Giovanna Aguiar de Castro
Marcos Vinícius Ribeiro Silva

Trabalho Final de Mineração de Dados

Goiânia-GO
Julho de 2018

0.1 Introdução

Neste estudo foi utilizado a base de dados *Predicting Cardiac Pathology*. Ela foi originada a partir de dados colhidos na Unidade de Cardiologia Materno Fetal (UCMF) no estado do Recife.

O problema de classificação aqui em questão é determinar se pacientes possuem ou não cardiopatias. Baseando em dados clínicos destes pacientes. O objetivo geral do trabalho nestes dados consiste em usar técnicas e processos de descoberta de conhecimento em base de dados (no inglês *Knowledge Discovery Databases KDD*) para que seja possível chegar a uma resposta satisfatória, para a questão levantada anteriormente. Com o objetivo principal de se chegar à uma hipótese que possa ser generalizada para novos casos do mesmo contexto apresentado.

Este trabalho está organizado em mais 5 seções. Seção 0.2 que detalha as ferramentas e os métodos utilizados para a manipulação dos dados. Seção 0.3 que mostra uma estatística descritiva das variáveis presentes na base de dados em questão. Seção 0.4 que mostra uma descrição do que foi feito com a base de dados e apresenta uma nova estatística descritiva referente à esses novos valores. Seção 0.5 que detalha todas as análises feitas neste trabalho. Por fim a Seção 0.6 que faz a conclusão deste trabalho.

0.2 Metódos

Para a manipulação da base de dados foi utilizado a Linguagem de Programação Python a partir da **IDE**[1] **JetBrains PyCharm Community Edition** versao de 02/01/2018 para sistemas operacionas 64Bits em conjunto com o framework **Weka** (Waikato Enviroment for Knowledge Analysis) e o **R**, a partir da **IDE RStudio**.

Os experimentos foram executados num computador com processador $Intel\ Core^{TM}$ $i7-8700K$ CPU @ $3.70GHz$ x 16, com 16GB de memória e sistema operacional Windows 10 Pro.

O processamento de dados foi realizado em quatro etapas, o pré-processamento, a normalização, o processamento e a discretização dos dados.

Para o processamento dos dados, foi criado um software na linguagem python com o uso das bibliotecas csv, os, unicodedata e dateutil com o intuito de auxiliar, flexibilizar e aumentar a produtividade e na assertividade do tratamento de dados, além de generalizar o tratamento de dados para outras bases de dados.

Para a análise explorátoria do dados, foi criado um software na linguagem R com o uso das bibliotecas Hmisc, arulesViz, arules, RSenal, ggplot2, plotly e ggpubr.

[1] Ambiente de Desenvolvimento Integrado

Os softwares possuem código aberto, e foram criados pelos autores deste documento e está disponibilizado na plataforma **GitHub**, e está disponível neste link https://github.com/marcosvsilva/DataMiningCardiacPathology

0.3 Dados Originais

0.3.1 Descrição das variáveis

A base de dados contém 21 atributos com 17873 instâncias. A Tabela 1 mostra uma breve descrição do significado e da tipagem dos atributos presentes na base de dados. Na Tabela 1 é possível encontrar o nome de cada atributo, o seu significado e o seu tipo.

Tabela 1: Significado e tipagem dos atributos encontrados na base de dados original

Atributo	Descrição do atributo	Tipo Atributo
ID	identificador numérico do paciente	numérico
Peso	peso do paciente	numérico
Altura	altura do paciente	numérico
IMC	índice de massa corpórea do paciente	numérico
Atendimento	data da visita do paciente ao medico	nominal
DN	data de nascimento do paciente	nominal
IDADE	idade do paciente	categórica
Convenio	convenio no qual o paciente esta vinculado	categórica
PULSOS	pulsação do paciente o	categórica
PA SISTOLICA	pressão sistólica do sangue do paciente	numérico
PA DIASTOLICA	pressão diastólica do sangue do paciente	numérico
PPA	resultado SBP/DBP	categórica
NORMALXANORMAL	ausência ou presencia de patologia	categórica
B2	tipo do segundo som do coração	categórica
SOPRO	tipo de murmúrio	categórica
FC	frequência cardíaca	categórica
HDA 1	histórico de doença 1	categórica
HD2	histórico de doença 2	categórica
SEXO	gênero do paciente	categórica
MOTIVO 1	primeira razão do encaminhamento para a clínica de cardiologia	categórica
MOTIVO 2	segunda razão do encaminhamento para a clinica de cardiologia	categórica

0.3.2 Análise Preliminar

Esta subseção de Descrição das variáveis apresenta uma estatística descritiva e uma análise preliminar dos atributos da base de dados em questão.

A Tabela 2 mostra algumas medidas resumo das variáveis Peso, Altura, IMC, PA SISTÓLICA e PA DIASTÓLICA que originalmente na base de dados considerada são

do tipo quantitativa contínua. As medidas resumo de estatística referentes as variáveis apresentadas na Tabela 2 são mínimo valor, máximo valor, valor médio, mediana, desvio padrão e variância. Estas medidas em conjunto com histogramas e boxplots são capazes de mostrar como está o comportamento dos dados referentes à cada atributo. Este tipo de análise é feito nas subseções seguintes que são identificadas pelo nome de cada variável analisada.

Em relação as variáveis do tipo categórico é apresentado distribuições de frequências nas subseções referentes a estas mesmas variáveis. É importante ressaltar que o atributo ID não será descrito nesta Seção. Uma vez que ele é apenas um identificador.

Tabela 2: Medidas Resumo das Variáveis quantitativas da base de dados

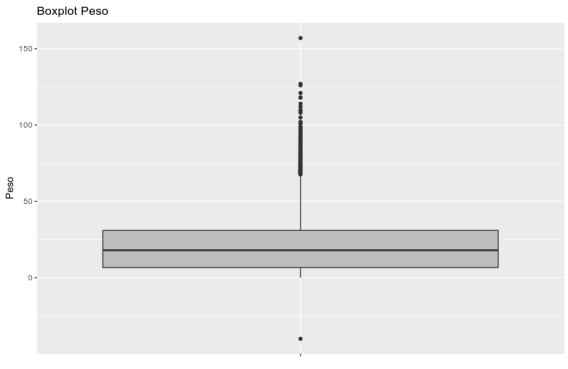
Nome	Mínimo	Máximo	Média	Mediana	Desvio Padrão	Variância
Peso	−40.00	157.00	21.16	18.00	18.07	326.52
Altura	0.00	198.00	83.86	99.0	56.58	3201.29
IMC	0.00	848.00	17.80	17.0	12.07	145.68
PA SISTÓLICA	10.00	990.00	101.34	100.0	15.51	240.56
PA DIASTÓLICA	6.00	120.00	62.30	60.00	8.88	78.85

0.3.2.1 Variável Peso

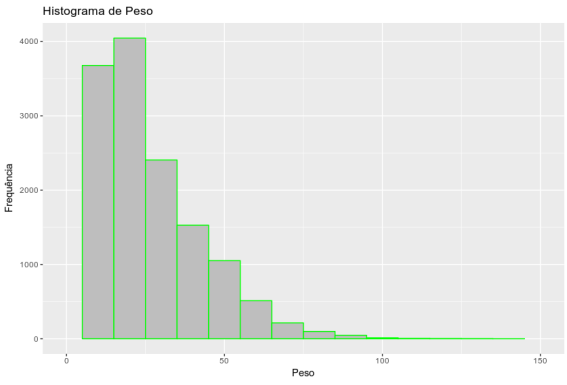
A variável Peso originalmente é do tipo quantitativa contínua e a sua unidade de medida é quilogramas (kg). Esta variável possui 319 valores faltantes (aproximadamente 2% da base de dados) e também apresenta 715 valores distintos.

Figura 1: Representação gráfica da distribuição dos valores de Peso

(a) Boxplot de Peso



(b) Histograma de Peso



A Figura 1 apresenta um boxplot (Figura 1a) e um histograma (Figura 1b) referente a variável Peso. A partir destas informações em conjunto com a Tabela 2 é possível notar as seguintes características:

- Os valores de peso se concentram em sua maioria entre os valores 0 kg e 50 kg. Pelo valor da mediana é possível notar que 50% das ocorrências de peso estão abaixo

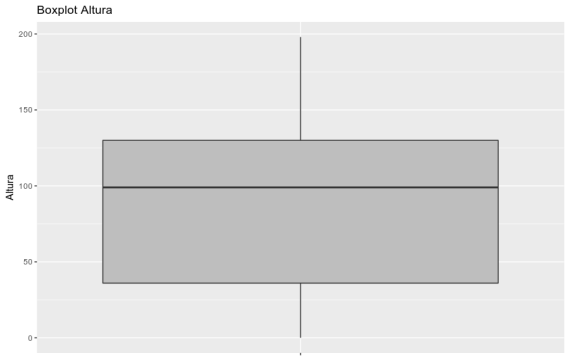
de 18 kg. Por este mesmo valor de mediana 18 kg e do boxplot (Figura 1a) é possível notar que alguns valores de peso se concentram próximo deste valor. A partir destas ilustrações também é possível observar alguns valores diferentes do esperado (*outliers*) tanto inferiormente quanto superiormente no boxplot. Pelo histograma (Figura 1b) e pelas medidas de média e mediana é possível notar que a variável peso possui uma distribuição assimétrica à direita (assimétrica positiva). Isto é, os valores estão concentrados a direita no histograma.

0.3.2.2 Variável Altura

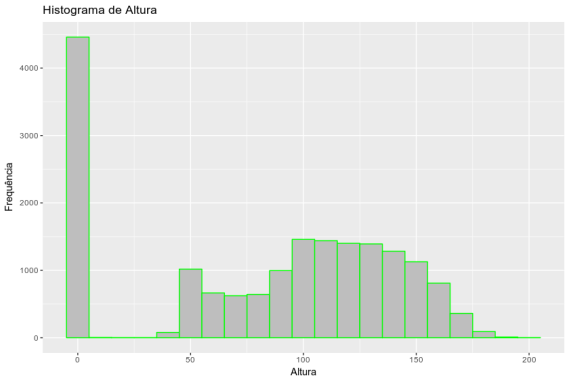
A variável Altura é originalmente é do tipo quantitativa contínua e sua unidade de medida é centímetros (cm). Esta variável não possui nenhum valor faltante e também apresenta 166 valores distintos.

