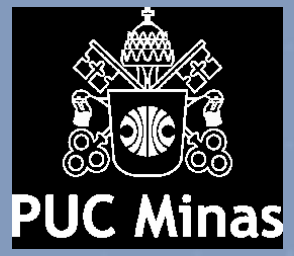


MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA

Luis Enrique Zárate

Exemplo Prático e Desafio

Uso do ambiente Knime

Parte 1

TÍTULO – Exemplo Prático e Desafio

Exemplo Prático e Desafio: Construção de um modelo de aprendizado para descrever o perfil dos indivíduos que sofrem com a doença do “Colesterol Alto”.

Objetivo: Aplicar os conceitos de modelagem e preparação de dados para obtenção da base de dados aplicada para descrição do perfil de pessoas que sofrem com “Colesterol Alto”.

Open for Innovation KNIME

https://www.knime.com/

IBGE
Instituto Brasileiro de Geografia e Estatística

https://www.ibge.gov.br/

LICAP
Laboratório de Inteligência Computacional Aplicada

licap.icei.pucminas.br

PUC Minas

Base de dados

Origem dos dados: Base de Dados sobre Pesquisa Nacional de Saúde - PNS 2013: percepção do estado de saúde, estilos de vida e doenças crônicas.
Instituto Brasileiro de Geografia e Estatística

- A base de dados original pode ser encontrada no seguinte endereço: https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?=&t=microdados

- O dicionário de dados pode ser encontrado em: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=PNS/2013/Microdados/Documentacao

- Informações referente ao estudo está disponível na biblioteca nacional do IBGE, pode ser acessado no seguinte endereço: https://biblioteca.ibge.gov.br/visualizacao/livros/liv94074.pdf

Módulos de informação da base de dados

Identificação do questionário	Informações do domicílio (A)	Visitas domiciliares de equipes de saúde (B)	Características gerais dos moradores (C)	Características de educação dos moradores (D)	Informações de trabalho dos moradores (E)
Rendimentos domiciliares (F)	Pessoas com deficiências (G)	Cobertura de Plano de saúde (I)	Utilização de serviços de saúde (J)	Saúde dos indivíduos com mais de 60 anos (K)	Informações de crianças com menos de 2 anos (L)
Características do trabalho (M)	Percepção do estado de saúde (N)	Acidentes e violências (O)	Estilos de vida (P)	Doenças crônicas (Q)	Saúde da mulher (R)
	Atendimento pré-natal (S)	Saúde bucal (U)	Atendimentos médicos (X)	Informações clínicas (W)	

Detalhamento da base de dados

- Dataset original: Atributos = 942, Registros = 205546

- A partir da base de dados original foi extraído um dataset contendo registros de pessoas diagnosticadas com “Colesterol Alto”, e para contrapor ao modelo a ser criado, foi inserida a mesma quantidade de registros de pessoas que não sofrem da doença do “Colesterol Alto”. O total de registros é de 14599.

Procedimentos adotados para Descoberta de conhecimento

- **Construir um mapa conceitual** para o problema de domínio considerado. Utilizar conhecimento tácito e explícito.

- **Selecionar conceitualmente os atributos mais relevantes** da base de dados de acordo ao mapa conceitual, previamente construído. Como resultado deste processo será gerado um conjunto de dados para iniciar o pré-processamento do conjunto de dados.

- **Realizar uma análise exploratória** prévia acerca do número de casos (Colesterol alto) por estado brasileiro e regiões do Brasil. O objetivo é tomar decisão se o modelo vai ser construído a nível brasil, região ou estado.

Procedimentos adotados – Método PICTOREA

- **Realizar uma análise estatística descritiva** (médias, medianas, desvio padrão, moda, histogramas, etc) dos atributos que compõem o conjunto de dados “alvo” do estudo. O objetivo desta etapa é explorar o conjunto de dados.

- Aplicar uma observação univariada para **detectear registros com inconsistências**. Tomar decisão acerca da eliminação ou não desses registros e/ou atributos.

- Aplicar uma **análise acerca da presença de dados vazíos e ausentes**. Avaliar a possibilidade de fusão de atributos de forma a diminuir a presença desses dados.

- Propor e aplicar estratégia univariada para **análise de outliers**. Eliminar registros ou atributos contendo outliers.

Procedimentos adotados – Método PICTOREA

- **Aplicar estratégias univariadas para seleção de atributos**. Sugere-se avaliar o poder discriminativo de cada atributo para separação de classes, se o problema escolhido for de classificação.

- Avaliar a posibilidade da **discretização de atributos** para aplicação de algoritmos de classificação ou clusterização. É sugerido que na presença de dados mixtos se opte por discretizar os dados numéricos, obtendo dados categóricos.

- Realizar uma avaliação final descrevendo as restrições adotadas durante o processo de preparação do conjunto de dados. Avaliar a representatividade do conjunto. Colocar restrições ao modelo caso houver necessidade.

Procedimentos adotados – Método PICTOREA

- Aplicar técnica de aprendizado de máquina para construção dos modelos de aprendizado.

- Validar e avaliar o modelo.

Observações:

A experiência prática mostra que diversas bases de dados, inclusive dentro no mesmo domínio de problema, demandam novas estratégias e procedimentos. A meta sempre deve ser a obtenção de um conjunto de dados relevante ao problema, de forma a aumentar a representatividade e confiabilidade do modelo.

É importante enfatizar que os modelos são construídos a partir de uma amostra de dados. Daí sua capacidade de generalização não é totalmente garantida.

