UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE ENGENHARIAS, CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
ENGENHARIA DE COMPUTAÇÃO

Gabriel Pinto Andrade
Leonardo Mulinari

Relatório Técnico - Mineração de Dados

PONTA GROSSA
OUTUBRO DE 2024

Gabriel Pinto Andrade
Leonardo Mulinari

Relatório Técnico - Mineração de Dados

Trabalho apresentado como requisito parcial para obtenção de nota na disciplina de Mineração de Dados em Engenharia de Computação na Universidade Estadual de Ponta Grossa

Orientador: Profª Alaine Margarete Guimarães

PONTA GROSSA
OUTUBRO DE 2024

SUMÁRIO

1 DESCRIÇÃO DA BASE DE DADOS

A base de dados analisada compreende notificações de síndrome gripal registradas no estado de Minas Gerais entre 2020 e 2024, oferecendo uma visão detalhada das ocorrências ao longo desse período. Originalmente composta por 64 atributos, para esta análise foram selecionados seis atributos principais: idade, sexo, raça/cor, sintomas, outros sintomas e status de vacinação para COVID-19. Esses dados foram coletados a partir do sistema de saúde e-SUS Notifica e refletem informações clínico-sintomáticas e de vacinação dos pacientes, possibilitando uma análise aprofundada da distribuição dos sintomas e possíveis relações entre os atributos dos casos notificados.

1.1 Descrição dos Atributos

Atributo	**Descrição**	**Tipo**	**Domínio**
Idade	Idade do paciente, calculada automaticamente a partir da data de nascimento.	Numérico Contínuo	0 a 120 anos (estimado)
Sexo	Sexo do paciente	Categórico	Masculino, Feminino
RacaCor	Raça/Cor do paciente	Categórico	Branca, Preta, Parda, Amarela, Indígena, Ignorado
Sintomas	Sintomas apresentados pelo paciente	Categórico Múltiplo	Febre, Tosse, Coriza, etc.
OutrosSintomas	Descrição de sintomas adicionais	Texto Livre	Descrição textual
CodigoRecebeuVacina	Status de vacinação do paciente contra COVID-19	Numérico Discreto	1 (Sim), 2 (Não), 3 (Ignorado)

Tabela 1 – Descrição dos Atributos da Base de Dados

1.2 Estatísticas Descritivas dos Dados

Essas informações descrevem a distribuição e a frequência de categorias dentro dos atributos principais, fornecendo uma visão geral dos dados e permitindo identificar eventuais inconsistências.

1.2.1 Idade

- **Média**: 28,99 anos
- **Desvio Padrão**: 14,77 anos
- **Mediana**: 30 anos

- **Intervalo Interquartil** (25% - 75%): 19 - 41 anos
- **Idade Mínima e Máxima**: 0 e 54 anos, respectivamente

1.2.2 Sexo

Categoria	Frequência
Feminino	262.244 (aproximadamente 55%)
Masculino	170.097 (aproximadamente 36%)
Inconsistências	2 registros com valores inválidos

Tabela 2 – Distribuição do Atributo Sexo

1.2.3 Raça/Cor

Categoria	Frequência
Branca	159.970
Parda	138.030
Ignorado	85.252
Amarela	25.105
Preta	22.135
Indígena	1.801

Tabela 3 – Distribuição do Atributo Raça/Cor

1.2.4 Código Recebeu Vacina (Status de Vacinação para COVID-19)

Categoria	Frequência
Sim (1)	269.117
Não (2)	35.012
Ignorado (3)	257
Valores Incorretos	“1.0"(27.155 registros) e “2.0"(3.309 registros)

Tabela 4 – Distribuição do Atributo Código Recebeu Vacina

1.2.5 Sintomas

Sintoma	Frequência
Tosse	248.293
Coriza	193.873
Dor de Cabeça	179.171
Febre	164.487
Dor de Garganta	146.223
Outros	126.044
Dispneia	46.415
Assintomático	41.984
Distúrbios Gustativos	22.096
Distúrbios Olfativos	19.309

Tabela 5 – Frequência dos Sintomas Notificados

2 PRÉ-PROCESSAMENTO DOS DADOS

Para garantir a qualidade dos dados e melhorar a capacidade de análise, várias etapas de pré-processamento foram realizadas, conforme detalhado abaixo:

- **Discretização da coluna *Sintomas***: Os sintomas listados na coluna foram discretizados em colunas binárias, permitindo a análise de cada sintoma individualmente.
- **Correlação entre *Sintomas* e *OutrosSintomas***: Os sintomas presentes na coluna *Sintomas* foram buscados na coluna *OutrosSintomas*. Se um sintoma estivesse em uma das colunas, era marcado como verdadeiro na coluna do sintoma.
- **Discretização de *OutrosSintomas***: Os sintomas mais frequentes na coluna *OutrosSintomas* foram identificados e também discretizados em colunas separadas, para uma análise mais precisa. Os seguintes sintomas foram separados em outras colunas: Mialgia, Congestão Nasal, Vômito, Náusea, Diarréia.
- **Ajuste da coluna *CodigoRecebeuVacina***: Os valores ausentes foram preenchidos com o número 3 (Ignorado). Registros com o nome de uma vacina foram marcados como 1 (Recebeu vacina), e entradas contendo "não realizou teste"foram atribuídas como 2 (Não recebeu vacina).
- **Limpeza da coluna *Sexo***: Remoção de instâncias com valores vazios ou inválidos (como o valor "MG").
- **Preenchimento de dados faltantes na coluna *Idade***: A interpolação linear foi aplicada para preencher os valores ausentes.
- **Limpeza da coluna *Sintomas***: Remoção da única instância com valor vazio.

- **Ajuste da coluna *RacaCor***: Os valores ausentes foram preenchidos com a categoria "Ignorado".
- **Discretização das colunas *RacaCor*, *Sexo* e *CodigoRecebeuVacina***: Cada categoria dentro dessas colunas foi transformada em uma coluna binária (0 ou 1) separada.
- **Discretização da coluna *Idade* em faixas etárias**: A coluna *Idade* foi dividida nas faixas etárias 'Infantil (0-9)', 'Adolescente (10-19)', 'Jovem (20-39)', 'Meia-idade (40-59)', e 'Idoso (60+)', representadas por colunas binárias.

3 DESCRIÇÃO DA BASE DE DADOS PÓS-PROCESSADA

O conjunto de dados analisado é composto por informações sobre sintomas, dados demográficos e status de vacinação, contendo 31 atributos discretizados em categorias binárias para facilitar a análise dos padrões clínicos. Originalmente, o arquivo incluía dados não estruturados que foram organizados em atributos discretos e específicos, permitindo análises mais detalhadas.

3.1 Estatísticas Descritivas dos Dados

Os dados foram categorizados em binários, com contagem de ocorrências e frequência de sintomas e categorias demográficas.

- Sintomas: A maioria dos sintomas foi representada em variáveis binárias (presença ou ausência), sendo que sintomas como "Tosse", "Febre"e "Coriza"estão entre os mais frequentes. Cerca de 390.290 registros indicam ausência de sintomas.
- Raça/Cor: As categorias incluem "Branca", "Parda", "Amarela", "Preta", "Indígena"e "Ignorado", sendo a mais frequente "Branca".
- Sexo: Dos registros, 262.243 são de pacientes do sexo feminino e 170.097 do sexo masculino.
- Status de Vacinação: Três categorias indicam se o paciente foi vacinado contra a COVID-19, com maior frequência de pacientes já vacinados.
- Faixa Etária: As categorias etárias foram divididas em "Infantil", "Adolescente", "Jovem", "Meia-idade"e "Idoso", sendo a faixa etária jovem a mais frequente.