Figura 2: Representação gráfica da distribuição dos valores de Altura

(a) Boxplot de Altura



(b) Histograma de Altura



A Figura 2 apresenta um boxplot (Figura 2a) e um histograma (Figura 6) referente a variável Altura. Com base na Figura 1a e na Tabela 1a é possível notar que:

- O boxplot (Figura 2a) mostra que a variável altura não possui valores fora da faixa esperada (*outliers*). A variável Altura possui uma distribuição assimétrica à esquerda (assimétrica negativa). Isto é os valores se concentram mais à esquerda do histograma (Figura 6). No entanto há uma considerável quantidade de valores negativos próximos do valor 0. 50% dos valores de altura estão abaixo de 100 cm.

0.3.2.3 Variável IMC

O valor de IMC é calculada a partir da altura e peso de cada paciente. A variável IMC é originalmente é do tipo quantitativa contínua. Esta variável possui 4727 valores faltantes (aproximadamente 26% da base de dados) e também apresenta 1695 valores distintos.

A Figura 3 apresenta um boxplot (Figura 8) e um histograma (Figura 3b) referentes ao atributo IMC. Pela Figura 3 e a Tabela 2 é possível notar que:

- A variável IMC apresenta uma grande quantidade de valores fora da faixa esperada (*outliers*) o que mostra que há inconsistências nesses valores. Uma vez que não é possível que um paciente tenha IMC com um valor de 848.00. Esta variável possui uma distribuição assimétrica à direita (assimétrica positiva) ou seja, os valores apresentados estão concentrados a direita do valor médio. Isto mostra que pode haver inconsistências nos valores das variáveis altura e peso, uma vez que o IMC é obtido a partir destes valores.

0.3.2.4 Variável Atendimento

A variável Atendimento é originalmente do tipo categórica e possui 983 valores faltantes (aproximadamente 5% da base de dados). Esta variável possui 2111 categorias distintas e por esta razão não será mostrado a sua distribuição de frequência. Esta variável será desconsiderada para a geração de classificadores pois ela apenas representa a data em que o paciente foi atendido.

0.3.2.5 Variável DN

A variável DN representa a data de nascimento de cada paciente é do tipo nominal e possui 1378 (aproximadamente 8% da base de dados) valores faltantes. Esta variável possui 6443 categorias distintas. Pelo fato de DN possuir muitas categorias distintas não será mostrado a sua distribuição de frequência.

0.3.2.6 Variável IDADE

A variável IDADE originalmente é do tipo categórica. Esta variável possui 1376 (aproximadamente 8% da base de dados) valores faltantes e também possui 1954 categorias distintas. Considerando o fato de que IDADE é do tipo categórica é notado um erro na sua tipagem pois IDADE deveria ser numérica uma vez que ela representa a idade dos pacientes.

0.3.2.7 Variável Convenio

A variável Convenio originalmente é do tipo categórica. Esta variável possui 5304 valores faltantes (aproximadamente 30% da base de dados) e também apresenta 439 categorias. Pelo fato de Convenio possui diversas categorias não será mostrado sua distribuição de frequência. Este atributo será desconsiderado para a criação de classificadores uma vez que o nome do convenio de um paciente não influencia no fato do paciente possuir ou não uma cardiopatia.

0.3.2.8 Variável PULSOS

A variável PULSOS é originalmente do tipo categórica e possui 1198 (aproximadamente 7% da base de dados) valores distintos A Tabela 3 mostra a distribuição de frequência da variável PULSOS.

Tabela 3: Distribuição de Frequência da variável PULSOS

PULSOS	Frequência Absoluta	Frequência Relativa [%]
Normais	16509	92.36
NORMAIS	2	0.01
Outro	45	0.25
Amplos	57	0.31
AMPLOS	1	0.005
Femorais diminuidos	43	0.24
Diminuídos	18	0.10

Com base na Tabela 3 é notório que há categorias com nomes equivalentes em PULSOS mas que seus significados são os mesmos. Estes casos sinônimos são Normais e NORMAIS, Amplos e AMPLOS. Ainda nesta mesma tabela é perceptível que em 92.37% (considerando a junção de Normais com NORMAIS) das ocorrências de PULSOS os pacientes apresentavam uma pulsação Normal.

0.3.2.9 Variável PA SISTOLICA

A variável PA SISTOLICA é originalmente do tipo quantitativa contínua e possui 7730 valores faltantes (aproximadamente 43% da base de dados) e também apresenta 20 valores distintos.

A Figura 4 mostra um boxplot (Figura 4a) e um histograma (Figura 4b) referentes a variável PA SISTOLICA. Baseado na Figura 4 em conjunto com a Tabela 2 é possível notar que:

- Assim como a variável IMC, PA SISTOLICA possui uma quantidade considerável de valores inesperados (*outliers*) este fator aponta para inconsistências na hora da imputação ou medição destes valores. É possível notar em PA SISTOLICA valores inconsitentes com a realidade. Para exemplo pode se usar paciente que apresenta uma PA SISTOLICA 10 e outro com uma PA SISTOLICA 990 uma vez que os limites inferiores e superiores de PA SISTOLICA em humanos variam aproximadamente entre 30 e 150. A variável PA SISTOLICA possui uma distribuição assimétrica à direita (assimétrica positiva) o que significa que seus valores estão concentrados a direita no histograma (Figura 4b). Além de que 50% dos seus valores estão abaixo de 100.

0.3.2.10 Variável PA DIASTÓLICA

A variável PA DIASTOLICA é originalmente do tipo quantitativa contínua e possui 7740 valores faltantes (aproximadamente 43% da base de dados) e apresenta também 37 valores distintos.

A Figura 5 mostra um boxplot (Figura 5a) e um histograma (Figura 5b) referentes a variável PA DIASTOLICA. Por meio destas figuras e da Tabela 2 é possível observar que:

- A partir do boxplot (Figura 5a) da variável PA DIASTOLICA é notável que esta variável possui valores fora do esperado (*outliers*) tanto inferiormente quanto superiormente. A partir disto é notável que houve inconsistências no calculo ou na imputação de alguns valores de PA DIASTOLICA na base de dados. Um exemplo a ser citado de um *outlier* é um paciente ter um PA DIASTOLICA 120, uma vez que um limite superior de PA DIASTOLICA em humanos gira em torno de 97. Esta variável possui uma distribuição assimétrica à direita (assimétrica positiva) o que quer dizer que seus valores estão concentrados a direita do valor médio no histograma (Figura 5b). Além de que 50% dos seus valores estão abaixo de 60.

0.3.2.11 Variável PPA

A variável PPA é uma variável agregadora que pode ser obtida por meio de PA SISTOLICA e PA DIASTOLICA. Este atributo é originalmente do tipo categórico e possui 217 valores faltantes (aproximadamente 1% da base de dados) e apresenta 9 categorias distintas. A Tabela 4 apresenta uma distribuição de frequência da variável PPA.

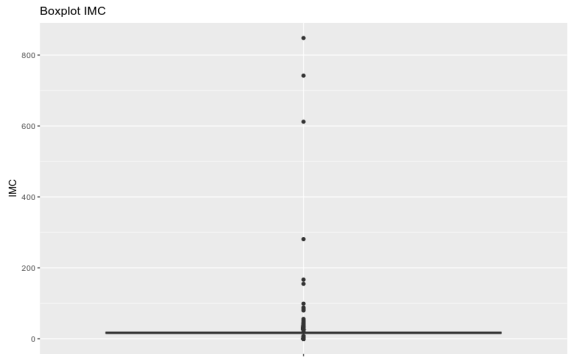
Tabela 4: Distribuição de Frequência da variável PPA

PPA	Frequência Absoluta	Frequência Relativa [%]
Não Calculado	9081	50.80
Normal	6141	34.35
Pre-Hipertensão PAS	193	1.07
HAS-2 PAS	215	1.20
Pre-Hipertensão PAD	233	1.30
#VALUE!	1496	8.37
HAS-1 PAS	153	0.85
HAS-2 PAD	58	0.32
HAS-1 PAD	86	0.48

A partir da Tabela 4 é notável que a categoria que mais se repete em PPA é a categoria Não calculado com 50.80% das ocorrências de PPA. Isto pode indicar inconsistências nos valores dos atributos PA SISTOLICA e PA DIASTOLICA. Uma vez que a PPA é obtida por meio da razão destas duas variáveis. Além de que em 8.37% das vezes

Figura 3: Representação gráfica da distribuição dos valores de IMC

(a) Boxplot de IMC



(b) Histograma de IMC

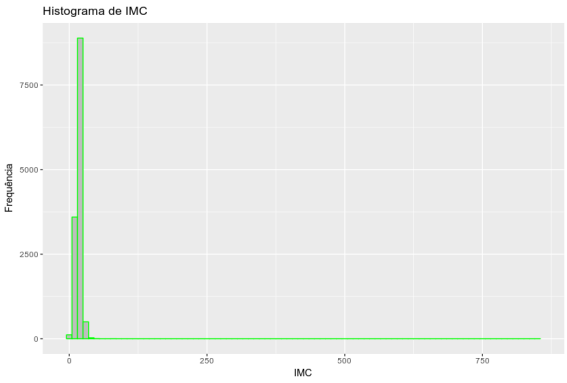
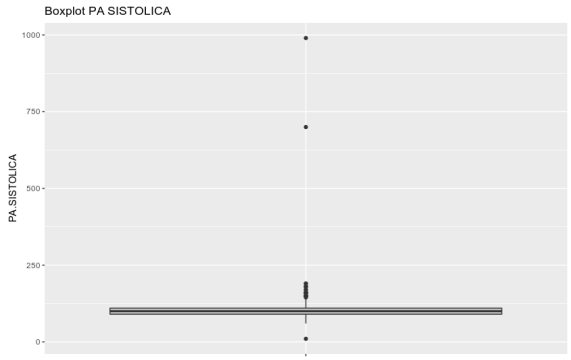


