal moderada a grave	S		39	2	75%
3	Pacientes com CLEARANCE entre 64 e 130 E troponina pós > 0 E presença moderada de cálcio na lesão	N		318	1	100%
4	Pacientes com IMC maior do que 34, com o valor da CKMB = 0 pré e pós procedimento CUJA soma do comprimento das lesões tratadas não ultrapassam 40 mm E com bifurcação	N		108	3	50%

ESPECIALISTA #3

Figura 91 - Avaliação das regras extraídas do modelo CN2 pelo especialista nº 3.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE (Modelo P – 2)

Estas regras classificam os pacientes pela classe MACE e que relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos nos pacientes na fase hospitalar e pós-hospitalar ao longo dos 13 anos de seguimento.

Classe: MACE - Modelo P-2 (6.377 pacientes sendo 1.614 com eventos ao longo de 13 anos)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de segunda geração	N	94%	1765	1	100
2	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira e segunda geração	N	87%	110	1	100
3	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com histórico cardiovascular e com estenose residual média das lesões pós procedimento $\leq 8,6\%$	N	69%	1248	1	100
4	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com quadro diferente de ISQUEMIA SILENCIOSA, com histórico cardiovascular e a estenose residual média das lesões pós procedimento $> 8,6\%$	S	60%	65	0	0
5	Idem a regra 3, só que em pacientes sem histórico cardiovascular	N	78%	1283	1	100
6	Se o valor da CKMB = 0 no exame pré e se for maior do 5 vezes no exame pós procedimento	S	100%	307	1	100
7	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MENOR do que 2,98	N	85%	937	1	100
8	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 mas SEM infarto recente	S	73%	221	1	75
9	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade > 83 anos	S	83%	12	1	100
10	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 32,78 em pacientes COM infarto recente MENOS_24H e idade MENOR do que 83 anos	S	65%	20	1	100
11	Idem a regra anterior exceto pelo valor do CKMB pós MENOR do que 32,78	N	82%	88	0	0

ESPECIALISTA #4

Figura 92 - Avaliação das regras extraídas do modelo P-2 pelo especialista nº 4.

Avaliação das regras geradas pela árvore de decisão

Classe: MFP (Modelo P – 5)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Modelo P-5 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO	N	95%	4662	1	100
2	Se o valor da CKMB = 0 no exame pré e for maior do 5 vezes a no exame pós procedimento	S	98%	307	1	100
3	Se o valor do CKMB no exame pré for diferente de ZERO e no pós procedimento estiver entre ZERO e 2,98	N	94%	937	1	100
4	Se o paciente apresentou-se com infarto recente MENOS_24H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver no intervalo entre 2,98 E 35,48	N	83%	98	0	0
5	Idem a regra anterior, mas com valor da CKMB pós procedimento MAIOR do que 35,48	S	73%	22	1	100
6	Se o paciente apresentou-se com infarto recente ENTRE_24_E_72H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver MAIOR do que 35,48	N	80%	76	0	0
7	Se o paciente não teve infarto e o valor do CKMB no exame pré for DIFERENTE de ZERO e no exame pós procedimento estiver MAIOR do que 2,98 e tiver overlap de stent	S	89%	46	1	100
8	Se o paciente apresentou-se com quadro de SCA_SSST e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU APENAS 1 VASO sem overlap de stent	S	69%	97	1	100
9	Se o paciente apresentou-se sem infarto recente e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU MAIS DE 1 VASO sem overlap de stent	S	75%	100	1	100

① INDEPENDENTE DE OVERLAP DO STENT
 ② INDEPENDENTE DE OVERLAP DO STENT

ESPECIALISTA #4

Figura 93 - Avaliação das regras extraídas do modelo P-5 pelo especialista nº 4.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE_REFERÊNCIA (Modelo P – 8)

Estas regras classificam os pacientes pela classe MACE_REFERENCIA e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase pós-hospitalar ao longo do 1º ano, sendo que estão excluídos os IAM_NQ da fase hospitalar.