Exemplo Prático e Desafio

Uso do ambiente Knime

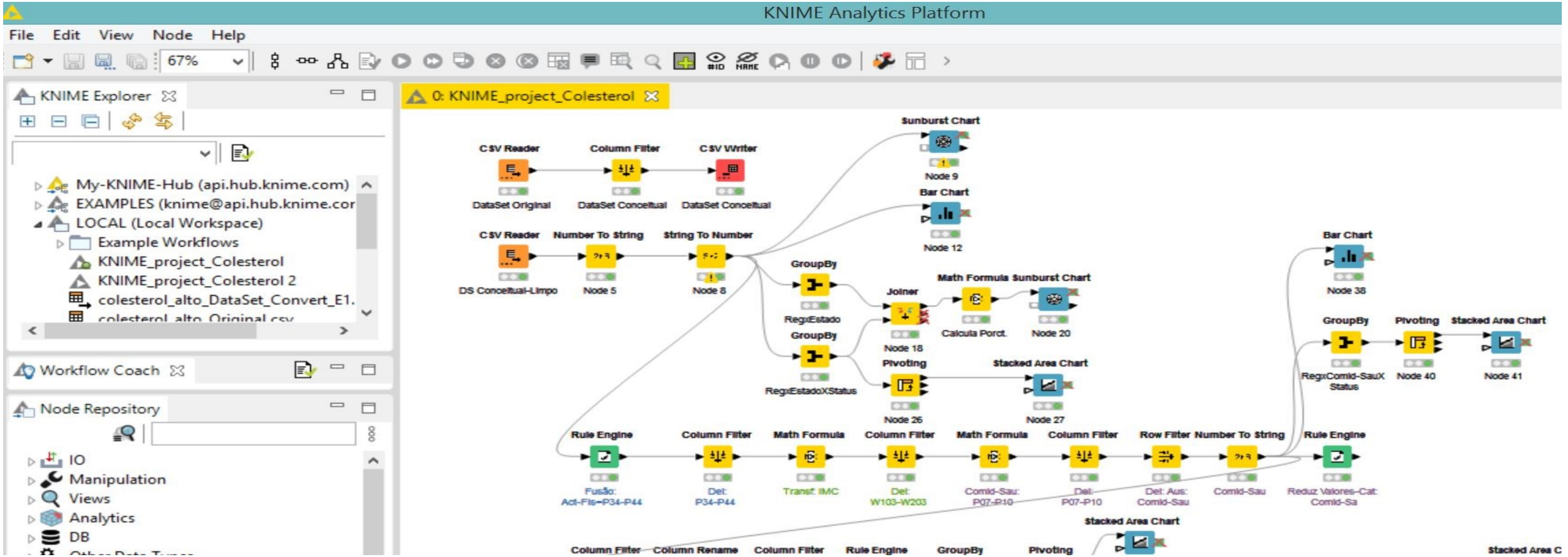
Parte 2

Open for Innovation
KNIME

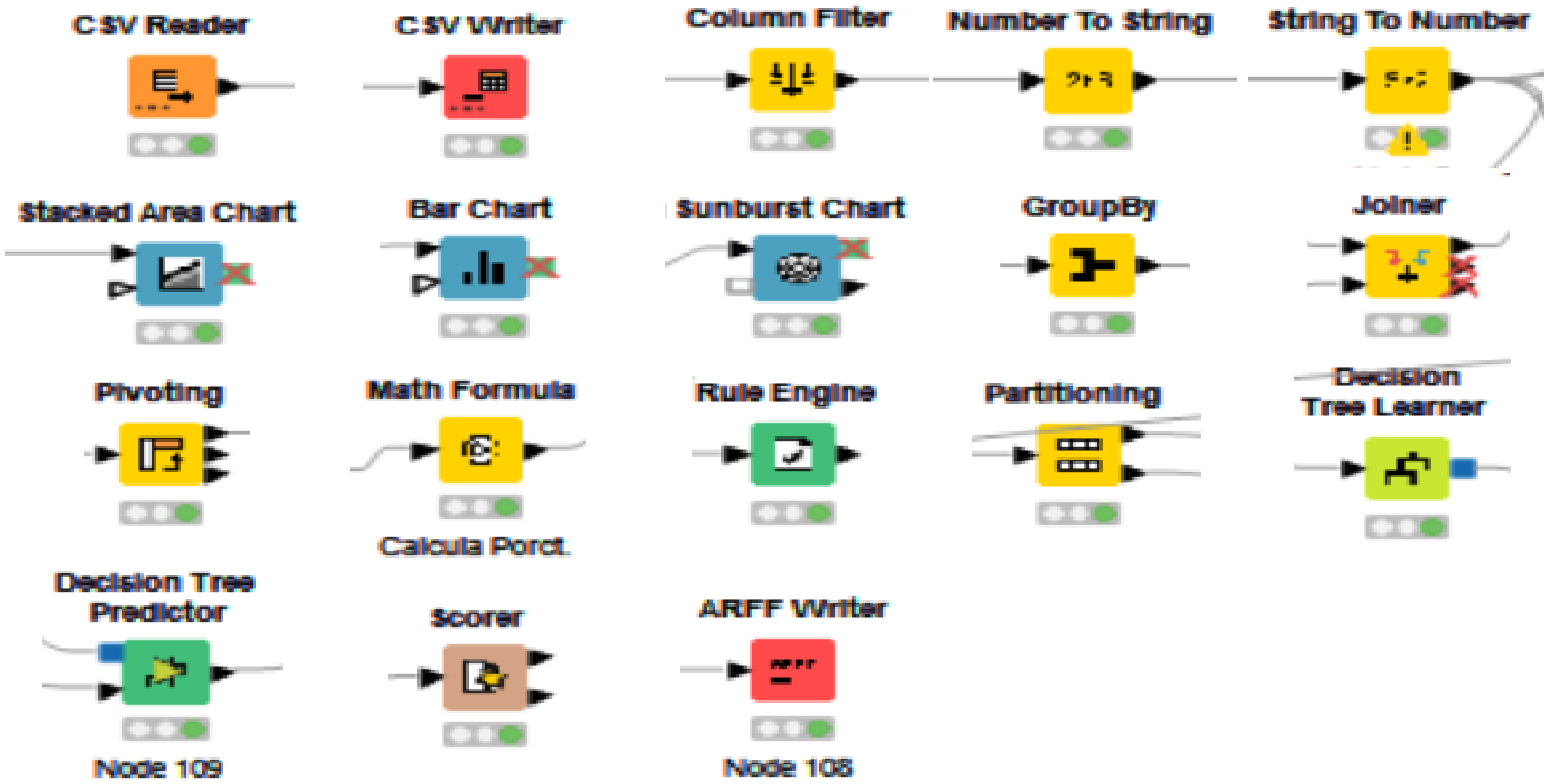
https://www.knime.com

Universität Konstanz, Alemanha, 2004

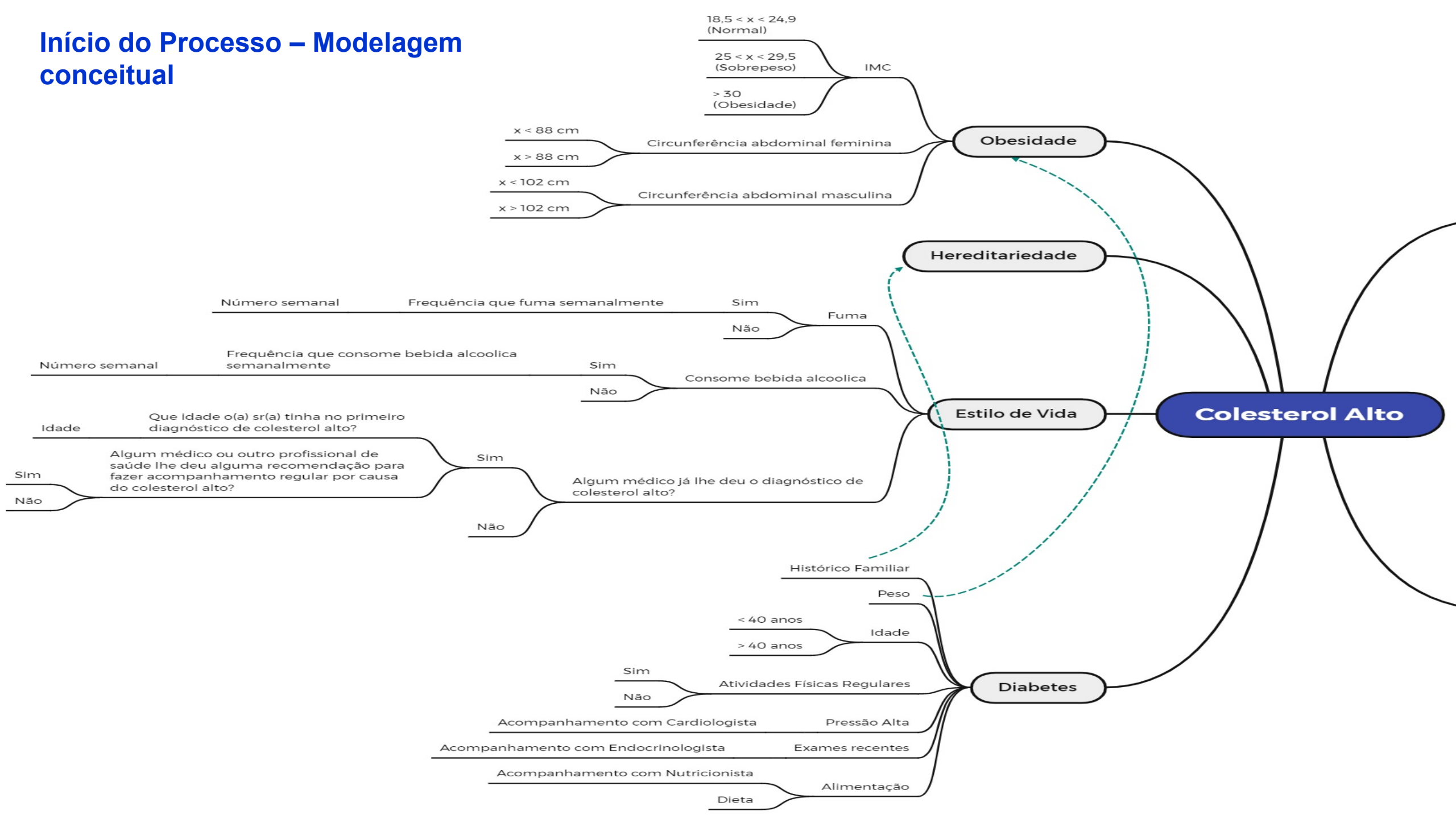
O KNIME (Konstanz Information Miner), é uma plataforma de código aberto para