Figura 4: Representação gráfica da distribuição dos valores de PA SISTOLICA

(a) Boxplot de PA SISTOLICA



(b) Histograma de PA SISTOLICA

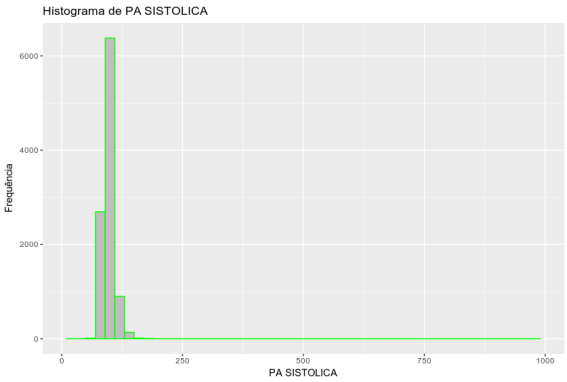
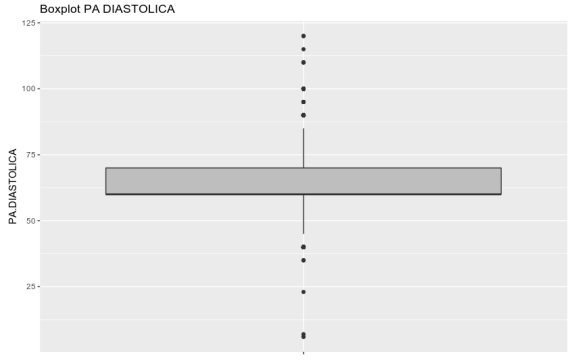
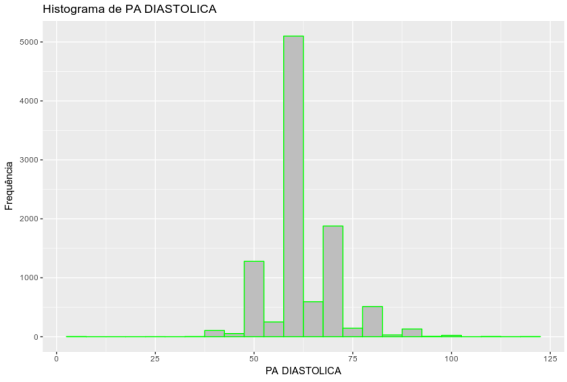


Figura 5: Representação gráfica da distribuição dos valores de PA DIASTOLICA

(a) Boxplot de PA DIASTOLICA



(b) Histograma de PA DIASTOLICA



em PPA houve a imputação incorreta de valores o que acarretou na geração de #VALUE! nos campos em questão.

0.3.2.12 Variável NORMAL X ANORMAL

A variável NORMAL X ANORMAL é o atributo classe. Isto é o problema em questão se resume em classificar um paciente com a presença ou a ausência de patologia cardiológica (categorias de NORMAL X ANORMAL). Esta variável é originalmente do tipo categórica com 1168 valores faltantes (aproximadamente 7% da base de dados) e possui também 4 distintas categorias. As categorias de NORMAL X ANORMAL com suas respectivas frequências são mostradas na Tabela 5.

Tabela 5: Distribuição de Frequência da variável NORMAL X ANORMAL

NORMAL X ANORMAL	Frequência Absoluta	Frequência Relativa [%]
Anormal	6712	37.55
Normal	9961	55.73
anormal	1	0.005
Normais	1	0.005

Por meio da Tabela 4 é perceptível que há algumas categorias com nomes sinônimos à outras como por exemplo a categoria Normal e Normais que possuem o mesmo sentido semântico. Ainda nesta tabela é visto que em 55.73% das ocorrências de NORMAL X ANORMAL é da categoria Normal. Isto é em 55.73% dos casos os pacientes foram declarados sem cardiopatias. Apesar de a diferença desta classe com a classe Anormal ser consideravelmente pouca isto pode trazer problemas na questão de classificação pelo fato de que as classes apresentadas estão desbalanceadas.

0.3.2.13 Variável B2

A variável B2 é originalmente do tipo categórica e apresenta 1179 valores faltantes (aproximadamente 7% da base de dados) com 5 classes distintas. A Tabela 6 mostra uma distribuição de frequência da variável B2.

Tabela 6: Distribuição de Frequência da variável B2

B2	Frequência Absoluta	Frequência Relativa [%]
Normal	15969	89.34
Desdob fixo	190	1.06
Outro	107	0.59
Hiperfonética	342	1.91
Única	86	0.48

Com base na Tabela 6 é perceptível que em 89.34% o tipo do segundo som do coração é Normal. As demais classes presentes em B2 tiveram uma baixa frequência em relação à classe Normal.

0.3.2.14 Variável SOPRO

A variável SOPRO é originalmente do tipo categórica e apresenta 1167 valores faltantes (aproximadamente 7% da base de dados) com 7 classes distintas. A Tabela 7 mostra uma distribuição de frequência da variável SOPRO.

Tabela 7: Distribuição de Frequência da variável SOPRO

SOPRO	Frequência Absoluta	Frequência Relativa [%]
Sistólico	4821	26.97
ausente	10727	60.01
sistólico	1090	6, 09
contínuo	30	0.16
Contínuo	23	0.13
diastólico	12	0.06
sistólico e diastólico	3	0.01

Por meio da Tabela 7 é notável que há classes com nomes sinônimos em relação à outras classes. Este fato aparece nas classes Sistólico e sistólico e em contínuo e Contínuo. Ainda nesta tabela é visível que em 60.01% há a ausência de murmúrio no coração dos pacientes.

0.3.2.15 Variável FC

A variável FC originalmente é do tipo categórica e apresenta 2041 valores faltantes (aproximadamente 11% da base de dados) com 127 categorias distintas. Por apresentar uma grande quantidade de categorias não será mostra a sua distribuição de frequência. No entanto há um erro na tipagem de FC uma vez que ela deve ser numérica pois representa a frequência cardíaca dos pacientes.

0.3.2.16 Variável HDA 1

A variável HDA 1 é originalmente do tipo categórica e possui 5414 (aproximadamente 30% da base de dados) valores faltantes com 8 categorias distintas. A Tabela 8 mostra uma distribuição de frequência da variável HDA 1.

A partir da Tabela 8 é visto que em aproximadamente 51.98% dos casos os pacientes se apresentam como assintomáticos em relação ao histórico da doença 1. Além de que as demais categorias presentes em HDA 1 possuem um pequeno percentual de frequência em relação a categoria Assintomático.

0.3.2.17 Variável HDA 2

A variável HDA 2 é originalmente do tipo categórica e possui 17221 (aproximadamente 96% da base de dados) valores faltantes com 8 categorias distintas. A Tabela 9 apresenta uma distribuição de frequência desta variável.

A grande quantidade de valores faltantes em HDA 2 pode impactar na relevância deste atributo em relação aos demais atributos. Isto é a pequena quantidade de valores presentes em HDA 2 podem não ser capazes de caracterizar determinados fatos dos pacientes. Dentre estes valores encontrados a categoria que se teve mais presente neste atributo foi Palpitação indicando que 0.83% dos pacientes tiveram a presença de Palpitação em relação ao histórico da doença 2.

0.3.2.18 Variável SEXO

A variável sexo é originalmente do tipo categórica e possui 4 (aproximadamente 0.022% da base de dados) valores faltantes com 6 categorias distintas. A Tabela 10 apresenta uma distribuição de frequência desta variável.

A partir da Tabela 10 é visto que há categorias com nomes sinônimos que possuem em seus valores o mesmo significado. Estes casos são apontados em M, Masculino e masculino que denotam pacientes do sexo masculino e em F e Feminino que denotam pacientes do sexo feminino. Considerando a junção desses sinônimos é notável que houve uma maior frequência de pacientes do sexo masculino. Ainda nessa mesma tabela é mostrado a categoria Indeterminado que posteriormente será tratada. Uma vez que um sexo indeterminado não interferirá na classificação de um paciente em relação se ele será diagnosticado com a presença ou não de uma cardiopatia.

0.3.2.19 Variável MOTIVO 1

A variável MOTIVO 1 é originalmente do tipo categórica e possui 1097 (aproximadamente 6% da base de dados) valores faltantes com 5 categorias distintas. A Tabela 11 mostra uma distribuição de frequência de MOTIVO 1.

Com base nas informações mostradas na Tabela 11 é analisado que a maior frequência relativa encontrada foi a da categoria 5-Parecer cardiológico com um valor de 44.65%. Isto quer dizer que em 7981 ocorrências o primeiro motivo de um paciente procurar um especialista é pelo fato do paciente estar em busca de um parecer cardiológico. Por esta mesma tabela também é perceptível a baixa frequência de pacientes que procuram especialistas para realizarem um *check-up*.

0.3.2.20 Variável MOTIVO 2

A variável MOTIVO 2 é originalmente do tipo categórica e possui 4778 (aproximadamente 27% da base de dados) valores faltantes com 16 categorias distintas. A Tabela 12 mostra uma distribuição de frequência de MOTIVO 2.

Por meio da Tabela 12 é visto que há duas categorias com nomes equivalentes que possuem o mesmo significado em seus valores. As categorias em questão são Outro e

Tabela 8: Distribuição de Frequência da variável HDA 1

HDA1	Frequência Absoluta	Frequência Relativa [%]
Palpitacao	576	3.22
Dispneia	764	4.27
Assintomático	9291	51.98
Dor precordial	873	4.88
Desmaio/tontura	270	1.51
Outro	207	1.15
Cianose	298	1.66
Ganho de peso	180	1.00

Tabela 9: Distribuição de Frequência da variável HDA 2

HDA2	Frequência Absoluta	Frequência Relativa [%]
Dispneia	138	0.77
Palpitacao	150	0.83
Desmaio/tontura	77	0.43
Ganho de peso	35	0.19
Outro	67	0.37
Dor precordial	98	0.54
Assintomático	1	0.005
Cianose	86	0.48

Tabela 10: Distribuição de Frequência da variável SEXO

SEXO	Frequência Absoluta	Frequência Relativa [%]
M	8930	49.96
F	6612	36.99
Indeterminado	1417	7.92
Masculino	584	3.26
Feminino	247	1.38
masculino	79	0.44

Tabela 11: Distribuição de Frequência da variável MOTIVO 1

MOTIVO 1	Frequência Absoluta	Frequência Relativa [%]
6- Suspeita de cardiopatia	5863	32.80
2- Check-up	1048	5.86
5- Parecer cardiológico	7981	44.65
1- Cardiopatia já estabelecida	1428	7.98
7- Outro	456	2.55

Outros. Ainda nesta tabela é perceptível que em 23.56% das vezes o motivo dos pacientes serem encaminhados para a clínica de cardiologia é uma cirurgia.