Classe: MACE_REFERENCIA - Modelo P-8 (6.377 pacientes sendo 341 com eventos)

Regra	Regra	Ocorrência de Evento	Probabilidade	Qtde	AVALIAÇÃO	
					Grau de novidade	Concorda %
1	Se houver presença de trombo nas lesões tratadas	S	72%	39	1	53%
2	Se não houver trombo nas lesões mas o paciente tiver doença vascular periférica	S	67%	33	1	75%
3	Paciente idoso COM iam recente SEM doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$	S	72%	31	0	0
4	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM valor da estenose residual média das lesões $\leq 6,13\%$. Se TODAS lesões novas e SEM a presença de cálcio	N	61%	212	1	75%
5	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$ com TODAS lesões novas, COM presença MODERADA de cálcio e função renal MODERADA_GRAVE	S	60%	35	1	100
6	Idem à anterior mas com função renal NORMAL_DISCRETA	N	72%	119	1	100
7	Idem à regra 4 diferindo pela presença ALTA do cálcio nas lesões.	S	63%	27	1	100
8	Paciente sem doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $> 6,13\%$.	S	69%	97	0	0

① DEPENDE DA CARGA DE TROMBOS

② DEPENDE DO GRAU DE LESÃO RESIDUAL, 6,1% É UM VALOR MÍNIMO

ESPECIALISTA #4

Figura 94 - Avaliação das regras extraídas do modelo P-8 pelo especialista nº 4.

Avaliação das regras geradas pelo algoritmo CN2

Classe: MFP (Orange CN2)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Orange CN2 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Pacientes com mais de 35 anos, com o valor da CKMB = 0 no exame pré e maior do que 5 vezes no exame pós procedimento E troponina pós > 29	S		184	1	100
2	Pacientes com mais de 69 anos, com o valor da CKMB = 0 no exame pré e maior do que 2 vezes no exame pós procedimento E função renal moderada a grave	S		39	1	100
3	Pacientes com CLEARANCE entre 64 e 130 E troponina pós > 0. E presença moderada de cálcio na lesão	N		318	1	100
4	Pacientes com IMC maior do que 34, com o valor da CKMB = 0 pré e pós procedimento CUJA soma do comprimento das lesões tratadas não ultrapassam 40 mm E com bifurcação	N		108	1	100

ESPECIALISTA #4

Figura 95 - Avaliação das regras extraídas do modelo CN2 pelo especialista nº 4.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE (Modelo P – 2)

Estas regras classificam os pacientes pela classe MACE e que relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos nos pacientes na fase hospitalar e pós-hospitalar ao longo dos 13 anos de seguimento.

Classe: MACE - Modelo P-2 (6.377 pacientes sendo 1.614 com eventos ao longo de 13 anos)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de segunda geração	N	94%	1765	2	75
2	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira e segunda geração	N	87%	110	2	75
3	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com histórico cardiovascular e com estenose residual média das lesões pós procedimento $\leq 8,6\%$	N	69%	1248	3 1	50
4	Se o valor da CKMB no exame pré e pós procedimento for ZERO e foram implantados somente stents de primeira geração em vasos nativos de pacientes com quadro diferente de ISQUEMIA SILENCIOSA, com histórico cardiovascular e a estenose residual média das lesões pós procedimento $> 8,6\%$	S	60%	65	3	25
5	Idem a regra 3, só que em pacientes sem histórico cardiovascular	N	78%	1253	2	75
6	Se o valor da CKMB = 0 no exame pré e se for maior do 5 vezes no exame pós procedimento	S	100%	307	1	100
7	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MENOR do que 2,98	N	85%	937	1	100
8	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 mas SEM infarto recente	S	73%	221	1	50
9	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade > 83 anos	S	83%	12	2	50
10	Se o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento MAIOR do que 2,98 em pacientes COM infarto recente MENOS_24H e idade MENOR do que 83 anos	S	65%	20	3	0
11	Idem a regra anterior exceto pelo valor do CKMB pós MENOR do que 32,78	N	82%	88	2	75

ESPECIALISTA #5

Figura 96 - Avaliação das regras extraídas do modelo P-2 pelo especialista nº 5.

Avaliação das regras geradas pela árvore de decisão

Classe: MFP (Modelo P – 5)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

PT não 100%

Classe: MFP - Modelo P-5 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se o valor da CKMB no exame pré e pós procedimento for ZERO	N	95%	4662	1	100
2	Se o valor da CKMB = 0 no exame pré e for maior do 5 vezes a no exame pós procedimento	S	98%	307	1	0
3	Se o valor do CKMB no exame pré for diferente de ZERO e no pós procedimento estiver entre ZERO e 2,98	N	94%	937	1	100
4	Se o paciente apresentou-se com infarto recente MENOS_24H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver no intervalo entre 2,98 E 35,48	N	83%	98	1	78
5	Idem a regra anterior, mas com valor da CKMB pós procedimento MAIOR do que 35,48	S	73%	22	2	75 50
6	Se o paciente apresentou-se com infarto recente ENTRE_24_E_72H e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento estiver MAIOR do que 35,48	N	80%	76	2	50
7	Se o paciente não teve infarto e o valor do CKMB no exame pré for DIFERENTE de ZERO e no exame pós procedimento estiver MAIOR do que 2,98 e tiver overlap de stent	S	89%	46	3	50
8	Se o paciente apresentou-se com quadro de SCA_SSST e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU APENAS 1 VASO sem overlap de stent	S	69%	97	2	50
9	Se o paciente apresentou-se sem infarto recente e o valor do CKMB no exame pré for DIFERENTE de ZERO e no pós procedimento for MAIOR do que 2,98 e TRATOU MAIS DE 1 VASO sem overlap de stent	S	75%	100	2	50