análise de dados, construção de relatórios e integração de dados. O KNIME integra componentes para acesso de dados de diferentes fontes, técnicas de aprendizado de máquina e mineração de dados.
O KNIME permite a construção de processos como *Workflows.*



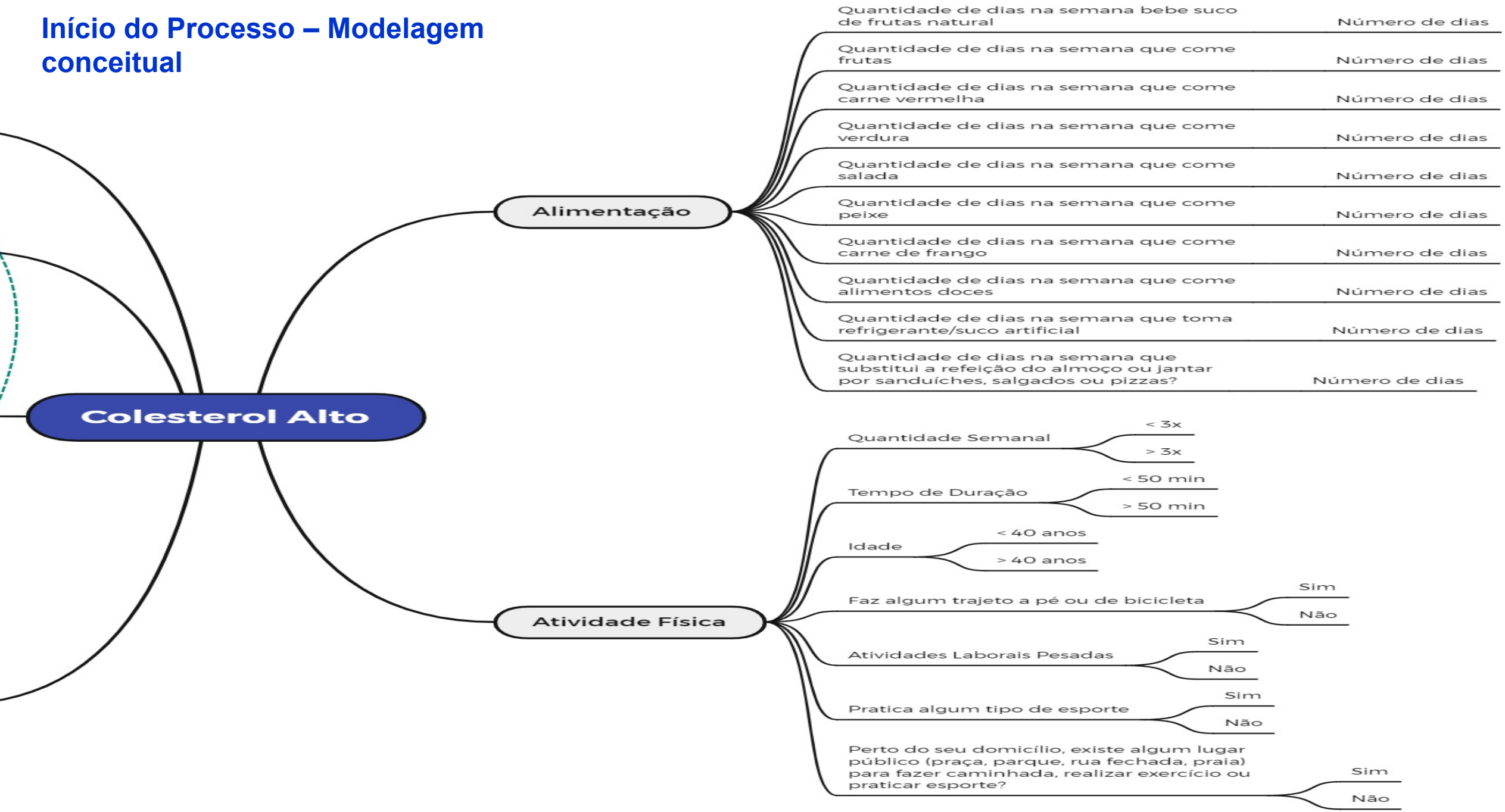
Componentes KNIME utilizados no presente exercício



