A Aplicação do Processo de KDD aos Dados da COVID-19: Um Estudo de Caso no Rio Grande do Sul, Brasil

GABRIEL V. HEISLER, PROF. DR. JOAQUIM V. C. ASSUNÇÃO

CT
CENTRO DE TECNOLOGIA - UFSM

Universidade Federal de Santa Maria
SEDES SAPIENTIÆ
1960

informática
Universidade Federal de Santa Maria

Objetivos

- **Objetivo Geral:**
  - Aplicar o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases - KDD*) para identificar padrões nos dados da pandemia de COVID-19 no Rio Grande do Sul.
- **Objetivos Específicos:**
  - Coletar e preparar dados abrangentes sobre casos de COVID-19 no estado.
  - Utilizar técnicas de mineração de dados para identificar associações entre sintomas e desfechos dos pacientes.
  - Apresentar visualizações e análises dos padrões identificados.
  - Contribuir com informações valiosas para futuras pesquisas em saúde pública.

Dados

- No início do ano de 2020 a Organização Mundial da Saúde (OMS) oficializou a COVID-19 como uma pandemia global
- Com isso, entidades começaram a coletar e divulgar dados sobre a pandemia.
- O governo do estado do Rio Grande do Sul desenvolveu um painel da COVID-19, apresentando um dashboard interativo e disponibilizando dados para download.
- Os conjuntos de dados incluem informações detalhadas sobre cada caso confirmado de COVID-19 no estado.
- Cada linha representa um paciente e inclui dados sobre sintomas apresentados e a evolução do caso (recuperação ou óbito).

Dados

- As opções de dados para download são por ano ou os dados completos.
- Neste trabalho são usados os dados entre 2020 e 2023.
- O *dataset* utilizado conta com aproximadamente 3 milhões de linhas e pesa mais de 600MB

Metodologia

- Neste trabalho a metodologia do Processo de Descoberta de Conhecimento em Bases de Dados *(KDD)* proposta por Fayyad et. al. (1996) foi utilizada.
- Todas as etapas foram realizadas, porém somente as mais relevantes serão mostradas.
- A linguagem R foi utilizada.