ESPECIALISTA #5

Figura 97 - Avaliação das regras extraídas do modelo P-5 pelo especialista nº 5.

Avaliação das regras geradas pela árvore de decisão

Classe: MACE_REFERÊNCIA (Modelo P – 8)

Estas regras classificam os pacientes pela classe MACE_REFERENCIA e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase pós-hospitalar ao longo do 1º ano, sendo que estão excluídos os IAM_NQ da fase hospitalar.

Classe: MACE_REFERENCIA - Modelo P-8 (6.377 pacientes sendo 341 com eventos)					AVALIAÇÃO	
Regra	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Se houver presença de trombo nas lesões tratadas	S	72%	39	1	100
2	Se não houver trombo nas lesões mas o paciente tiver doença vascular periférica	S	67%	33	2	50
3	Paciente idoso COM iam recente SEM doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$	S	72%	31	3	25
4	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM valor da estenose residual média das lesões $\leq 6,13\%$. Se TODAS lesões novas e SEM a presença de cálcio	N	61%	212	3	25
5	Paciente SEM doença vascular periférica SEM iam recente SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $\leq 6,13\%$ com TODAS lesões novas, COM presença MODERADA de cálcio e função renal MODERADA_GRAVE	S	60%	35	2	50
6	Idem à anterior mas com função renal NORMAL_DISCRETA	N	72%	119	2	75
7	Idem à regra 4 diferindo pela presença ALTA do cálcio nas lesões.	S	63%	27	2	50
8	Paciente sem doença vascular periférica SEM trombo nas lesões tratadas COM o valor da estenose residual média das lesões $> 6,13\%$.	S	69%	97	2	25

ESPECIALISTA #5

Figura 98 - Avaliação das regras extraídas do modelo P-8 pelo especialista nº 5.

Avaliação das regras geradas pelo algoritmo CN2

Classe: MFP (Orange CN2)

Estas regras classificam os pacientes pela classe MFP e relatam a probabilidade de ocorrência ou não de eventos cardíacos adversos na fase hospitalar e pós-hospitalar ao longo do 1º ano.

Classe: MFP - Orange CN2 (6.377 pacientes sendo 814 com eventos até o 1º ano)					AVALIAÇÃO	
Linha	Regra	Ocorrência de Evento	Probabilidade	Qtde	Grau de novidade	Concorda %
1	Pacientes com mais de 35 anos, com o valor da CKMB = 0 no exame pré e maior do que 5 vezes no exame pós procedimento E troponina pós > 29	S		184	1	100
2	Pacientes com mais de 69 anos, com o valor da CKMB = 0 no exame pré e maior do que 2 vezes no exame pós procedimento E função renal moderada a grave	S		39	1	75
3	Pacientes com CLEARANCE entre 64 e 130 E troponina pós > 0 E presença moderada de cálcio na lesão	N		318	2	50
4	Pacientes com IMC maior do que 34, com o valor da CKMB = 0 pré e pós procedimento CUJA soma do comprimento das lesões tratadas não ultrapassam 40 mm E com bifurcação	N		108	1	75

ESPECIALISTA #5

Figura 99 - Avaliação das regras extraídas do modelo CN2 pelo especialista nº 5.