Início do Processo – Modelagem conceitual



Início do Processo – Modelagem conceitual



Selecionar conceitualmente os atributos relacionadas com o Mapa Conceitual – Total 91 atributos

C006	P012	P021	P034	P03903	P04401	P05402	P05408	P05416	Q061	W00203
C00702	P013	P022	P035	P040	P04403	P05401	P05409	P05417	Q06201	W00303
C00703	P015	P025	P03701	P04101	P04404	P05402	P05410	P05418	Q06202	W00203
P007	P016	P026	P03702	P04102	P050	P05403	P05411	P05419	Q06203	V001
P008	P017	P02601	P038	P042	P051	P05404	P05412	P05421	Q06204	
P009	P018	P027	P039	P04301	P052	P05405	P05413	P05422	Q06205	
P010	P019	P028	P03901	P04302	P053	P05406	P05414	P058	Q06206	
P011	P020	P029	P03902	P044	P05401	P05407	P05415	Q060	W00103	

V001: Estado UF

Q060: Diagnóstico clínico para Colesterol Alto

PUC Minas

Módulo C - Características gerais dos moradores

C1. Quantas pessoas moram neste domicílio: C001

C3. Número de ordem: C00301

Nome: C00302

C4. Condição no domicílio: C004

1. Pessoa responsável pelo domicílio
2. Cônjuge ou companheiro(a) de sexo diferente
3. Cônjuge ou companheiro(a) do mesmo sexo
4. Filho(a) do responsável e do cônjuge
5. Filho(a) somente do responsável
6. Enteado(a)
7. Genro ou nora
8. Pai, mãe, padrasto ou madrasta
9. Sogro(a)
10. Neto(a)
11. Bisneto(a)
12. Irmão ou irmã
13. Avô ou avó
14. Outro parente
15. Agregado(a) - Não parente que não compartilha despesas
16. Convivente - Não parente que compartilha despesas
17. Pensionista
18. Empregado(a) doméstico(a)
19. Parente do(a) empregado(a) doméstico(a)

(siga C6)

C6. Sexo: C006

1. Masculino
2. Feminino

(siga C7)

C7. Data de nascimento: C00701 / C00702 / C00703

dia mês ano

(siga C8)

C8. Idade: C008

(siga C9)

C9. Cor ou raça:(Leia as opções de resposta) C009

1. Branca
2. Preta
3. Amarela
4. Parda
5. Indígena

(Se C008>=10 anos, siga C10.
Se C008<10, passe ao C12.)

Módulo P. Estilos de Vida

Neste módulo, vou lhe fazer perguntas sobre o seu estilo de vida, como hábitos de alimentação, prática de atividade física, uso de bebidas alcoólicas e fumo.

P1. O(A) sr(a) sabe seu peso? (*mesmo que seja valor aproximado*) P001

☐ 1. Sim, qual? P00101 ☐☐☐ Quilograma ☐ 2. Não sabe

(siga P2)

P2. Quanto tempo faz que o(a) sr(a) se pesou da última vez? (Leia as opções de resposta) P002

☐ 1. Menos de 1 semana
☐ 2. Entre 1 semana e menos de 1 mês
☐ 3. Entre 1 mês a menos de 3 meses
☐ 4. Entre 3 meses e menos de 6 meses
☐ 5. Há 6 meses ou mais
☐ 6. Nunca se pesou

(Se C008 (idade) ≥ 30, siga P3. Caso contrário, passe ao P4.)

P3. O(A) sr(a) lembra qual seu peso aproximado por volta dos 20 anos de idade? (*somente para pessoas com 30 anos ou mais*) P003

☐ 1. Sim, qual? P00301 ☐☐☐ Quilograma ☐ 2. Não lembra / Não sabe

(siga P4)

P4. O(A) sr(a) sabe sua altura? (*mesmo que seja valor aproximado*) P004

☐ 1. Sim, qual? P00401 ☐☐☐ Centímetros ☐ 2. Não sabe

(Se C006 = 1, passe ao P6.) (Se C006 = 2, siga P5.)

Ser for mulher com idade entre 18 e 49 anos de idade

P5. A sra está grávida no momento? P005

☐ 1. Sim ☐ 2. Não ☐ 3. Não sabe

(siga P6)

Agora vou lhe fazer perguntas sobre sua alimentação.

P6. Em quantos dias da semana o(a) costuma comer feijão? P006

☐☐ Dias ☐ 0. Nunca ou menos de uma vez por semana

P7. Em quantos dias da semana, o(a) sr(a) costuma comer salada de alface e tomate ou salada de qualquer outra verdura ou legume cru? P007

☐☐ Dias ☐ 0. Nunca ou menos de uma vez por semana

IBGE
Instituto Brasileiro de Geografia e Estatística

Módulo Q. Doenças crônicas

As perguntas deste módulo são sobre doenças crônicas. Vamos fazer perguntas sobre diagnóstico de doenças, uso dos serviços de saúde e tratamento dos problemas.