0.3.3 Main changes in the database

Nesta subseção será mostrado apenas as principais mudanças que foram feitas na base de dados de um modo geral. Isto é alterações que refletiram no comportamento dos dados como um todo. As demais mudanças feitas à nível de atributo são apontadas em suas respectivas subseções na Seção 0.4.

A primeira etapa foi o pré-processamento de dados, que consiste em remover todas as instâncias que não possuíam classificação para o atributo de interesse, logo, como não tinham classificação, não poderiam fornecer nenhuma informação quanto a sua influência no contexto como um todo.

Como na primeira etapa, também foi removido todas as instâncias que possuíam idade inferior a zero superior a dezoito anos, pois neste trabalho, há o intuito de predizer problemas cardíacos entre crianças e adolescentes com os atributos disponibilizados pela base de dados. Logo registros que não contém idade, ou a idade está fora do intervalo mencionado, pode evidenciar características que nem sempre são verdade para o público a qual o trabalho é realizado.

Após a primeira etapa de retirar os registros da base de dados que não convêm, foi gerada uma nova base de dados mais enxuta com um total de 15103 registros.

Os atributos ID, convênio, atendimento e DN foram removidos da base de dados. Onde o campo ID é utilizado apenas para identificar o paciente e esta relação de identificação não faz sentido na questão do poder de classificação de pacientes com cardiopatia ou não. Visto que a ideia aqui é ter um maior poder de generalização na questão da rotulação de pacientes. O atributo convênio carrega apenas a informação de qual convênio está vinculado determinado paciente deste modo não há porque considerá-lo na base de dados a ser utilizada para a geração de classificadores. Os atributos data de atendimento e nascimento não são necessários, pois a única informação importante expressa por ambas para o contexto do problema a ser solucionado, é a idade do paciente no momento da coleta de dados, no entanto, este dado é expresso na base de dados no atributo idade.

O campo HDA 2 foi retirado da base de dados, uma vez que ele possuía dados relevantes para apenas 4% da base de dados e deste modo seu poder de influência para contribuir em uma determinação é drasticamente reduzido pelo fato de ele ter poucas ocorrências.

Após a limpeza de instâncias e atributos, foi realizada uma limpeza de todos valores inválidos, estes que não são reconhecidos pela codificação binária UTF-8, que é a representação que foi utilizada em todo o trabalho, estes valores são expressos por

‘$\#VALUE!'$

Ainda na etapa de pré-processamento geral da base de dados, foram convertidos todos os valores de ponto flutuante que falharam na conversão do formato *.xls* para *.csv* que representaram a divisão da casa decimal com a casa inteira por virgula, para uma leitura limpa e correta de todos valores da base de dados. Também foram convertidos todos os valores do tipo texto em textos maiúsculos e também foi removido todas as acentuações para evitar que possam ser tratados de forma diferente, textos iguais expressos com letras acentuadas ou maiúsculas e minúsculas.

A etapa de normalização, consiste em delimitar um limite máximo e mínimo para o atributo a ser normalizado, está definição de intervalo leva em consideração o valor do atributo, os valores possíveis reais para este atributo e a importância e relação dele para o contexto do problema. Valores que de atributos que não estejam dentro do intervalo delimitado ficam sem valor, entrando para a estatística de análise como valor faltante, e será descrita detalhadamente variável a variável na sessão 0.4.

Agora, já na etapa de processamento dos dados, o primeiro processo realizado foi o recalculo da idade, realizado com os atributos **Atendimento e DN**[2], calculando-se a idade do paciente com a diferença do dia que o paciente foi atendido, para o dia que o paciente nasceu. Para os casos onde não se possuí valores válidos para os atributos DN e atendimento nenhum valor é atribuído a idade do paciente.

O segundo processo realizado pela etapa de processamento, é a conversão da altura para centímetros, como a unidade não é expressa na base de dados, é determinado um limiar, que é definido como quatro, que todos valores abaixo deste limiar é assumido como unidade metro e é convertido para centímetros e valores acima deste limiar é assumido como centímetros e não é convertido, o valor quatro foi assumido pois não existem crianças ou adolescentes com altura acima de quatro metros e nem abaixo de quatro centímetros.

A terceira etapa do processamento, é o recalculo do IMC, com os atributos idade e peso, devidamente limpos pelo pré-processamento e normalizados pela etapa de normalização. Quando os atributos idade e peso, possuem valores validos, o recalculo o IMC é realizado com sua fórmula matemática padrão, com o resultado atribuído ao IMC, e para os casos onde os atributos idade ou peso não possuem valores validos, não é atribuído valor ao atributo.

No processamento de dados, foi realizado a reclassificação do atributo PPA, utilizando os atributos **pressão sistólica, pressão arterial diastólica, altura, sexo e idade**[3] do paciente. A reclassificação só é realizada quando todos os atributos possuem valores válidos. É realizado verificando se a pressão arterial sistólica e diastólica são maiores que os valores normal para o percentil 95 da tabela de níveis de pressão arterial (disponí-

[2] A remoção dos atributos no pré-processamento só é finalizada após a primeira etapa do processamento.
[3] Atributos: PA SISTOLICA, PA DIASTOLICA, ALTURA, SEXO e IDADE respectivamente.

vel em: https://ead.inf.ufg.br/pluginfile.php/95790/mod_resource/content/1/Tables%20of%20blood%20pressure%20for%20children.pdf) após identificar o a idade e o sexo do paciente, e utilizando sua altura como percentil de classificação de altura, com os valores abaixo de 5 utilizando o percentil 5 e valores de altura acima de 95 utilizando o percentil 95 da tabela. Para crianças de zero a 11 meses de idade é utilizado os valores de um ano expresso pela tabela. Para entender mais sobre os cálculos e a classificação realizada de pressão arterial, pode se atentar ao documento de explicação de classificação (disponível em: https://ead.inf.ufg.br/pluginfile.php/95789/mod_resource/content/1/Explanation%20about%20the%20PPA%20variable.pdf)

Para as instâncias que não foram viáveis a reclassificação da PPA por possuir algum atributo dependente com valor nulo, foi utilizado o atributo já normalizado existente, caso este exista. Com isto, ao final do processo de reclassificação da PPA, notou-se que mesmo com todos os trabalhos e esforços para tentar obter um atributo relevante, este ainda não possuía valor em 50% dos registros do banco de dados, ficando ainda inviável seu uso pelo mesmo motivo do atributo HDA1 e sendo este o último atributo retirado da base de dados.

Ainda na sessão de processamento realizamos a junção dos atributos MOTIVO1 e MOTIVO2. Conforme análises anteriores foi descoberto que os dois atributos trabalhavam como classe e subclasse não sendo necessários dois atributos para expressar tais informações. A junção gerou um atributo MOTIVO como resultado, e tomou por prioridade o atributo MOTIVO2, onde o atributo MOTIVO é preenchido com o valor do atributo MOTIVO2 sempre que válido, e com o valor do atributo MOTIVO1 quando for inválido. Quanto o valor do atributo MOTIVO1 também for inválido o valor do atributo resultante da junção também fica sem valor. Logo tal processamento retira dois atributos da base de dados e adiciona um atributo, finalizando a base de dados com 13 atributos e 1 atributo classe.

Sendo a última etapa do processamento de dados, simplesmente a realocação do atributo de classificação para o último atributo da lista, pois o software de análise, utiliza como padrão o ultimo atributo da lista como classe de interesse e classificador. Esse processo não altera em nada a base de dados, somente auxilia a análise arquitetando a base de dados igualmente a arquitetura esperada pelo software de análise mencionado na sessão 0.2.

Agora já na etapa de discretização dos dados, todos atributos foram discretizados em intervalos iguais com 5 categorias distintas com o detalhe da discretização de cada atributo relatado na 0.4.

Todas as etapas aqui citadas de pré-processamento, processamento, normalização e discretização dos dados, é realizado por um software mencionado na sessão 0.2

0.4 Descrição Básica

Esta seção apresenta uma estatística descritiva dos atributos da base de dados que já se encontram pré processados. Além de que também é descrito as principais mudanças produzidas nos atributos da base de dados original. Esta seção tem o intuito de analisar se houve uma diferença em alguns resultados estatísticos em relação a análise preliminar feita na Subseção 0.3.2.

Após as mudanças descritas na Subseção 0.3.3 a base de dados em questão passou a ter 14 atributos com 15103 instâncias.

0.4.1 Variável PESO

A etapa de normalização desta variável retirou todos os valores abaixo de 0 e acima de 175, pois foi subentendido que os pesos estão na unidade quilograma e pela variância dos dados e a comparação dos mesmos com a idade do paciente do registro. Então foi aberto uma margem bem grande para este atributo até mesmo pelo menos não ter muito ruído e pela importância desta variável na nossa análise, onde um problema cardíaco pode estar altamente ligado ao excesso ou falta de peso de um paciente.

O atributo peso foi discretizado em 5 intervalos distintos. A partir disto seu tipo passou a ser categórico. Este atributo na nova base de dados possui 1742 (aproximadamente 12% da base de dados) valores faltantes. Sua unidade de medida ainda foi deixada em quilogramas kg. A Tabela 13 apresenta uma distribuição de frequência de PESO.

Por meio da Tabela 13 é importante observar que 10082 pacientes possuem seus pesos no intervalo $[0, 35)$, isto é entre 0 e menor que 35kg. Este fato é predominante pela característica de que a maioria dos pacientes encontrados na base de dados são crianças. Tanto que as duas classes que possuem os maiores pesos ($[105, 140)$ e $[140, 175)$) dispõem de uma pequena frequência absoluta em relação as demais classes encontradas.

0.4.2 Variável ALTURA

A variável altura foi normalizada com valores de 0 a 200, assumindo a unidade centímetro como padrão e já convertido na etapa de processamento. A análise feita na base de dados pode apresentar que valores acima de dois metros de altura, mesmo que possíveis no mundo real, podem interferir muito na predição pela sua alta influencia, no IMC e na PPA, por exemplo. Pela expectativa de generalização esperada do modelo, logo, realizar o corte é o mais eficiente.