7 REFERÊNCIAS

7 REFERÊNCIAS

1. Portal Brasil. [Online]. Doenças crônicas não transmissíveis são a maior causa de morte no mundo, diz OMS; 2011 [cited 21 09 2014]. Available from: <http://www.brasil.gov.br/saude/2011/09/doencas-cronicas-nao-transmissiveis-sao-a-maior-caoa-de-morte-no-mundo-diz-oms>.
2. Portal Brasil. [Online]. Doenças cardiovasculares causam quase 30% das mortes no País; 2011 [cited 2014 09 10]. Available from: <http://www.brasil.gov.br/saude/2011/09/doencas-cardiovasculares-causam-quase-30-das-mortes-no-pais>.
3. Bastos AS, Beccaria LM, Contrim LM, Cesarino CB. Tempo de chegada do paciente com Infarto agudo do miocárdio em unidade de emergência. *Rev Soc Bras Cir Cardiov*. 2012;27(3):411-8.
4. Shlomo S. Andreas Gruentzig - the life and death of a pioneer. *Cardiol J*. 2006;13(4):348-50.
5. Editor Andreas Gruentzig, M.D. (1939–85). SCAI (Society for Cardiovascular Angiography and Interventions). [Online].; 2014 [cited 2014 10 16]. Available from: <http://www.scai.org/About/History/Legends/Detail.aspx?cId=907997e5-519b-4723-9fd4-3dddb97b2072>.
6. Piva e Mattos LA. Portal da Sociedade Brasileira de Hemodinâmica e Cardiologia Intervencionista. [Online]. Cardiologia intervencionista brasileira é sinônimo mundial de excelência; 2014 [cited 2014 09 10]. Available from: <http://sbhci.org.br/publico-leigo/cardiologia-intervencionista-brasileira-e-sinonimo-mundial-de-excelencia/>.
7. Gonçalves BKD, Tedeschi AL, Sena MA, Peixoto RTS, Tedeschi BF. Evolução da intervenção coronariana percutânea: visão de um centro especializado. *Rev Bras Cardiol*. 2007 jan-fev;20(1):47-52
8. Sousa JEMR, Sousa AGMR. *Uma década (2002-2012) de emprego clínico dos stents farmacológicos no tratamento da Doença Arterial Coronária*. São Paulo: Atheneu; 2012.
9. Abu-Mostafa YS, Magdom-Ismail M, Lin HT. *Learning form data*. AMLBook.com; 2012.

10. Fayyad U, Piatetsky-Shapiro G, Padhraic S. From data mining to knowledge discovery in knowledge databases. *Art Int Magazine*. 1996; 17(3):37-54.
11. Rezende SO. *Sistemas inteligentes: fundamentos e aplicações*. Barueri: Manole; 2003.
12. Than PN, Steinbach M, Kumar V. *Introduction to data mining*. Addison Wesley; 2006.
13. Wasan SK, Bhatnagar V, Kaur H. The impact of data mining techniques on medical diagnosis. *Data Sci J*. 2006;5(19):119-26.
14. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. *Int J Med Inform*. 2008 Feb;77(2):81-97.
15. Ferro M, Lee HD. O processo de KDD knowledge discovery in database para aplicações na medicina. In: *Semana de Informática de Cascavel*; 2001; Cascavel. p. 57-62.
16. Steiner MTA, Soma NY, Nievola JC, Steiner Neto PJ. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. *Gest Prod*. 2006;13(2):325-37.
17. Cavalcante PF. *A importância de fatores de risco na obstrução das artérias coronárias utilizando técnicas de mineração de dados* [Tese]. Pontifícia Universidade Católica de Goiás, 2009.
18. Maciel TV, Seus VR, Machado KS, Borges EN. Mineração de dados em triagem de risco de saúde. *Rev Bras Comp Aplicada*. 2015 maio;7(2):26-40.
19. Vianna RCXF, Moro CMCB, Moysés SJ, Carvalho D, Niévola JC. Mineração de dados e características da mortalidade infantil. *Cad Saúde Pública*. 2010 mar;26(3):535-42.
20. Trinadade CM, Souza DD, Moro CMC, Aldenuci MG, Nievola JC, Moys SJ. Aplicação de KDD na descoberta de comportamento das Hepatites Virais no Município de Curitiba. In: *IX Congresso Brasileiro de Informática em Saúde*; 2004; Ribeirão Preto.
21. Carvalho DR, Dallagasa MR, Silva SH. Uso de técnicas de mineração de dados para a identificação automática de beneficiários propensos ao diabetes mellitus tipo 2. *Informação & Informação*. 2015 set/dez; 20(3):274-96.
22. Gayathri P, Jaisankar N. Comprehensive study of heart disease diagnosis using data mining and soft computing techniques. *Int J Engineer Technol*. 2013 jun-jul;5(3):29471-58.