Q1. Quando foi a última vez que o(a) sr(a) teve sua pressão arterial medida?(Leia as opções de resposta)

Q001

- 1. Há menos de 6 meses
- 2. Entre 6 meses e menos de 1 ano
- 3. Entre 1 ano e menos de 2 anos
- 4. Entre 2 anos e menos de 3 anos
- 5. 3 anos ou mais
- 6. Nunca

(Se Q1=1 a 5, siga Q2. Se Q1=6, passe ao Q29.)

Q2. Algum médico já lhe deu o diagnóstico de hipertensão arterial (pressão alta)?

Q002

- 1. Sim
- 2. Apenas durante a gravidez (só para mulheres)
- 3. Não

(Se Q2=1, siga Q3. Se Q2=2 ou 3, passe ao Q29.)

Q3. Que idade o(a) sr(a) tinha no primeiro diagnóstico de hipertensão arterial (pressão alta)?

Q003

|_|_| Anos

- 0. Menos de 1 ano

(siga Q4)

Q4. O(A) sr(a) vai ao médico/serviço de saúde regularmente por causa da hipertensão arterial (pressão alta)?

Q004

- 1. Sim
- 2. Não, só quando tem algum problema
- 3. Nunca vai

(Se Q4 = 2 ou 3, siga Q5.
Se Q4 = 1, passe ao Q6.)

Q5. Qual o principal motivo do(a) sr(a) não visitar o médico/serviço de saúde regularmente por causa da hipertensão arterial (pressão alta)?

Q005

- 1. O serviço de saúde é muito distante
- 2. O tempo de espera no serviço de saúde é muito grande
- 3. Tem dificuldades financeiras
- 4. Não acha necessário
- 5. O horário de funcionamento do serviço de saúde é incompatível com suas atividades de trabalho ou domésticas
- 6. O plano de saúde não cobre as consultas
- 7. Não sabe quem procurar ou aonde ir
- 8. Dificuldade de transporte
- 9. Outro (Especifique:_____ Q00501 _____)

(siga Q6)

Q56. Alguma vez o(a) sr(a) se internou por causa do diabetes ou de alguma complicação? Q056

1. Sim
2. Não

(Se Q56=1, siga Q57. Se Q56=2, passe ao Q58.)

Q57. Há quanto tempo foi a última internação por causa do diabetes ou de alguma complicação? (Leia as opções de resposta) Q057

1. Há menos de 6 meses
2. Entre 6 meses e menos de 1 ano
3. Entre 1 ano e menos de 2 anos
4. Entre 2 anos e menos de 3 anos
5. Há 3 anos ou mais

(siga Q58)

Q58. Em geral, em que grau o diabetes ou alguma complicação do diabetes limita as suas atividades habituais (*tais como trabalhar, realizar afazeres domésticos, etc.*)?(Leia as opções de resposta) Q058

1. Não limita
2. Um pouco
3. Moderadamente
4. Intensamente
5. Muito intensamente

(siga Q59)

Q59. Quando foi a última vez que o(a) sr(a) fez exame de sangue para medir o colesterol e triglicerídeos?(Leia as opções de resposta) Q059

1. Há menos de 6 meses
2. Entre 6 meses e menos de 1 ano
3. Entre 1 ano e menos de 2 anos
4. Entre 2 anos e menos de 3 anos
5. Há 3 anos ou mais
6. Nunca fez

(Se Q59=1 ao 5, siga Q60. Se Q59=6, passe ao Q63.)

Q60. Algum médico já lhe deu o diagnóstico de colesterol alto? Q060

1. Sim
2. Não

(Se Q60=1, siga Q61. Se Q60=2, passe ao Q63.)

Q61. Que idade o(a) sr(a) tinha no primeiro diagnóstico de colesterol alto? Q061

Anos

0. Menos de 1 ano

(siga Q62)

Q62. Algum médico ou outro profissional de saúde lhe deu algumas das seguintes recomendações por causa do colesterol alto?(Leia as opções de resposta)

a. Manter uma alimentação saudável (com frutas e vegetais) Q06201 — 1. Sim 2. Não (siga Q62b)
b. Manter o peso adequado Q06202 — 1. Sim 2. Não (siga Q62c)
c. Prática de atividade física Q06203 — 1. Sim 2. Não (siga Q62d)
d. Tomar medicamentos Q06204 — 1. Sim 2. Não (siga Q62e)
e. Não fumar Q06205 — 1. Sim 2. Não (siga Q62f)
f. Fazer acompanhamento regular Q06206 — 1. Sim 2. Não

Q63. Algum médico já lhe deu o diagnóstico de uma doença do coração, tais como infarto, angina, insuficiência cardíaca ou outra?(Leia as opções de resposta) Q063

1. Sim 2. Não

(Se Q63= 2, passe ao Q68. Caso contrário, siga para os itens abaixo.)

a. Infarto Q06301 — 1. Sim 2. Não (siga Q63b)
b. Angina Q06302 — 1. Sim 2. Não (siga Q63c)
c. Insuficiência cardíaca Q06303 — 1. Sim 2. Não (siga Q63d)
d. Outra (Especifique: Q06304 Q06305) — 1. Sim 2. Não

IBGE
Instituto Brasileiro de Geografia e Estatística

OBSERVAÇÕES

Módulo W - Informações Laboratoriais

W00101 - Informe o seu peso (primeira pesagem em kg)

W00102 - Informe o seu peso (segunda pesagem em kg)

W00201 - Informe a sua altura (primeira medida em cm)

W00202 - Informe a sua altura (segunda medida em cm)

W00301 - Informe a sua circunferência da cintura (primeira medida em cm)

W00302 - Informe a sua circunferência da cintura (segunda medida em cm)

W00401 - Informe a pressão arterial sistólica (primeira medida)

W00402 - Informe a pressão arterial diastólica (primeira medida

W00403 - Informe a pressão arterial sistólica (segunda medida)

W00404 - Informe a pressão arterial diastólica (segunda medida)

W00405 - Informe a pressão arterial sistólica (terceira medida)

W0i0406 - Informe a pressão arterial diastólica (terceira medida)

Muito obrigado pela sua participação! As informações que o(a) sr(a) nos forneceu serão valiosas para a formulação de políticas para a melhoria da assistência á saúde no Brasil.