Do mesmo modo que a variável Peso o atributo Altura foi discretizado em 5 categorias distintas. Deste modo o seu tipo passou a ser categórica. Este atributo apresenta 3086 (aproximadamente 20% da base de dados) valores faltantes. É importante ressaltar

que a unidade de medida de Altura ainda continua centímetros cm. A Tabela 14 mostra uma distribuição de frequência de Altura.

Com base na Tabela 14 é visto que cerca de 60% dos pacientes possuem altura entre 80 e 160 cm. O que denota uma possível normalidade em relação a frequência dos valores de pesos encontrados na Tabela 13.

0.4.3 Variável IMC

A variável IMC foi normalizada de 0 a 60, pois valores acima de 60 já entram em um limiar de problemas acima de cardíacos, se possível ser registrados esses valores, como o IMC já depende de dois atributos que foram normalizados, o intervalo foi amplamente aberto, limitando em um valor alto, para que também sejam evitados valores fora da faixa esperada.

O atributo IMC foi recalculado com base nos atributos altura e peso. Quando era apresentado inconsistências em altura ou peso o valor de IMC se tornava um valor faltante, uma vez que não é possível obter um valor de IMC quando há inconsistências nos atributos que são usados para calculá-lo.

Após todas as transformações o atributo IMC foi discretizado em 5 intervalos distintos e passou a possuir 3422 (aproximadamente 23% da base de dados) valores faltantes. Deste modo o tipo de IMC passou a ser categórico. A Tabela 15 apresenta uma distribuição de frequência de IMC.

Por meio da Tabela 15 é notável a diferença dos valores presentes em IMC em relação aos valores originais. Isto revela o grande número de inconsistências apresentadas nos dados originais. Após este recalculo os valores de IMC passaram a ser mais próximos do esperado em relação aos pacientes considerados. A maior concentração de instâncias está entre os valores 12 e 24 o que confirma a afirmação anterior.

0.4.4 Variável IDADE

Por apresentar algumas inconsistências nos valores originais foi decidido realizar o recalculo da idade, uma vez que os demais atributos da base de dados permitiam realizar este novo cálculo. Este recalculo foi feito utilizando os atributos atendimento e DN.

Por questões de restrições de domínio foi decidido analisar apenas pacientes que possuíam idade no intervalo fechado de 0 a 18 anos. Deste modo foram eliminadas todas as instâncias de pacientes que tinham suas idades fora do intervalo descrito. A partir disto esta variável passou a não apresentar nenhum valor faltante. Além de que estes valores foram discretizados em 5 categorias distintas. A Tabela 16 apresenta uma distribuição de frequência de Idade.

Tabela 12: Distribuição de Frequência da variável MOTIVO 2

MOTIVO 2	Frequência Absoluta	Frequência Relativa [%]
6-Palpitação/taquicardia/arritmia	592	3.31
6- Dispnéia	335	1.87
5- Atividade física	1137	6.36
5- Cirurgia	4212	23.56
6- Sopro	2997	16.76
1- Cardiopatia adquirida	133	0.74
1- Cardiopatia congenica	1262	7.06
6- Dor precordial	650	3.63
6- HAS/dislipidemia/obesidade	424	2.37
6- Cianose	166	0.92
Outro	1097	6.13
6- Alteração de pulso/perfusão	5	0.02
6- Cardiopatia na familia	53	0.29
6- Cansaço	18	0.10
7- Outro	456	2.55
5- Uso de cisaprida	9	0.05
6- Cianose e dispnéia	5	0.02

Tabela 13: Distribuição de Frequência da variável Peso

PESO	Frequência Absoluta	Frequência Relativa [%]
$[0, 35)$	10082	66.750
$[35, 70)$	3059	20.250
$[70, 105)$	203	1.340
$[105, 140)$	11	0.070
$[140, 175)$	1	0.006

Tabela 14: Distribuição de Frequência da variável ALtura

ALTURA	Frequência Absoluta	Frequência Relativa [%]
$[0, 40)$	10	0.066
$[40, 80)$	2104	13.931
$[80, 120)$	4462	29.543
$[120, 160)$	4640	30.722
$[160, 200)$	801	5.303

Tabela 15: Distribuição de Frequência da variável IMC

IMC	Frequência Absoluta	Frequência Relativa [%]
$[0, 12)$	166	1.099
$[12, 24)$	10776	71.350
$[24, 36)$	715	4.734
$[36, 48)$	21	0.139
$[48, 60)$	3	0.019

A Tabela 16 denota as diferenças apontadas anteriormente. Deste modo é analisado que cerca de 65% dos pacientes tem idade entre 0 e 7 anos.

0.4.5 Variável PULSOS

Pelas mudanças realizadas na base de dados o atributo PULSOS passou a ter apenas 33 valores faltantes. O tipo de PULSOS foi mantido, apenas as categorias que possuem nomes sinônimos que foram unificadas e por consequência restaram apenas 5 categorias. A Tabela 17 mostra a distribuição de frequência desta variável.

Do mesmo modo que a variável original sem mudanças, a categoria de PULSOS que apresenta maior prevalência é a categoria NORMAIS. Isto implica que cerca de 98% dos pacientes não apresentam anomalias na verificação de seus pulsos.

0.4.6 Variável PA SISTOLICA

A normalização da pressão sistólica segue a mesma linha da frequência cardíaca, pois parta da mesma ideia, logo, segue de uma mesma normalização. No tópico de descrição da variável de frequência cardíaca será dado uma melhor descrição sobre.

Após as mudanças citadas esta variável passou a ter 5645 (aproximadamente 37% da base de dados) valores faltantes. Além de que este atributo foi discretizado em 5 categorias distintas. No entanto pelas restrições impostas na idade dos pacientes presentes na base de dados esta variável passou a ter apenas 3 categorias. A Tabela 18 mostra uma distribuição de frequência de PA SISTOLICA.

A partir da Tabela 18 é notável uma diferença em PA SISTOLICA em relação aos valores originais. Devido as restrições de idade a maior concentração de PA SISTOLICA passou a ser no intervalo de 100 a menor que 150 o que representa cerca de 44.11% das ocorrências desta variável. Além de que o número de valores que eram inconsistentes diminui drasticamente restando apenas valores que podem ser considerados como possíveis para seres humanos.

0.4.7 Variável PA DIASTOLICA

Do mesmo modo que ocorreu em PA SISTOLICA, logo, segue de uma mesma normalização. No tópico de descrição da variável de frequência cardíaca o assunto será melhor descrito.

Após estas mudanças citadas e as demais alterações feitas na base de dados está variável ficou com 5657 (aproximadamente 37% da base de dados) valores faltantes e 3 categorias distintas. A Tabela 19 apresenta uma distribuição de frequência de PA DIASTOLICA.

Com base na Tabela 19 é visto que cerca de 62% das manifestações de pressão arterial diastólica estão no intervalo fechado em 50 e aberto em 100.

0.4.8 Variável B2

Em relação à esta variável as mudanças feitas foram apenas unificadas as categorias que possuíam nomes sinônimos mas que seus valores tinham o mesmo significado. Deste modo restaram apenas 5 categorias. Por questões das restrições dadas a idade dos pacientes esta variável passou de 1179 para 16 valores faltantes. A Tabela 20 apresenta uma distribuição de frequência de B2.

Com base nos dados apresentados na Tabela 20 foi analisado que a categoria NORMAL ainda continua predominante com uma frequência relativa de aproximadamente 96% das ocorrências de B2. Isto implica que o tipo do segundo som do coração de 14473 pacientes é do tipo normal.

0.4.9 Variável SOPRO

As mudanças em SOPRO se basearam basicamente na unificação de categorias que apresentavam nomes sinônimos. Devido as restrições impostas nos valores que são descritas neste trabalho restaram apenas 9 valores faltantes e 5 categorias distintas nesta variável. A Tabela 21 mostra uma distribuição de frequência de SOPRO.

A Tabela 21 aponta que cerca de 65% dos pacientes não apresentam murmurio no coração. Isto implica em uma porcentagem um pouco maior em relação aos dados originais.

0.4.10 Variável FC

A normalização de dados da variável de frequência cardíaca foi realizada entre 0 e 250, pois segundo o artigo de [7], ele relata que a menor frequência cardíaca em um humano acordado é de 40 batimentos por minuto e a máxima considerada pela equipe médica é de 190 batimentos por minuto para crianças do nascimento ao primeiro mês de vida. Foi definido o intervalo máximo até ao valor de 250 batimentos por minuto por que é imprescindível que seja captado todas as pressões artérias que estejam acima da normalidade, mas ainda tentando prevenir os valores fora do esperado.

A variável FC foi convertida para o tipo numérico quantitativa contínua. Deste modo foi possível discretizá-la para diminuir o número de categorias presentes. Além de que só foram considerados frequências que estavam no intervalo fechado de 40 a 250 pois este intervalo reflete melhor os valores reais de frequência cardíaca em humanos. Após as restrições usadas na base de dados esta variável passou a ter 773 valores faltantes o

Tabela 16: Distribuição de Frequência da variável Idade

IDADE	Frequência Absoluta	Frequência Relativa [%]
$[0, 4)$	5620	37.211
$[4, 8)$	4221	27.948
$[8, 12)$	3455	22.876
$[12, 16)$	1592	10.540
$[16, 20)$	215	1.423

Tabela 17: Distribuição de Frequência da variável PULSOS

PULSOS	Frequência Absoluta	Frequência Relativa [%]
NORMAIS	14941	98.927
OUTRO	38	0.251
FEMORAIS DIMINUIDOS	37	0.245
DIMINUIDOS	11	0.072
AMPLOS	43	0.284

Tabela 18: Distribuição de Frequência da variável PA SISTOLICA

PA SISTOLICA	Frequência Absoluta	Frequência Relativa [%]
$[50, 100)$	2758	18.261
$[100, 150)$	6663	44.117
$[150, 200)$	37	0.244

Tabela 19: Distribuição de Frequência da variável PA DIASTOLICA

PA DIASTOLICA	Frequência Absoluta	Frequência Relativa [%]
$[0, 50)$	151	0.999
$[50, 100)$	9266	61.352
$[100, 150)$	29	0.192

Tabela 20: Distribuição de Frequência da variável B2

B2	Frequência Absoluta	Frequência Relativa [%]
NORMAL	14473	95.828
DESDOB FIXO	174	1.152
OUTRO	94	0.622
HIPERFONETICA	274	1.814
UNICA	72	0.476

Tabela 21: Distribuição de Frequência da variável SOPRO

SOPRO	Frequência Absoluta	Frequência Relativa [%]
SISTOLICO	5260	34.827
AUSENTE	9778	64.742
CONTINUO	44	0.291
DIASTOLICO	9	0.059
SISTOLICO E DIASTOLICO	3	0.019

que é uma grande diferença em relação ao número original. A Tabela 22 mostra uma distribuição de frequência de FC.