23. Karaolis M, Moutiris JA, Papaconstantinou L, Pattichis CS. Association rule analysis for the assessment of the risk of coronary heart events. In: *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2009; Minneapolis, Minnesota, USA.
24. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Trans Inf Technol Biomed*. 2010 May;14(3):559-66.
25. Abdullah AS, Rajalaxmi RR. A data mining model for predicting the coronary heart disease using random forest classifier. in international conference on recent trends in computational methods, communication and controls. *Proceed Int J Comp Appl*; 2012;3(C):22-25.
26. Rajkuma A, Reena GS. Diagnosis of heart disease using datamining algorithm. *Global J Comp Sci Technol*. 2010 Sept;10(10):38-43.
27. Anbrasi M, Anupriya E, Iyengar NCSN. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *Int J Engineer Sci Technol*. 2010;2(10):5370-5376.
28. Kumar DS, G.Sathyadevi G, Sivanesh S. Decision support system for medical diagnosis using data mining. *Int J Comp Sci*. 2011 May;8(3):147-153.
29. Alizadehsani R, Habibi J, Bahadorian B, Mashayekhi H, Ghandeharioun A, Boghrati R, et al. Diagnosis of coronary arteries stenosis using data mining. *J Med Signals Sens*. 2012 Jul;2(3):153-9.
30. Lichman M. *UCI- machine learning repository*. [Online].; 2013 [cited 2014 09 10. Available from: <http://archive.ics.uci.edu/ml>.
31. Tavares S, Sousa AGMR, Costa RA, Moreira A, Costa Jr R, Maldonado G, et al. Impacto de stents farmacológicos em pacientes com doença arterial coronária estável submetidos a intervenção coronária percutânea na prática diária do mundo real. *Rev Bras Cardiol Invasiva*. 2010;18(4):392-99.
32. Costa RA, Sousa AGMR, Costa Jr R, Moreira A, Maldonado G, Cano MN, et al. Evolução tardia de pacientes com infarto agudo do miocárdio tratados com stents farmacológicos na prática diária do mundo real - subanálise do Registro DESIRE (Drug-Eluting Stent In the REal World). *Rev Bras Cardiol Invasiva*. 2011 jul/set;19(3):244-54.
33. Von Zuben F, Attux RRF. *Árvores de decisão* [Apostila]. Campinas: Unicamp, DCA-Departamento de Engenharia da Computação e Automação Industrial. Report N°. Tópico 7.

34. Prati RC, Batista GEAPA, Monard MC. Curvas ROC para avaliação de classificadores. *Latin America Transaction. IEEE America Latina*. 2008 Jun;6(2):215-22.
35. Witten IH, Frank. E, Hall MA. *Data mining: practical machine learning tools and techniques*. 3rd ed.: Morgan Kaufmann; 2011.
36. Quinlan JR. Induction of decision trees. *Machine Learning*. 1986;1:81-106.
37. Quinlan JR. *C4.5: programs for machine learning*. San Francisco-CA: Morgan Kaufmann; 1993.
38. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees (wadsworth statistics/probability)*. 1st ed. Florida: CRC Press; 1984.
39. Amo S. Técnicas de mineração de dados. In: *XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informatica*; 2004; Salvador. p. 43.
40. Wahbeh AH, Al-Radaideh QA, Al-Kabi MN, Al-Shawakfa EM. A comparison study between data mining tools over some classification methods. *Int J Adv Comp Sci Appl*. 2011;1:18-26.
41. Moghimipour I, Ebrahimpour M. Comparing decision tree method over three data mining software. *Int J Statist Probabil*. 2014 Jul;3(3): 147-156.
42. Vieira AJ, Garrett JM. Understanding interobserver agreement: the kappa statistics. *Fam Med*. 2005 May;37(5):360-3..
43. Demsar J, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, et al. Orange: data mining toolbox in python. *J Mach Learn Res*. 2013;14(1)2349-2353.
44. Thygesen K, Alpert JS, Simoons ML, Chaitman BR, White HD. Terceira definição universal de enfarte do miocárdio. *Eur Heart J*. 2013;1:2551-67.
45. Boente ANP, Goldschmidt RR, Estrela VV. Uma metodologia para o apoio à realização do processo de descoberta de conhecimento em base de dados. In: *Simpósio de Excelência em Gestão e Tecnologia*, 2008.
46. Prati RC, Batista GEAPA, Monard MC. A study with class imbalance and random sampling for a decision tree learning system. In: Baumer M. *Artificial intelligence and practice II*. Boston: Springer; 2008. p. 131-140.

47. Freitas AA. Comprehensible classification models - a position paper. *SIGKDD Explorations*. 2014;15(1):1-10.
48. Carvalho VO. *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado* [Tese]. Inst. Ciências Matemáticas e de Computação – USP/São Carlos, 2007.
49. Carvalho D, Escobar L, Tsunoda D. Pontos de atenção para o uso da mineração de dados na saúde. *Informação & Informação*. 2014 jan; 19(1):249-72.
50. Medeiros LR, Stein A. Medicina baseada em evidências e análise de decisão na clínica cirúrgica. *Rev Ass Med Rio Grande do Sul*. 2001 jan-jun; 45(1-2):45-50.