(Encerre a entrevista)

Exemplo Prático e Desafio

Uso do ambiente Knime

Parte 3

Fusão de variáveis sobre Atividade Física

C006	P012	P021	P034	P03903	P04401	P05402	P05408	P05416	Q061	W00203
C00702	P013	P022	P035	P040	P04403	P05401	P05409	P05417	Q06201	W00303
C00703	P015	P025	P03701	P04101	P04404	P05402	P05410	P05418	Q06202	W00203
P007	P016	P026	P03702	P04102	P050	P05403	P05411	P05419	Q06203	V001
P008	P017	P02601	P038	P042	P051	P05404	P05412	P05421	Q06204	Act-Fisic
P009	P018	P027	P039	P04301	P052	P05405	P05413	P05422	Q06205	
P010	P019	P028	P03901	P04302	P053	P05406	P05414	P058	Q06206	
P011	P020	P029	P03902	P044	P05401	P05407	P05415	Q060	W00103	

P35. Quantos dias por semana o(a) sr(a) costuma praticar exercício físico ou esporte?
P38. No seu trabalho, o(a) sr(a) anda bastante a pé?
P39. No seu trabalho, o(a) sr(a) faz faxina pesada, carrega peso ou faz outra atividade pesada que requer esforço físico intenso?
P42. Nas suas atividades habituais (tais como ir a algum curso, escola ou clube ou levar alguém a algum curso, escola ou clube), quantos dias por semana o(a) sr(a) faz alguma atividade que envolva deslocamento a pé ou bicicleta?
P44. Nas suas atividades domésticas, o(a) sr(a) faz faxina pesada, carrega peso ou faz outra atividade pesada que requer esforço físico intenso?

$P035$ > 2 OR $P038$ = "1" OR $P039$ = "1" OR $P042$ > 2 OR $P044$ = "1" => Act-Fis="S“

NOT ($P035$ > 2 OR $P038$ = "1" OR $P039$ = "1" OR $P042$ > 2 OR $P044$ = "1") => Act-Fis="N"

Transformação de Variáveis – Cálculo do IMC

C006	P012	P021	P034	P03903	P04401	P05402	P05408	P05416	Q061	W00203
C00702	P013	P022		P040	P04403	P05401	P05409	P05417	Q06201	W00303
C00703	P015	P025	P03701	P04101	P04404	P05402	P05410	P05418	Q06202	V001
P007	P016	P026	P03702	P04102	P050	P05403	P05411	P05419	Q06203	**Act-Fisic**
P008	P017	P02601			P051	P05404	P05412	P05421	Q06204	**IMC**
P009	P018	P027		P04301	P052	P05405	P05413	P05422	Q06205	
P010	P019	P028	P03901	P04302	P053	P05406	P05414	P058	Q06206	
P011	P020	P029	P03902		P05401	P05407	P05415	Q060	W00103	

W00101 - Informe o seu peso (primeira pesagem em kg)
W00102 - Informe o seu peso (segunda pesagem em kg)
W00103 – Média do peso
W00201 - Informe a sua altura (primeira medida em cm)
W00202 - Informe a sua altura (segunda medida em cm)
W00202 – Média da altura

IMC = 10000*($W00103$/($W00203$*$W00203$))

Fusão de variáveis relacionados à freq. de alimentar-se com comida saudável

C006	P012	P021	P034	P03903	P04401	P05402	P05408	P05416	Q061	
C00702	P013	P022		P040	P04403	P05401	P05409	P05417	Q06201	
C00703	P015	P025	P03701	P04101	P04404	P05402	P05410	P05418	Q06202	V001
P007	P016	P026	P03702	P04102	P050	P05403	P05411	P05419	Q06203	**Act-Fisic**
P008	P017	P02601			P051	P05404	P05412	P05421	Q06204	**IMC**
P009	P018	P027		P04301	P052	P05405	P05413	P05422	Q06205	Comid-Saud
P010	P019	P028	P03901	P04302	P053	P05406	P05414	P058	Q06206	
P011	P020	P029	P03902		P05401	P05407	P05415	Q060		

P7. Em quantos dias da semana, o(a) sr(a) costuma comer salada de alface e tomate ou salada de qualquer outra verdura ou legume cru?)
P8. Em geral, quantas vezes por dia o(a) sr(a) come este tipo de salada?
P9. Em quantos dias da semana, o(a) sr(a) costuma comer verdura ou legume cozido, como couve, cenoura, chuchu, berinjela, abobrinha? (*sem contar batata, mandioca ou inhame*)
P10. Em geral, quantas vezes por dia o(a) sr(a) come verdura ou legume cozido?

Comid_Saud = max_in_args($P007$,$P010$)
$Comid_Saud$ = "7" => "Todos os dias"
NOT ($Comid_Saud$ = "7") => "Algumas vezes"

Eliminação de variáveis com dados aussentes significativos e/ou irrelevantes após análise de conhecimento com especialista de domínio

C006	P012	P021	P034	P03903	P04401	P05402	P05408	P05416	Q061	
C00702	P013	P022		P040	P04403	P05401	P05409	P05417	Q06201	
C00703	P015	P025	P03701	P04101	P04404	P05402	P05410	P05418	Q06202	V001
	P016	P026	P03702	P04102	P050	P05403	P05411	P05419	Q06203	Act-Fisic
	P017	P02601			P051	P05404	P05412	P05421	Q06204	IMC
	P018	P027		P04301	P052	P05405	P05413	P05422	Q06205	Comid-Saud
	P019	P028	P03901	P04302	P053	P05406	P05414	P058	Q06206	
P011	P020	P029	P03902		P05401	P05407	P05415	Q060		

Potenciais atributos (removidos) que poderiam aumentar a representatividade

PUC Minas

Categorização da variável Idade

C006			P034						Q061	
	P013			P040						
	P015	P025	P03701							
	P016	P026	P03702		P050					**Act-Fisic**
					P051					**IMC**
	P018	P027								Comid-Saud
		P028	P03901							
P011	P020							Q060		

$Idade$ <= 25 => "Jovem"
$Idade$ >= 26 AND $Idade$ <= 50 => "Adulto"
$Idade$ >= 51 AND $Idade$ <= 80 => "Adulto Maior"
$Idade$ >= 81 => "Idoso"