Tabela 22: Distribuição de Frequência da variável FC

FC	Frequência Absoluta	Frequência Relativa [%]
$[0, 50)$	3	0.019
$[50, 100)$	10252	67.880
$[100, 150)$	3944	26.114
$[150, 200)$	126	0.834
$[200, 250)$	5	0.033

Com base na Tabela 22 é observado que cerca de 68% dos pacientes apresentam uma frequência cardíaca no intervalo fechado de 50 a 100. Esta visualização nos dados originais não era possível devido a distribuição de categorias que era utilizada.

0.4.11 Variável HDA 1

As mudanças em HDA 1 basearam basicamente na mudança de valores faltantes. Esta mudança de valores faltantes é refletida por causa das restrições impostas na base de dados que foram descritas neste trabalho. Deste modo esta variável passou a ter 3771 (aproximadamente 25% da base de dados) valores faltantes com 8 categorias distintas.

Tabela 23: Distribuição de Frequência da variável HDA 1

HDA1	Frequência Absoluta	Frequência Relativa [%]
Palpitacao	539	3.562
Dispneia	669	4.429
Assintomático	8440	55.882
Dor precordial	805	5.33
Desmaio/tontura	252	1.668
Outro	197	1.304
Cianose	258	1.708
Ganho de peso	172	1.138

A Tabela 23 mostra que 8440 dos pacientes se mostraram assintomáticos em relação ao histórico da primeira doença apresentada. Apesar das mudanças refletidas em HDA 1 ainda é visto que a diferença percentual entre a categoria Assintomático com as demais ainda é extremamente considerável.

0.4.12 Variável SEXO

A categoria de sexo indeterminada foi desconsiderada e seus valores foram considerados como valores faltantes. Uma vez que acredita-se que uma indeterminação na descrição de um sexo de um paciente não irá trazer implicações em definições de cardiopatias. Além de que as categorias que apresentavam nomes sinônimos foram unificadas

restando apenas 2 categorias. Após as mudanças feitas na base de dados e nesta variável ela passou a dispor de 420 (aproximadamente 3% da base de dados) valores faltantes. A Tabela 24 apresenta uma distribuição de frequência de Sexo.

Tabela 24: Distribuição de Frequência da variável SEXO

SEXO	Frequência Absoluta	Frequência Relativa [%]
MASCULINO	8612	57.021
FEMININO	6071	40.197

A Tabela 24 aponta que a base de dados possui mais pacientes do sexo masculino do que feminino.

0.4.13 Variável MOTIVO

Esta é uma nova variável que representa a substituição dos valores de MOTIVO 2 em MOTIVO 1. Esta ação foi realizada devido ao fato de que os valores de MOTIVO 2 são apenas uma especificação do que é apresentado em MOTIVO 1. Desta forma foi decidido realizar a junção destes atributos. Após estas ações a variável MOTIVO apresentou 138 (aproximadamente 1% da base de dados) valores com 19 categorias distintas. A Tabela 25 apresenta a distribuição de frequência de MOTIVO.

Tabela 25: Distribuição de Frequência da variável MOTIVO

MOTIVO	Frequência Absoluta	Frequência Relativa [%]
6-Palpitação/Taquicardia/Arritmia	536	3.548
6-Dispneia	288	1.906
6-ParecerCardiologico	2303	15.248
6-AtividadeFisica	1010	6.687
2-Check-UP	899	5.952
5-Cirurgia	3904	25.849
1-CardiopatiaAdquirida	122	0.807
1-CardiopatiaCongenica	1045	6.919
6-DorPrecordial	593	3.926
6-Has/Dislipidemia/Obesidade	387	2.562
6-SuspeitaDeCardiopatia	102	0.675
6-Sopro	2631	17.42
6-Cianose	126	0.834
Outro	943	6.243
6-AlteracaoDePulso/Perfusao	4	0.026
6-Cansaco	16	0.105
6-CardiopatiaNaFamilia	45	0.297
5-UsoDeCisaprida	8	0.052
6-CianoseEDispneia	3	0.019

A Tabela 25 aponta que em cerca 26% das vezes o motivo dos pacientes serem encaminhados para a clinica de cardiologia é a necessidade de se realizar uma cirurgia.

0.4.14 Variável NORMALXANORMAL

O atributo classe NORMAL x ANORMAL não teve o seu tipo alterado o que implica que seu tipo ainda continua categórico. Neste atributo classe NORMAL X ANORMAL foi unificado todas as categorias que possuíam nomes sinônimos foram mescladas e restaram apenas duas categorias. Além de que todas as instâncias que tinham o campo de NORMALXANORMAL como valor faltante foram retiradas da base de dados. Deste modo esta variável não apresenta valores faltantes. O resultado final após as mudanças é mostrado na Tabela 26.

Tabela 26: Distribuição de Frequência da variável NORMAL X ANORMAL

NORMAL X ANORMAL	Frequência Absoluta	Frequência Relativa [%]
Anormal	5976	39.568
Normal	9127	60.431

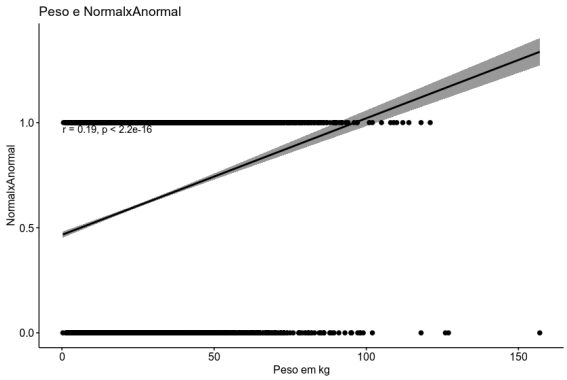
A Tabela 26 mostra que há 5976 ocorrências da classe negativa (categoria Anormal) e 9127 ocorrências da classe positiva (categoria Normal). Por meio destes valores é visto que ainda há um desbalanceamento entre as classes positiva e negativa mas que é ligeiramente menor que o que fora encontrado na base de dados original.

0.5 Análise

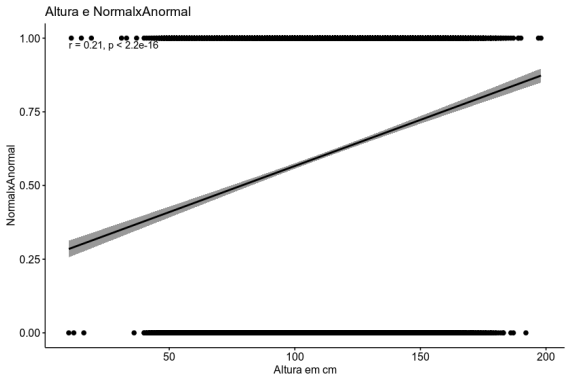
0.5.1 Análise Bivariada

Para a análise bivariada foi decidido utilizar a Correlação de Pearson. A Correlação de Pearson é uma medida de associação bivariada do grau de relacionamento entre duas variáveis. Além de que ela é uma medida da variância que é compartilhada entre duas variáveis [3]. Este coeficiente varia entre o intervalo fechado dos valores -1 e 1, onde 1 implica em um grau de correlação perfeito e de forma análoga -1 significa um grau de relação totalmente inexistente.

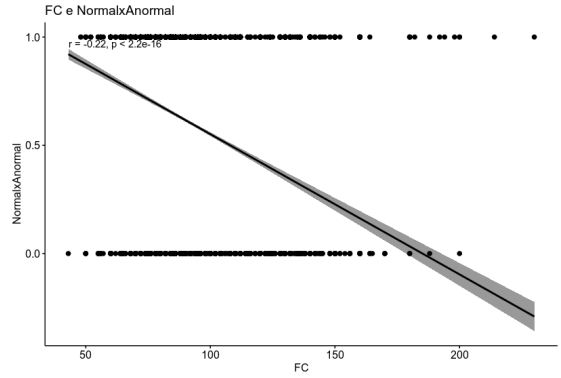
As Figuras 6b a 6g mostram a correlação de Pearson com entre as variáveis Peso, Altura, IMC, Idade, PA SISTOLICA, PA DIASTOLICA e FC com NORMALXANORMAL. Estas relações possuem graus de correlação 0.19, 0.21, -0.22, 0.073, 0.18, 0.00059, 0.024 com NORMALXANORMAL. A partir destes valores é observado que apenas IMC que possui uma correlação negativa com NORMALXANORMAL. Além de que os maiores graus de correlacionamento com NORMALXANORMAL são das variáveis Altura e Peso.