Metodologia

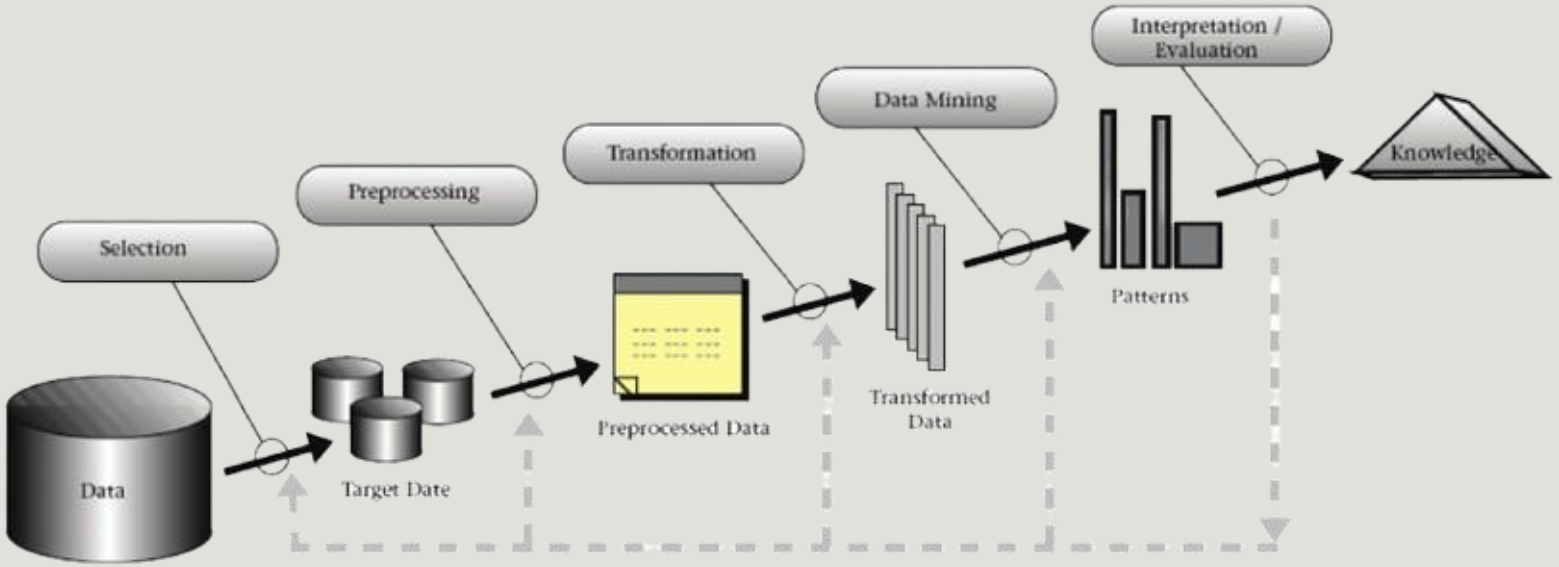


Figura do artigo “From data mining to knowledge discovery in databases.”, Fayyad et. Al (1996)

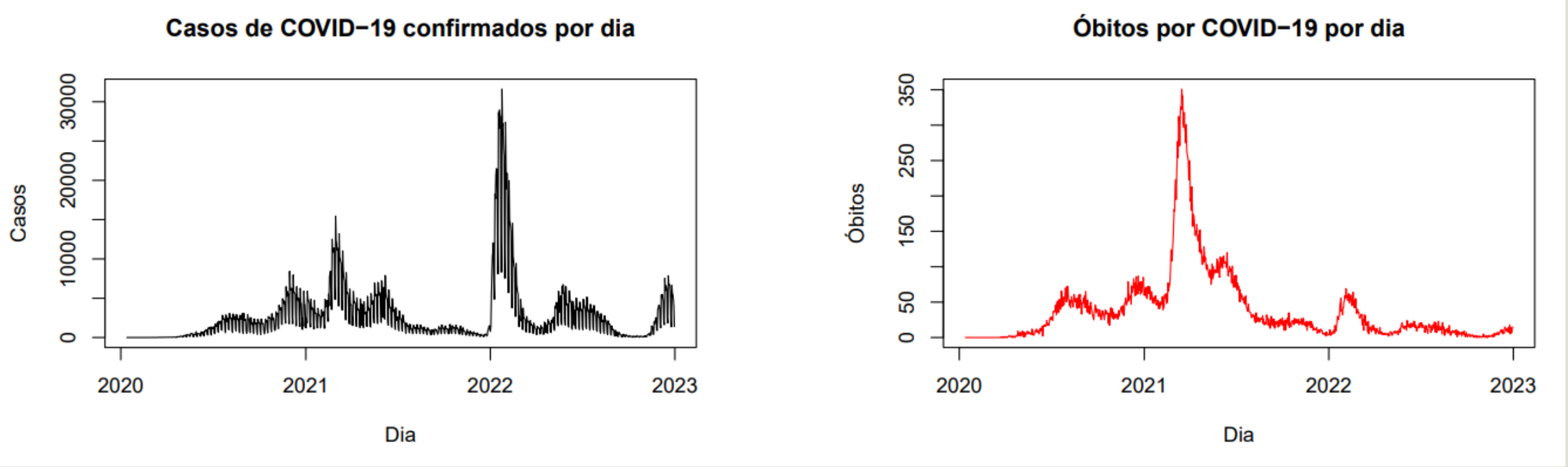
Extração e seleção dos dados

- Os dados foram extraídos (em csv) do painel fornecido pelo Estado do Rio Grande do Sul.
- Não existem dados relevantes faltantes do *dataset*.
- Os dados contam com as informações do caso específico de COVID-19 (como cidade, idade do paciente, sintomas, evolução do paciente, entre outros).
- Foram selecionados para este trabalho os dados de sintomas e evolução do paciente.

Visualização

- Primeiramente ter uma visão geral dos dados foram criados novos conjuntos de dados para criar algumas visualizações.
- As visualizações mais relevantes criadas nesta etapa são os gráficos de quantidade de casos e quantidade de óbitos diários.

