Mineração de dados

Aplicação em ‘Clinical Big Data’

Grupo:
Pedro Leandro
Rychardson Ribeiro

O que é Mineração de dados?

A mineração de dados é uma ferramenta que pode ser utilizada descobrir padrões ou informações valiosas em grandes datasets. Isso é feito combinando técnicas de aprendizado de máquina, análise estatística.

Benefícios da Mineração de dados

- Descoberta de insights
- Descoberta tendências ocultas;
- Economia no orçamento e ganho de eficiência.

Desafios da Mineração de dados

- Complexidade;
- Custo;
- Incerteza.

Aplicação

Artigo da Military Medical Research (MMR)

Data mining in clinical big data: the frequently used databases, steps, and methodological models

Disponível em: https://mmrjournal.biomedcentral.com/articles/10.1186/s40779-021-00338-z

O foco do artigo

Analisar as técnicas de mineração de dados aplicadas a big data clínica, com foco em métodos que suportam análise preditiva e descritiva. Isso é feito:

1. Introduzindo conceitos e datasets públicos como SEER, MIMIC e NHANES.
2. Descrevendo modelos, tarefas e métodos de data mining em análises clínicas.
3. Demonstração de exemplos práticos de aplicação, por exemplo: modelos preditivos e análises de padrões de saúde para apoiar médicos e pesquisadores no uso de big data na tomada de decisões clínicas.

Dataset SEER (Surveillance, Epidemiology, and End Results)

Foco: Dados sobre câncer.

Informações disponíveis:

- Estatísticas de incidência e sobrevivência para vários tipos de câncer.
- Detalhes sobre localização do tumor, estágio da doença, tratamento e resultados.

Aplicações:

- Estudo de prognósticos em pacientes com câncer.
- Identificação de fatores de risco e tendências epidemiológicas.

Origem: Base de dados dos Estados Unidos, mantida pelo National Cancer Institute (NCI).

Dataset NHANES (National Health and Nutrition Examination Survey)

Foco: Saúde e nutrição de crianças e adultos.

Informações disponíveis:

- Dados sobre estado de saúde, histórico médico, hábitos alimentares e fatores ambientais.
- Exames físicos e laboratoriais abrangentes, incluindo análises sanguíneas e nutricionais.

Aplicações:

- Estudo de padrões nutricionais e suas associações com doenças.
- Identificação de fatores de risco populacionais, como obesidade e diabetes.

Origem: Mantido pelo CDC (Centers for Disease Control and Prevention) nos Estados Unidos.

Dataset MIMIC (Medical Information Mart for Intensive Care)

Foco: Dados de cuidados intensivos em hospitais.

Informações disponíveis:

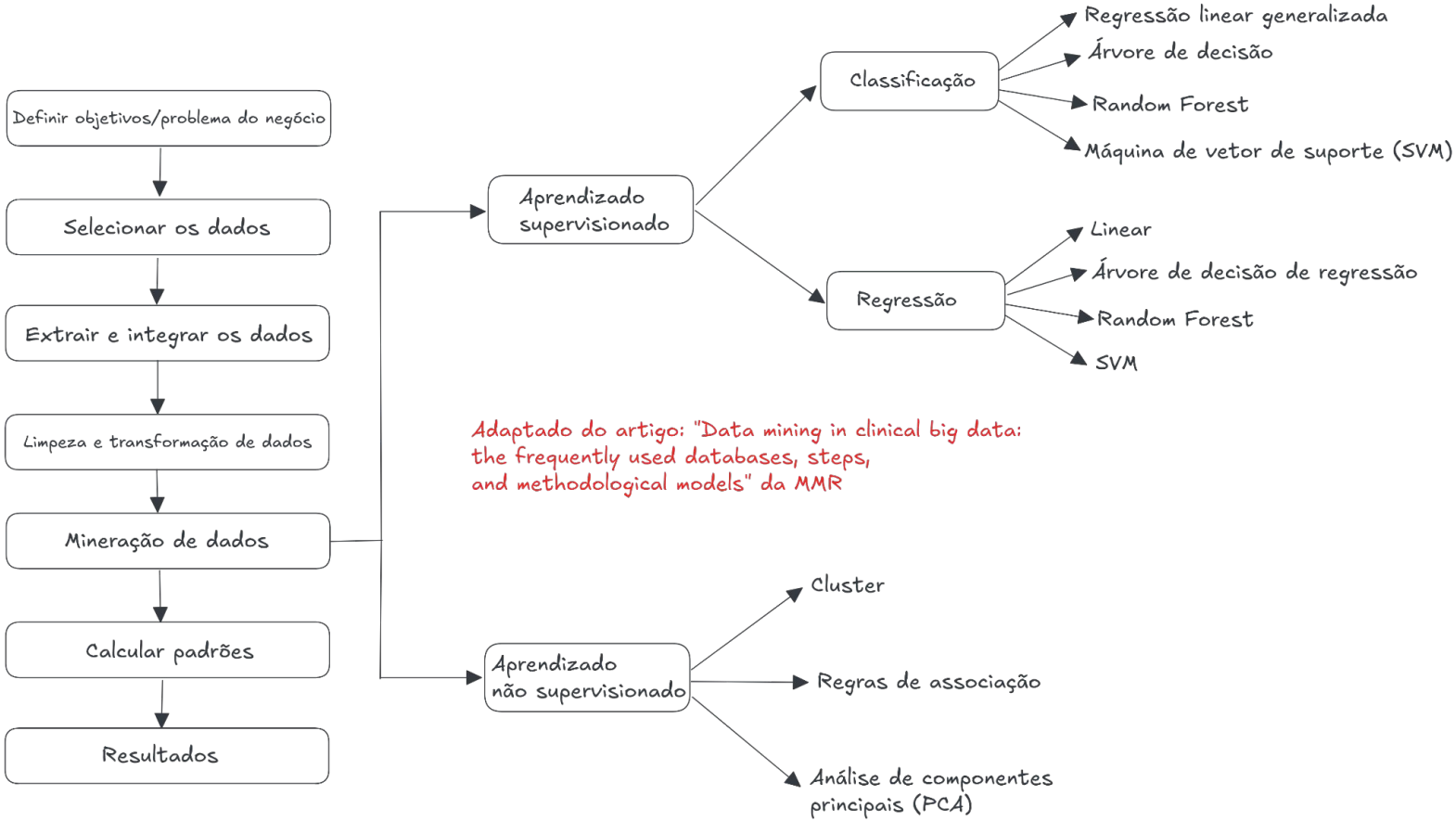
- Dados clínicos de pacientes de UTIs, incluindo sinais vitais, exames laboratoriais, medicamentos, intervenções médicas e desfechos hospitalares.
- Informações demográficas e temporais, como tempo de internação e evolução do quadro clínico.

Aplicações:

- Desenvolvimento de modelos preditivos para mortalidade e complicações em UTIs.
- Estudos sobre gerenciamento de tratamentos críticos.

Origem: Desenvolvido pelo MIT, contém dados de hospitais dos EUA.

Passo a passo de mineração de dados em datasets públicos de medicina



Técnicas utilizadas

1. **Aprendizado supervisionado**:
   - **Random Forest (RF)**: Usada para predição de mortalidade em UTIs com base em dados do banco MIMIC.
   - **Máquinas de Vetores de Suporte (SVM)**: Utilizadas para criar modelos preditivos para adesão a medicamentos e para predizer diabetes com base em variáveis de risco.
   - **Regressão logística e de Cox**: Aplicadas para análise de fatores prognósticos.
2. **Aprendizado não supervisionado**:
   - **Análise de Agrupamento (Clustering)**:
     - Exemplos incluem o agrupamento de pacientes hipertensos em subgrupos para melhor gerenciamento.
   - **Regras de Associação (Apriori e FP-Growth)**: Aplicadas para identificar fatores de risco de doenças e padrões de tratamento.
   - **Análise de Componentes Principais (PCA)**: Usada para redução de dimensionalidade e eliminação de variáveis ruidosas.
3. **Modelos de risco competitivo**: Analisam a coexistência de múltiplos desfechos em dados de sobrevivência, como no caso de pacientes com câncer.

Aplicações

1. **Previsão de Sepsis em UTIs**: Usando modelos baseados em RF e SVM no banco MIMIC para predizer prognósticos.
2. **Fatores prognósticos de câncer**: Utilizando dados do banco SEER para investigar riscos competitivos em pacientes com câncer.
3. **Padrões alimentares e nutrição**: PCA aplicada em dados do NHANES para avaliar a relação entre alimentos ultraprocessados e qualidade nutricional.