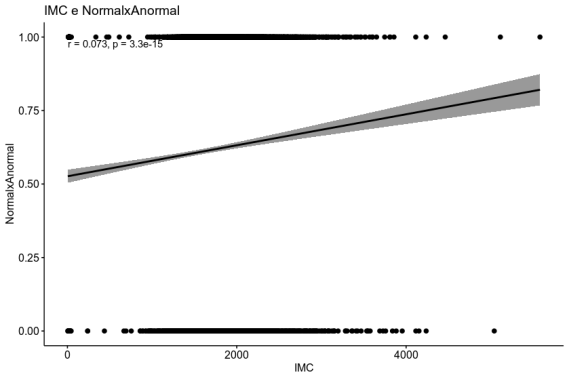
(a) PESO E NORMALXANORMAL



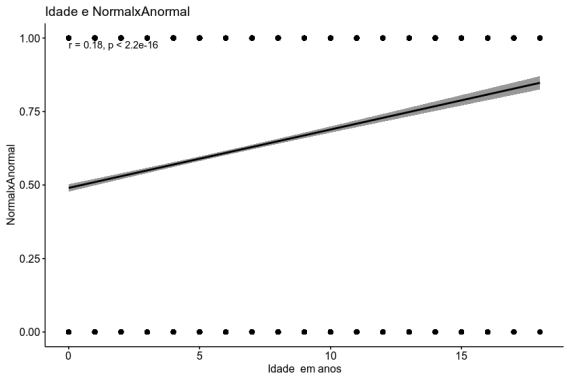
(b) ALTURA E NORMALXANORMAL



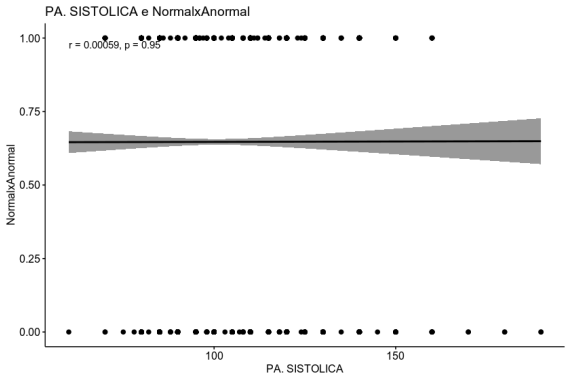
(c) FC E NORMALXANORMAL



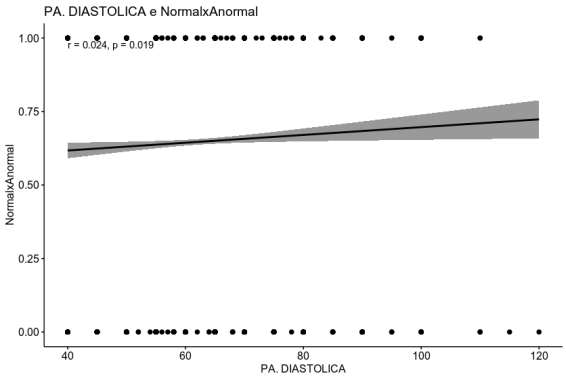
(d) IMC E NORMALXANORMAL



(e) IDADE E NORMALXANORMAL



(f) PA SISTOLICA E NORMALXANORMAL



(g) PA DIASTOLICA E NORMALXANORMAL

0.5.2 Análise Multivariada

0.5.2.1 Regras de Associação

Para a análise multivariada foi utilizado Regras de Associação. Regras de Associação é uma técnica de mineração de dados. Pode ser considerada como uma das formas mais comuns de descobertas de padrões dentro da abordagem de aprendizado não supervisionado. Tem em seu uso geral a realização de uma análise exploratória dos dados [5].

Regras de associação são implicações que assumem a forma:

$$A \Rightarrow B,$$

que é interpretada como "**Se** A **então** B", na qual:

- A é chamado de *antecedente* da regra.
- B é chamado de de *consequente* da regra.
- $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ e $A \cap B = \emptyset$.

As métricas utilizadas para avaliar as Regras de Associação aqui serão Suporte, Confiança e Lift. O Suporte pode ser definido como a frequência de $A \cup B$. A Confiança pode ser definida como a probabilidade de B ocorrer dado que houve ocorrência de A. O Lift de uma regra indica o quanto mais frequente se torna B dado que A ocorreu.

Foi decidido utilizar o algoritmo Apriori proposto por [1]. Ele é um algoritmo utilizado para o aprendizado de Regras de Associação. Ele se baseia na questão de que se conjuntos de itens são frequentes então a combinação destes também pode gerar novos conjuntos que também sejam frequentes [5].

0.5.2.2 Regras Geradas

Todas as variáveis da base de dados pré-processada foram utilizadas para gerar as Regras de Associação. Os parâmetros utilizados foram: suporte mínimo: 60%, confiança mínima: 90%. Além de que não é possível definir um valor mínimo para lift. O motivo da escolha se baseia nas circunstancias de regras que possuíam uma frequência superior a metade das instâncias da base de dados e que estas regras pudessem ser consideravelmente confiáveis dado o valor de 90%. Foram geradas 13 regras de associação que são mostradas nas Tabelas 27 e 28.

A Tabela 27 representa as regras quem implicam em pulsos normais. Por meio destas regras é observado que as categorias que aparecem como antecedente das regras estão intimamente relacionadas com categoria NORMAIS da variável Pulso. Estas regras dizem que quando tiver a ocorrências individuais de NORMALXANORMAL=NORMAL,

PA.DISTOLICA= $[50, 100)$, SOPRO=AUSENTE, PESO= $[0, 35)$, FC= $[50, 100)$, IMC= $[12, 24)$ e B2=NORMAL, elas irão acarretar em pulsos normais em 99.52%, 99.44%, 99.50%, 98.98%, 99.46%, 99.28% e em 99.22% das vezes respectivamente.

Tabela 27: Regras que estão associadas à pulsos normais

Regra				Conf.[%]	Sup.[%]	Lift
$NORMAL.X.ANORMAL = NORMAL$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	99.52	60.14	1.00
$PA.DIASTOLICA = Class[50, 100)$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	99.44	61.01	1.00
$SOPRO = AUSENTE$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	99.50	64.42	1.00
$PESO = Class[0, 35)$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	98.98	66.07	1.00
$FC = Class[50, 100)$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	99.46	67.51	1.00
$IMC = Class[12, 24)$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	99.28	70.84	1.00
$B2 = NORMAL$	$\Rightarrow$	$PULSOS\ NORMAIS$	=	99.22	95.08	1.00

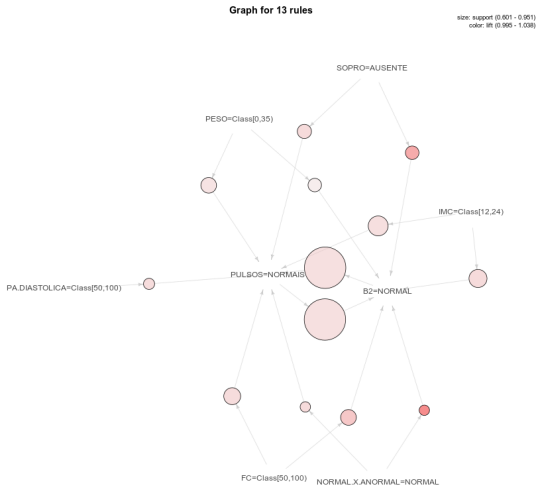
Tabela 28: Regras que estão associadas à B2 normais

Regra				Conf.[%]	Sup.[%]	Lift
$NORMAL.X.ANORMAL = NORMAL$	$\Rightarrow$	$B2\ NORMAL$	=	99.44	60.08	1.03
$SOPRO = AUSENTE$	$\Rightarrow$	$B2\ NORMAL$	=	98.42	63.72	1.02
$PESO = Class[0, 35)$	$\Rightarrow$	$B2\ NORMAL$	=	95.30	63.62	0.99
$FC = Class[50, 100)$	$\Rightarrow$	$B2\ NORMAL$	=	97.47	63.72	1.01
$IMC = Class[12, 24)$	$\Rightarrow$	$B2\ NORMAL$	=	96.42	68.80	1.00
$PULSOS = NORMAIS$	$\Rightarrow$	$B2\ NORMAL$	=	96.11	95.08	1.00

A Tabela 28 representa as regras que implicam em B2 normais. As categorias apresentadas nos antecedentes das regras estão intimamente ligados com a ocorrência do segundo tipo do som do coração normal. Além disto também é possível abstrair destas regras que quando houver ocorrências individuais de NORMALXANORMAL=NORMAL, PA.DISTOLICA= $[50, 100)$, SOPRO=AUSENTE, PESO= $[0, 35)$, FC= $[50, 100)$, IMC= $[12, 24)$ e B2=NORMAL, elas irão acarretar em um segundo som do coração normal em 99.44%, 63.72%, 63.62%, 95.30%, 97.47%, 96.42% e em 96.11% das vezes respectivamente.

Ao fazer a analise das Tabelas 27 e 28 em conjunto, foi visto que as categorias apresentadas como antecedentes da regra são as mesmas em ambas as tabelas com apenas

Figura 6: Grafo gerado a partir das regras mostradas. Onde o tamanho do nó é o suporte da regra e o gradiente de cores é representado pelo valor do lift da regra



a exceção do antecedente PA.DIASTOLICA $= Class[50, 100)$ que aparece apenas na Tabela 27. Isto implica que estas categorias estão fortemente relacionadas a pulsos normais e a um segundo som do coração normal. Além disto também foi observado a relação de biimplicação entre as categorias de pulsos normais e um segundo som do coração normal. Esta relação acarreta em que sempre a ocorrência de uma destas duas categorias irá implicar na aparição da outra.

A Figura 6 mostra um grafo que representa como as regras estão relacionadas entre si. O tamanho dos nós é dado pelo suporte de cada regra em questão. O gradiente de cores apresentado nos nós é dado pelo valor do lift de cada regra.

0.5.3 Aprendizado de Máquina

Para o processo de Aprendizado de Máquina foi usado os respectivos algoritmos **K-vizinhos mais próximos; Arvore de decisão; Redes Bay-seanas**. Sendo eles representados no framework Weka como: **IBK; J48; NaiveBayes** e o **BayesNet**, que é o algoritmo da **Rede Bayseana TAN**.

0.5.4 K-Vizinhos mais próximos

É um algoritmo baseado em distâncias. Este algoritmo classifica um novo objeto baseado nos exemplos do conjunto de treinamento que são próximos a ele. É considerado um algoritmo preguiçosos porque não aprende um modelo para os dados a serem testados ele apenas memoriza a posição dos objetos de treinamento [2]. O parâmetro de número de vizinhos utilizado foi 5.