Visualização



Pré-processamento

- Algumas etapas básicas de pré-processamento foram feitas, como transformação das datas para o formato correto.

Transformação

- Para posteriormente aplicar os algoritmos de mineração de dados é necessário que os dados sejam transformados, para “servirem” ao algoritmo.
- Como mais de um algoritmo de mineração de dados foi aplicado, as etapas de transformação serão expostas juntamente dos algoritmos.

Mineração de Dados

- Foram aplicados algoritmos de associação e de classificação de dados.
- Primeiramente serão mostradas as etapas referentes ao método de associação, e posteriormente o método de classificação.

Associação

- O algoritmo de associação escolhido para este trabalho foi o Apriori, visto que os dados disponíveis tem um formato adequado.
- Este algoritmo trabalha com dados preferencialmente binários (verdadeiro ou falso). Os dados selecionados são basicamente binários (são definidos como “SIM” e “NÃO”, no caso dos sintomas, e “ÓBITO” e “RECUPERADO”, no caso da evolução do paciente).
- O algoritmo funciona com base em podas baseadas em suporte. Os conjuntos são verificados, e se não tem um suporte mínimo são descartados.

Associação

- Os dados foram transformados para o formato necessário.

EVOLUCAO	FEBRE	TOSSE	GARGANTA	DISPNEIA	OUTROS
RECUPERADO	NAO	NAO	NAO	NAO	SIM
RECUPERADO	SIM	SIM	SIM	NAO	NAO
RECUPERADO	NAO	SIM	NAO	NAO	NAO
RECUPERADO	NAO	SIM	SIM	NAO	NAO
RECUPERADO	NAO	SIM	NAO	SIM	SIM
RECUPERADO	NAO	NAO	NAO	NAO	NAO
OBITO	SIM	SIM	NAO	SIM	NAO
RECUPERADO	SIM	NAO	NAO	SIM	NAO
RECUPERADO	SIM	SIM	NAO	SIM	SIM
RECUPERADO	SIM	SIM	NAO	NAO	SIM

Apriori

- O algoritmo Apriori, da biblioteca Arules da linguagem R, foi empregado nesta análise.
- Configuração inicial:
  - Suporte Mínimo: Inicialmente definido como 0.1.
  - Apenas regras com a evolução do paciente no lado direito (rhs) foram consideradas.
- Notou-se que nenhuma regra foi gerada onde o paciente tinha como evolução “Óbito”.
- Isto acontece devido a quantidade de casos nos quais a evolução é “Recuperado” ser muito maior que a de óbitos (>98.5% de recuperados).

Apriori

- Como o desbalanceamento nos tipos de evolução dos casos impactou na geração de regras associativas foram feitas duas abordagens:

1. Ignorar todos os casos onde a evolução do paciente foi “Recuperado”.
2. “Balancear” o *dataset.*

- Ambas alternativas foram utilizadas.

Apriori

lhs	rhs	support
GARGANTA=NÃO	EVOLUCAO=OBITO	0.858
DISPNEIA=SIM	EVOLUCAO=OBITO	0.831
OUTRO=NÃO	EVOLUCAO=OBITO	0.725
FEBRE=SIM	EVOLUCAO=OBITO	0.590
TOSSE=NÃO	EVOLUCAO=OBITO	0.520

Regras geradas omitindo os casos onde a evolução é “Recuperado”
(Suporte mínimo foi configurado como 0.5)

Apriori: *dataset* balanceado

- O “Balanceamento” do conjunto de dados foi realizado juntando a totalidade dos dados nos quais a evolução é óbito com uma amostra do mesmo tamanho de dados de evolução recuperado.