3.2 Descrição dos Atributos

Atributo	Descrição	Tipo	Domínio
Dor de Cabeça	Indica se o paciente relatou dor de cabeça	Categórico Binário	False, True
Tosse	Indica se o paciente apresentou tosse	Categórico Binário	False, True
Febre	Indica se o paciente apresentou febre	Categórico Binário	False, True
Dor de Garganta	Presença de dor de garganta	Categórico Binário	False, True
Coriza	Presença de coriza	Categórico Binário	False, True
Dispneia	Indica se houve relato de dispneia	Categórico Binário	False, True
Assintomático	Indica ausência de sintomas	Categórico Binário	False, True
Distúrbios Olfativos	Indica alteração olfativa	Categórico Binário	False, True
Distúrbios Gustativos	Indica alteração gustativa	Categórico Binário	False, True
Mialgia	Presença de mialgia	Categórico Binário	False, True
Congestão Nasal	Presença de congestão nasal	Categórico Binário	False, True
Vômito	Indica se houve relato de vômito	Categórico Binário	False, True
Náusea	Indica se houve relato de náusea	Categórico Binário	False, True
Diarréia	Indica se houve relato de diarreia	Categórico Binário	False, True
racaCor_Amarela	Paciente de raça/cor amarela	Categórico Binário	False, True
racaCor_Branca	Paciente de raça/cor branca	Categórico Binário	False, True
racaCor_Ignorado	Informação de raça/cor ignorada	Categórico Binário	False, True
racaCor_Indígena	Paciente de raça/cor indígena	Categórico Binário	False, True
racaCor_Parda	Paciente de raça/cor parda	Categórico Binário	False, True
racaCor_Preta	Paciente de raça/cor preta	Categórico Binário	False, True
sexo_Feminino	Sexo feminino	Categórico Binário	False, True
sexo_Masculino	Sexo masculino	Categórico Binário	False, True
codigoRecebeuVacina_1	Recebeu vacina contra COVID-19	Categórico Binário	False, True
codigoRecebeuVacina_2	Não recebeu vacina contra COVID-19	Categórico Binário	False, True
codigoRecebeuVacina_3	Status de vacinação ignorado	Categórico Binário	False, True
faixa_etaria_infantil (0-9)	Faixa etária infantil (0 a 9 anos)	Categórico Binário	False, True
faixa_etaria_adolescente (10-19)	Faixa etária adolescente (10 a 19 anos)	Categórico Binário	False, True
faixa_etaria_jovem (20-39)	Faixa etária jovem (20 a 39 anos)	Categórico Binário	False, True
faixa_etaria_meia-idade (40-59)	Faixa etária de meia-idade (40 a 59 anos)	Categórico Binário	False, True
faixa_etaria_idoso (60+)	Faixa etária idosa (60+ anos)	Categórico Binário	False, True

Tabela 6 – Descrição dos Atributos da Base de Dados

4 RESULTADOS E REGRAS DE ASSOCIAÇÃO

Tabela 7 – Regras de Associação e seus Valores de LIFT

Regras de Associação	**LIFT**
Dor de Cabeça → códigoRecebeuVacina_1	1.022
Febre → Tosse	1.241
Dor de Garganta → Tosse	1.238
Coriza → Tosse	1.316
raçaCor_Branca → códigoRecebeuVacina_1	1.073
faixa_etaria_jovem(20-39) → códigoRecebeuVacina_1	1.081
sexo_Feminino, Coriza → Tosse	1.312
Coriza, códigoRecebeuVacina_1 → Tosse	1.303
sexo_Feminino, faixa_etaria_jovem(20-39) → códigoRecebeuVacina_1	1.085

Com 0.8 de confiança e 0.3 de suporte minímos não foram encontradas regras, portanto esses valores foram ajustados para 0.7 e 0.2, respectivamente, de modo que as regras na tabela acima foram encontradas.

A relação entre a presença de Coriza e a Tosse apresenta um LIFT de 1.316, oque sugere que a coriza pode ser um preditor relevante para o aparecimento de tosse.

A análise também revela que sintomas como Dor de Cabeça, Dor de Garganta, e Coriza estão frequentemente associados à administração de vacinas (códigoRecebeuVacina_1), com LIFTs que variam de 1.022 a 1.316. Isso indica que a presença desses sintomas pode estar correlacionada com uma maior chance de o indivíduo ter recebido a vacina e ter sintomas como efeitos colaterais, mas não necessariamente que se o indivíduo toma vacina ele terá alguma efeito colateral

A análise das variáveis demográficas, como raça e faixa etária, mostra que tanto a raça branca quanto a faixa etária de 20 a 39 anos têm uma associação positiva com a vacinação, com LIFTs de 1.073 e 1.081, respectivamente. Isso sugere que esses grupos demográficos podem estar mais propensos a receber vacinas.

A combinação de sexo feminino e coriza também mostra um LIFT de 1.312, reforçando a ideia de que a presença de determinados sintomas pode ser mais pronunciada em mulheres.