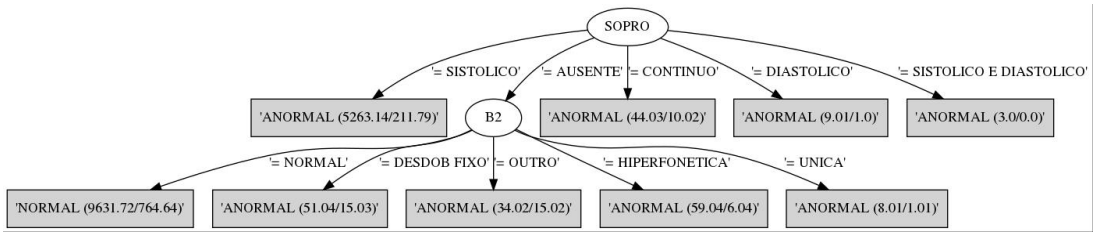
0.5.5 Árvore de Decisão

A Árvore de decisão é uma representação simples de conhecimento e ainda é uma forma bem sucedida de algoritmo de aprendizado [4, 11]. Para realizar a construção de uma árvore de decisão é preciso calcular o **ganho de informação**, com o objetivo de determinar quais os atributos que caracterizam melhor o problema em questão.

A árvore de decisão obtém como entrada um objeto e retorna uma decisão. Esta decisão é tomada através de uma sequência de testes. Cada nó interno corresponde a um teste de valor de uma das propriedades de um determinado atributo e os ramos de cada nó são fixados como os possíveis valores do nó em questão. Cada folha na árvore especifica uma certa decisão [11].

O **ganho de informação** é baseado em uma medida denominada **entropia** que caracteriza a impureza ou pureza de determinado conjunto de atributos [9, 10].

Figura 7: Árvore de Decisão gerada a partir da base de dados



A Figura 7 mostra uma árvore de decisão gerada a partir dos dados pré processados. A partir desta árvore é possível notar o alto ganho de informação que as variáveis SOPRO e B2 carregam com elas. Isto é elas conseguem influenciar fortemente no poder de classificação de um paciente com uma cardiopatia ou não.

0.5.6 Naive Bayes

Os modelos gráficos probabilísticos, ou redes bayseanas, utilizam o conceito de *independência condicional* entre variáveis para obter um equilíbrio entre os números de parâmetros de parâmetros a calcular e a representação de dependências entre as variáveis. Esses modelos representam a distribuição de probabilidade conjunta de um grupo de variáveis aleatórias em um domínio específico [2].

O classificador Naive Bayes é o mais simples desses modelos, na medida em que pressupõe que todos os atributos do exemplo são independentes uns dos outros, dado o contexto da classe [8].

0.5.7 Rede Baysena TAN

É uma extensão do Naive Bayes, pois permite o relaxamento da hipótese de independência condicional entre atributos dado a classe. O Classificador TAN foi proposto por Friedman e Goldszmidt (1997) e possibilita representar dependências entre pares de atributos. No Classificador TAN a dependência entre atributos devem ser representadas pela estrutura de uma árvore, ou seja, cada atributo deve ter no máximo um pai [6].

0.5.8 Performance dos algoritmos

A Figura 8 mostra uma tabela que foi gerada por meio do framework Weka. Estes valores apresentam os desempenho dos 4 algoritmos apresentados na base de dados preprocessada. As colunas da tabela mostrada na Figura 8 representam as métricas que foram utilizadas para avaliar o desempo dos algoritmos propostos. Uma descrição sobre cada métrica é dada nesta mesma subseção.

Figura 8: Performance dos algoritmos escolhidos

	accuracy	sd	recall	sd	precision	sd	specificity	sd	f_measure	sd	auc_roc	sd
Decision Tree (J48)	0.93	0.06	0.87	0.01	0.95	0.01	0.97	0.01	0.91	0.01	0.92	0.01
Naive Bayes	0.91	0.06	0.87	0.01	0.91	0.01	0.94	0.01	0.89	0.01	0.93	0.01
Bayes Net (TAN)	0.91	0.06	0.87	0.01	0.91	0.01	0.94	0.01	0.89	0.01	0.94	0.01
Lazy.IBK (KNN)	0.92	0.06	0.85	0.01	0.96	0.01	0.98	0.01	0.90	0.01	0.93	0.01

As proximas subseções irão descrever as métricas utilizadas para a avaliação dos classificadores produzidos. Além de que também será discutido o desempenho destes algoritmos em relação a cada métrica aqui apresentada.

0.5.8.1 Accuracy

A métrica **accuracy** é a taxa de acertos total de determinado algoritmo de mineração de dados. Em relação ao número de acertos total o algoritmo que se sobresaiu foi à arvore de decisão, isto mesmo levando em consideração os valores de desvios padrão. No entanto todos os outros algoritmos tiveram uma taxa de acerto acima de 90%. Isto implica em um bom desempenho geral dado que os dados em questão tem uma natureza médica.

Esta alta taxa de acertos pode refletir na questão de que estes classificadores não apresentam uma baixa taxa de erros tanto na classe positiva como na negativa.

0.5.8.2 Recall

A métrica **recall** é a taxa de acertos na classe positiva. Em relação à esta métrica os algoritmos que se sobressaíram foram à arvore de decisão e as duas redes bayseanas. No entanto a diferença apresenta em relação ao KNN não é tão expressiva.

Esta métrica em todos os algoritmos apresentou uma pequena diferença percentual em relação ao desempenho destes classificadores em relação a métrica de acertos totais. Este valor pode refletir na questão de que os classificadores podem estar "engolindo"alguns falsos positivos. Isto é predizendo uma instância como positiva quando ela na verdade é negativa. Por esta diferença percentual ser pequena e a base de dados apresentar uma quantidade considerável de instâncias pode se considerar que não há muitas perdas neste sentido.

0.5.8.3 Precision

A métrica **precision** é a taxa de valores positivos classificados corretamente dentre todos aqueles preditos como positivos. Basicamente a precision é a taxa de acerto em relação a rotulação dos valores. Em relação a esta métrica o algoritmo que se sobressaiu foi o KNN mesmo sendo levando em consideração o desvio padrão. Apesar de que os demais algoritmos apresentam uma considerável performance em precision. Isto quer dizer que geralmente a taxa de exemplos positivos classificados corretamente dentre todos os preditos como positivos é alta mesmo que passem alguns falsos positivos na avaliação.

0.5.8.4 Specificity

A métrica **specificity** é taxa de acertos na classe negativa. O que corresponde aos verdadeiros negativos, que foram preditos como negativos. realmente o são. Em relação a esta métrica o algoritmo que obteve melhor desempenho foi o KNN. O que implica que dos valores que foram preditos como negativos pelo KNN, 97% deles são realmente negativos. Em relação aos demais algoritmos é visto que eles obtiveram valores de specificity acima de 90% o que acarreta em uma boa taxa de acertos da classe negativa entre todos os algoritmos.

0.5.8.5 F_measure

A métrica *F_measure* é a média Harmônica ponderada de **precision** e **recall**. É uma métrica que tem o intuito de trazer um número único que ira indicar a qualidade geral do classificador e geralmente é uma boa métrica para questões de classes desproporcionais [2]. Em relação a esta medida o algoritmo que obteve um melhor desempenho foi à arvore de decisão. O que implica que este classificador possui uma boa combinação de acertos na classe positiva com uma boa taxa de rotulação das classes.

0.5.8.6 Area under curve roc

Na métrica **auc_roc** há a seguinte característica: quanto mais próximo do valor 1 melhor o classificador (sendo 1 o valor ideal) do mesmo modo quanto mais próximo de 0.5 há uma maior chance de valores inconsistentes. Uma vez que uma taxa próxima de

0.5 seria o equivalente a uma comparação com uma moeda. Isto é na moeda só ha duas possibilidades (uma cara e outra coroa) o que acarreta em 50% de chances de acerto em uma destas possibilidades.

Dos algoritmos em questão o que apresentou uma maior taxa auc_roc foi a rede bayseana TAN, isto mesmo levando em consideração o valor de desvio padrão. Da mesma forma que ocorreu em outras métricas, ná ha uma diferença considerável em cada algoritmo. Deste modo é admissível assumir que eles tiveram desempenhos parecidos.

0.6 Conclusões

O processo de KDD se feito de forma correta contribui em muito nos resultados apresentados pelos algoritmos de aprendizado de máquina. Além de que este correto tratamento dos dados pode impactar a descoberta de padrões e as tomadas de decisões a serem feitas em diversas áreas do conhecimento.

O objetivo deste trabalho foi aplicar as técnicas de KDD previamente conhecidas para a classificação de pacientes em relação a presença ou não de cardiopatias. Para isto foi utilizado a base de dados da UCMF que depois de tratada passou a ter 15103 instâncias com 14 atributos. Os seguintes algoritmos Árvore de decisão, redes bayseanas e k-vizinhos mais próximos foram aplicados para a geração de classificadores. Além de que o algoritmo Apriori foi aplicado para realizar uma análise exploratória dos dados para a descoberta de padrões.

Os resultados obtidos apresentaram valores satisfatórios em relação à área de conhecimento usada nos dados. O que pode apontar que a mineração de dados é uma ferramenta a ser considerada como um auxílio aos profissionais que atuam nesta área.

Referências

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207–216, 1993. 26

[2] K. Faceli, A. L. Carolina, and J. Gama. *Inteligência artificial uma abordagem de aprendizado de máquina*. Ed. LTC, 2015. 28, 29, 31

[3] D. B. F. Filho and J. A. da Silva Júnior. Desvendando os mistérios do coeficiente de correlação de pearson (r). *Revista Política Hoje*, 18(1):115–146, 2009. 24

[4] S. C. GARCIA. O uso de árvores de decisão na descoberta de conhecimento na área de saúde. Master's thesis, Universidade Federal do Rio Grande do Sul, Oct. 2003. 29

[5] M. Kantardzic. *Data Mining: Concepts, Models, Methods, and Algoritms*. Wiley-IEEE Press, 2011. 26

[6] C. Karcher. Redes bayesianas aplicadas À anÁlise do risco de crÉdito, (2009). *Universidade de São Paulo*, 2009. 30

[7] M. MacGill. *What should my heart rate be?*, 11 2017. 20

[8] A. McCallum and K. Niga. A comparison of event models for naive bayes text classification. 29

[9] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. 29

[10] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. 29

[11] S. J. Russel and Norvig Peter. *Artficial Intellingence: A Modern Approach*. 2 edition, 2003. 29