Suporte	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Nº de regras	0	0	0	0	0	0	1	3	7	27

Número de regras geradas com diferentes suportes
(e confiança 0.8)

Apriori: *dataset* balanceado

lhs	rhs	suporte	confiança
GARGANTA=NÃO, DISPNEIA=SIM, OUTRO=NÃO	ÓBITO	0.265	0.944
GARGANTA=NÃO, DISPNEIA=SIM	ÓBITO	0.355	0.935
DISPNEIA=SIM, OUTRO=NÃO	ÓBITO	0.309	0.916
TOSSE=NÃO, DISPNEIA=SIM	ÓBITO	0.215	0.907
DISPNEIA=SIM	ÓBITO	0.415	0.903
TOSSE=SIM, DISPNEIA=SIM	ÓBITO	0.200	0.898
FEBRE=SIM, DISPNEIA=SIM	ÓBITO	0.252	0.887

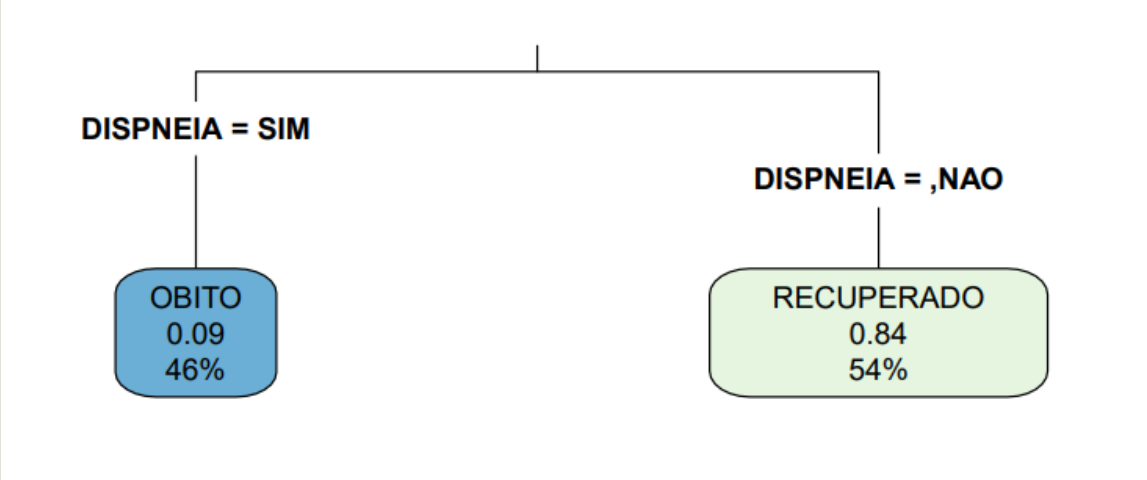
Regras geradas com o conjunto balanceado
(suporte de 0.2 e confiança de 0.8)

Classificação

- Uma árvore de classificação foi selecionada como o algoritmo de classificação principal para esta análise.
- Este algoritmo funciona construindo uma árvore de decisão de forma recursiva, dividindo os dados em subconjuntos cada vez mais homogêneos com base nos atributos
- A biblioteca RPART, disponível na linguagem R, foi utilizada para implementar e treinar a árvore de classificação.

Árvore de Classificação

- Para esta etapa o conjunto de dados também foi balanceado (da mesma maneira que feito na etapa de Associação)
- A árvore foi criada utilizando os sintomas como atributos e a evolução do paciente como classe alvo.