PUC Minas

Renomeação das variáveis

Sexo			P034						Q061	
	P013 (Dias-Carne-Ave)			P040						
	P015 (Dias-Peixe)	P025 (Dias-Alim-Doce)	P03701 (Hr-Exec-Fis)							
	P016 (Dias-Suco-Natural)	P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**
					P051 (Fumou no passado)					**IMC**
	P018 (Dias-Come-Frutas)	P027 (Consumo-Alcoo)								Comid-Saud
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade
P011 (Dias-Carne-Vermelha	P020 (Dias-BB-Refri)							Q060 (Diag-Col-Alto)		

PUC Minas

Analisa como a frequência de comer carne de boi e de ave se relacionam com o Diagnóstico

Sexo			P034						Q061	
	P013 (Dias-Carne-Ave)			P040						
	P015 (Dias-Carne-Peixe)	P025 (Dias-Alim-Doce)	P03701 (Hr-Exec-Fis)							
	P016 (Dias-Suco-Natural)	P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**
					P051 (Fumou no passado)					**IMC**
	P018 (Dias-Come-Frutas)	P027 (Consumo-Alcoo)								Comid-Saud
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade
P011 (Dias-Carne-Vermelha	P020 (Dias-BB-Refri)							Q060 (Diag-Col-Alto)		Dias-Carne-Av-BBo

Dias-Carne-Av-Bo = $Dias-Carne-Ave$ + $Dias-Carne-Vermelha$

$Dias-Carne-Av-Bo$ <= 5 => "Até5dias"
NOT ($Dias-Carne-Av-Bo$ <= 5) => "Mais5dias"

Analisa como a frequência de comer peixe se relaciona com o Diagnóstico

Sexo			P034						Q061	
				P040						
	P015 (Dias-Come-Peixe)	P025 (Dias-Alim-Doce)	P03701 (Hr-Exec-Fis)							
	P016 (Dias-Suco-Natural)	P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**
					P051 (Fumou no passado)					**IMC**
	P018 (Dias-Come-Frutas)	P027 (Consumo-Alcoo)								Comid-Saud
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade
	P020 (Dias-BB-Refri)							Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo

Remover: Dias-Come-Peixe

Analisa como a frequência de consumo de sucos naturais e frutas se relacionam com o Diagnóstico

Sexo			P034						Q061	
				P040						
		P025 (Dias-Alim-Doce)	P03701 (Hr-Exec-Fis)							**Dias-Comid-Natur**
	P016 (Dias-Suco-Natural)	P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**
					P051 (Fumou no passado)					**IMC**
	P018 (Dias-Come-Frutas)	P027 (Consumo-Alcoo)								Comid-Saud
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade
	P020 (Dias-BB-Refri)							Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo

Dias-Comid-Natur = $Dias-Suco-Natural$ + $Dias-Come-Frutas$

$Dias-Comid-Natur$ <= 6 => "Até6dias"
NOT ($Dias-Comid-Natur$ <= 6) => "Mais6dias"

Analisa como a frequência de consumo de refrigerantes se relaciona com o Diagnóstico

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							**Dias-Comid-Natur**
		P025 (Dias-Alim-Doce)	P03701 (Hr-Exec-Fis)								
		P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**	
					P051 (Fumou no passado)					**IMC**	
		P027 (Consumo-Alcoo)								Comid-Saud	
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade	
	P020 (Dias-BB-Refri)							Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

$Dias-Bebe-Refri$ <= 3 => "Até3dias"
$Dias-Bebe-Refri$ >= 4 AND $Dias-Bebe-Refri$ <=6 => "De4a6dias"
$Dias-Bebe-Refri$ >= 7 => "TodosDias"

Analisa como a frequência de consumo de doces relaciona com o Diagnóstico

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
		P025 (Dias-Alim-Doce)	P03701 (Hr-Exec-Fis)								**Dias-Comid-Natur**
		P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**	
					P051 (Fumou no passado)					**IMC**	
		P027 (Consumo-Alcoo)								Comid-Saud	
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

$Dias-Alim-Doces$ <= 2 => "Até2dias"
$Dias-Alim-Doces$ >= 3 AND $Dias-Alim-Doces$ <=5 => "De3a5dias"
$Dias-Alim-Doces$ >= 6 => "TodosDias"

Analisa como a frequência de troca de comida saudável por industrializada se relaciona com o Diagnóstico

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
			P03701 (Hr-Exec-Fis)								Dias-Comid-Natur
		P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**	Dias-Comid-SauxIndus
					P051 (Fumou no passado)					**IMC**	
		P027 (Consumo-Alcoo)								Comid-Saud	
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

$Dias-Comid-SauxIndus$ <= 1 => "Até1dias"
$Dias-Comid-SauxIndus$ >= 2 AND $Dias-Comid-SauxIndus$ <=6 => "De2a6dias"
$Dias-Comid-SauxIndus$ >= 6 => "TodosDias"

Analisa como se o hábito de consumo de bebida alcoólica se relacionam com o Diagnóstico

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
			P03701 (Hr-Exec-Fis)								Cons-Bebida-Alc
		P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**	**Dias-Comid-Natur**
					P051 (Fumou no passado)					**IMC**	
		P027 (Consumo-Alcoo)								Comid-Saud	
		P028 (Dias-cons-BB-Alcoo)	P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

$Cons-Bebida-Alc$ = 1 => "Nunca"
$Cons-Bebida-Alc$ = 2 => "UmaMes"
$Cons-Bebida-Alc$ = 3 => "MaisMes"

Analisa como o hábito de fumar se relacionam com o Diagnóstico

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
											Cons-Bebida-Alc
		P026 (Dias-Comid-SaudxIndust)			P050 (Fuma-Atualmente)					**Act-Fisic**	**Dias-Comid-Natur**
					P051 (Fumou no passado)					**IMC**	
										Comid-Saud	
			P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

$Fuma-Atualmen$ = "1" => "Diariamente"
$Fuma-Atualmen$ = "2" => "MenosQ-Diariamente"
$Fuma-Atualmen$ = "3" => "Não-Fuma"

Remove informação sobre exercícios físicos e de comida saudável

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
			P03701 (Hr-Exec-Fis)								**Cons-Bebida-Alc**
		P026 (Dias-Comid-SaudxIndust)	P03702 (Min-Exec-Fis)		P050 (Fuma-Atualmente)					**Act-Fisic**	**Dias-Comid-Natur**
										IMC	
										Comid-Saud	
			P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