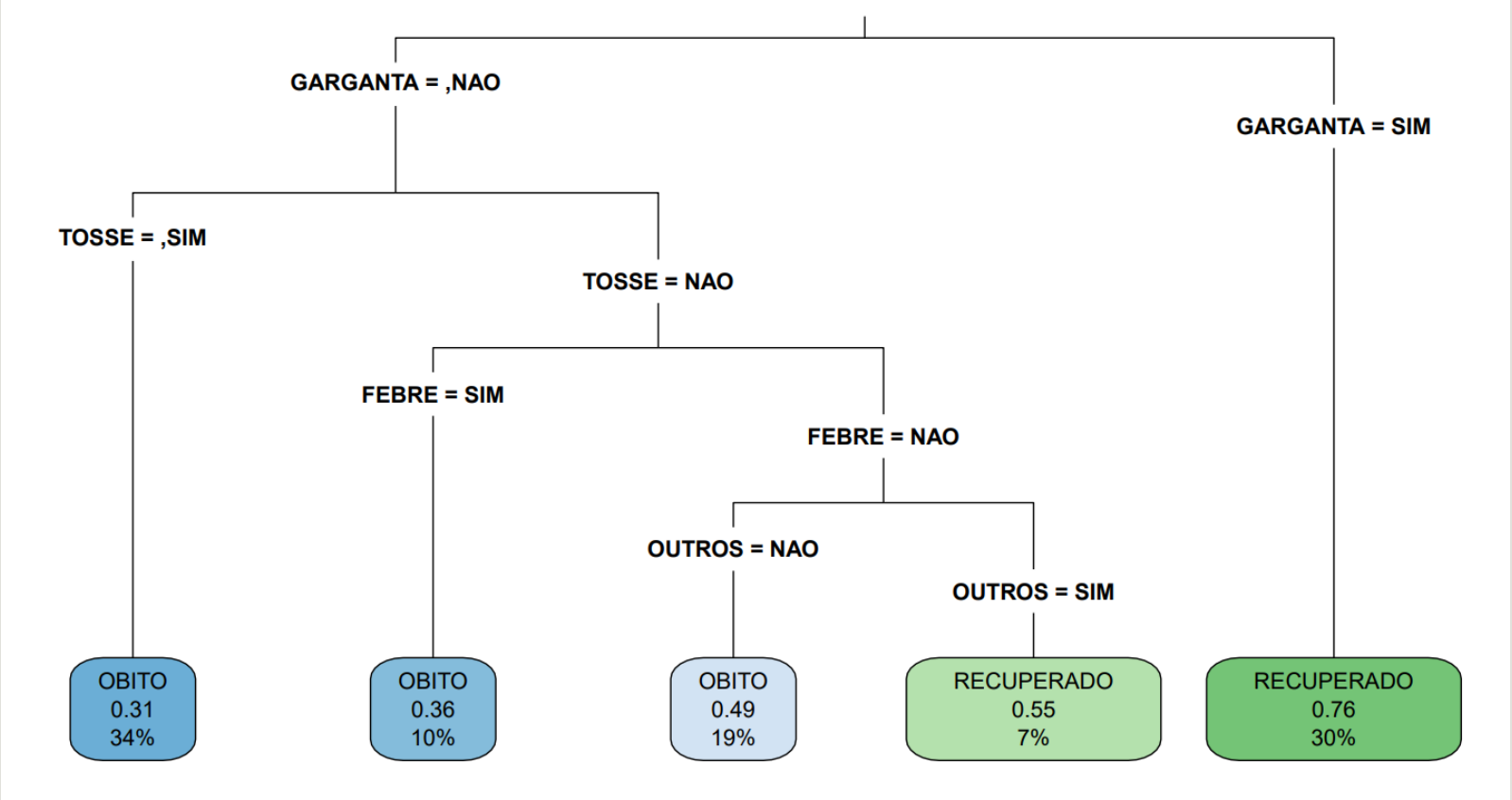


Primeira árvore gerada

Árvore de Classificação

- Como a dispneia é um sintoma muito determinante para a evolução do paciente a árvore foi gerada dependendo fortemente dele.
- Para ter uma visualização mais ampla, o sintoma foi omitido e uma nova árvore foi gerada.

Árvore de Classificação



Árvore gerada omitindo a coluna de dispneia

Resultados obtidos

- Os resultados obtidos com os métodos de mineração de dados mostraram padrões nos dados que não são triviais, como por exemplo que alguns sintomas prevalecem em comparação a outros quando se trata de pacientes que foram a óbito.
- Essas descobertas podem ser valiosas, por exemplo, em situações clínicas, auxiliando na identificação precoce e manejo de pacientes com suspeita de COVID-19.
- Ressaltamos que este trabalho não tem como objetivo ser um guia médico ou instrumento de diagnóstico. Também não buscamos fornecer respostas definitivas, mas sim explorar os dados disponíveis.

Trabalhos futuros

- Em trabalhos futuros buscamos explorar ainda mais os dados da doença no estado.
- Investigar a evolução dos padrões de sintomas ao longo do tempo.
- Combinar conjuntos de dados da COVID-19 com informações como ocupação de leitos hospitalares e dados de vacinação