Remover: Hr-Exec-Fis & Min-Exec-Fis

Remover: Comid-Saud

PUC Minas

Categoriza IMC

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
											Cons-Bebida-Alc
		P026 (Dias-Comid-SaudxIndust)			P050 (Fuma-Atualmente)					**Act-Fisic**	**Dias-Comid-Natur**
										IMC	
			P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

IMC < 18.5 => "Magreza"
IMC >= 18.5 AND IMC <=24.9 => "Saudável"
IMC >= 25.0 AND IMC <=29.9 => "Sobre-Peso"
IMC >= 30.0 AND IMC <=34.9 => "Obesidade 1"
IMC >= 35.0 AND IMC <=39.9 => "Obesidade 2"
IMC > 40 => "Obesidade 3"

Atributos removidos por estarem já representados

Sexo			P034						Q061		Dias-Bebe-Refri-Faixa
				P040							Dias-Alim-Doces
											Cons-Bebida-Alc
		P026 (Dias-Comid-SaudxIndust)			P050 (Fuma-Atualmente)					**Act-Fisic**	**Dias-Comid-Natur**
										IMC	
			P03901 (Dias-Ativ-Fis-Trab)							Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

P034: P34. Nos últimos três meses, o(a) sr(a) praticou algum tipo de exercício físico ou esporte? (*não considere fisioterapia*)

P03901:P39a. Em uma semana normal, em quantos dias o(a) sr(a) faz essas atividades no seu trabalho?

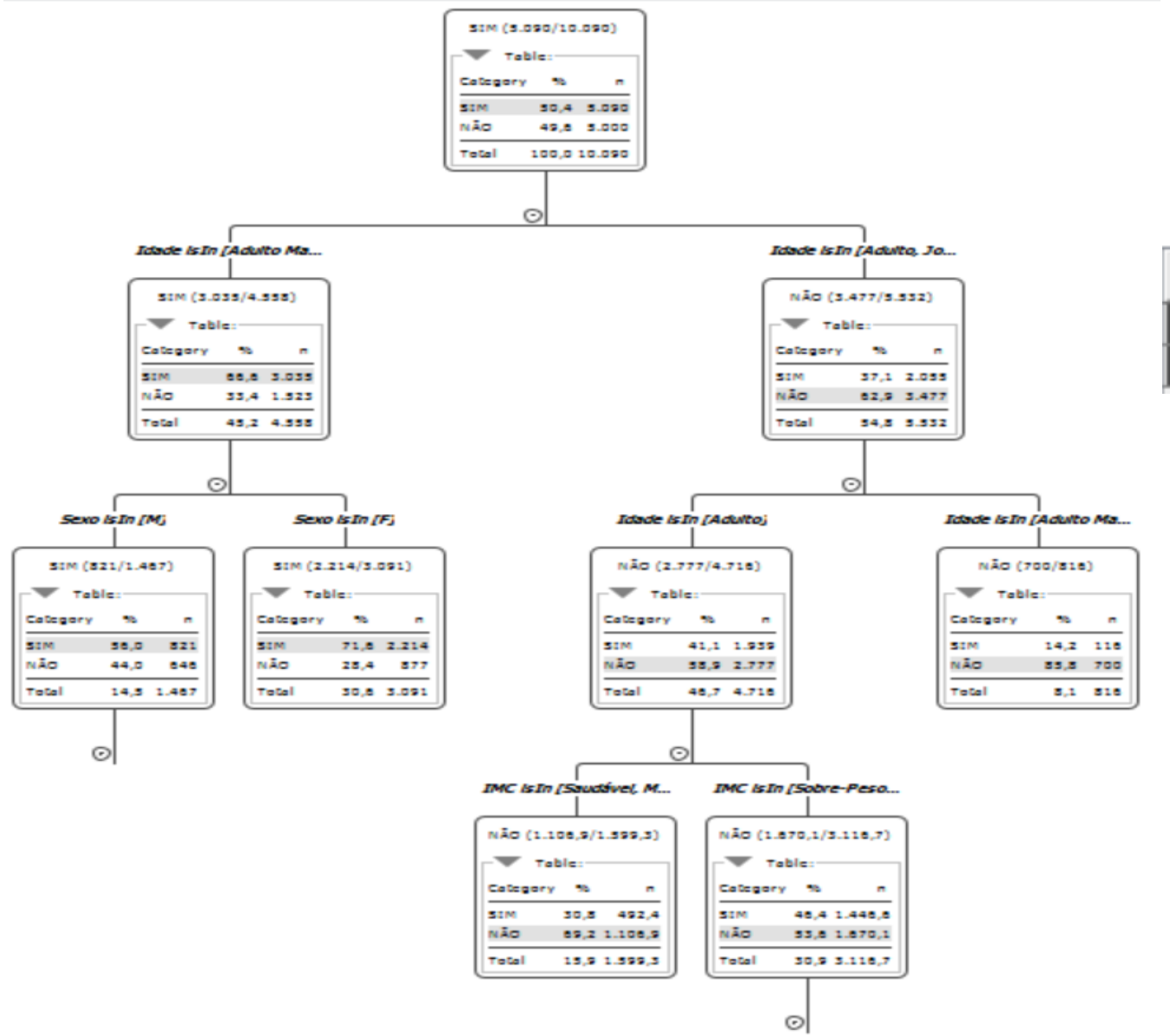
P040:P40. Para ir ou voltar do trabalho, o(a) sr(a) faz algum trajeto a pé ou de bicicleta?

Q061: Q61. Que idade o(a) sr(a) tinha no primeiro diagnóstico de colesterol alto?

Conjunto de Dados Resultantes

Sexo											Dias-Bebe-Refri-Faixa
											Dias-Alim-Doces
											Cons-Bebida-Alc
		P026 (Dias-Comid-SaudxIndust)			P050 (Fuma-Atualmente)					**Act-Fisic**	**Dias-Comid-Natur**
										IMC	
										Idade	
								Q060 (Diag-Col-Alto)		Dias-Carne-Av-Bo	

PUC Minas



Row ID	SIM	NÃO
SIM	880	378
NÃO	506	759

O desafio continua!!

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...
SIM	880	506	759	378	0.7	0.635	0.7	0.6	0.666
NÃO	759	378	880	506	0.6	0.668	0.6	0.7	0.632