Relatório em Látex do DECAT à SEMAC

Franklin Vitor Soares Nascimento
Publicado: 28 de novembro, 2025

Resumo

Os Dados do mundo real raramente vêm prontos para análise, eles são desorganizados e inconsistentes. Portanto, fazendo assim a necessidade de arsenais de técnicas e métodos para à manipulação e tratamento dos dados. A principal medida para a confecção e tratamento desses dados é a utilização de uma ferramenta que seja capaz de lidar com as adversidades, neste presente caso, a Linguagem R juntamente ao ecossistema Tidyverse auxiliaram neste trabalho. Foi utilizado um conjunto de dados real à respeito do câncer de mama. O processo iniciou-se com a importação do conjunto de dados, seguido de uma etapa crucial de limpeza e organização. Foram aplicadas técnicas para lidar com valores ausentes, corrigir inconsistências nos registros e padronizar os formatos, garantindo a integridade e a qualidade das informações presentes.

Introdução

A análise de dados em saúde tem se tornado uma ferramenta essencial para apoiar decisões clínicas e aprimorar a detecção precoce de doenças. Entre os conjuntos de dados amplamente utilizados para esse fim, destaca-se o Breast Cancer Wisconsin (Diagnostic), que reúne informações numéricas extraídas de imagens de biópsias mamárias. Cada registro no banco descreve características morfológicas das células — como raio, textura, compactação e concavidade — permitindo identificar padrões associados a tumores benignos ou malignos.

Nesta aplicação, realizamos um tratamento completo desses dados, envolvendo limpeza, padronização, exploração estatística e preparação para modelos preditivos. O objetivo é transformar os atributos brutos em informações estruturadas e confiáveis, capazes de servir como base para análises mais profundas e algoritmos de aprendizado de máquina. Com isso, buscamos demonstrar como a engenharia e a análise de dados podem contribuir para soluções inteligentes voltadas ao auxílio do diagnóstico de câncer de mama.

O avanço das tecnologias de coleta e processamento de informações abriu novas possibilidades para a área da saúde, permitindo transformar grandes volumes de dados em conhecimento valioso. Dentro desse cenário, o conjunto de dados Breast Cancer Wisconsin (Diagnostic) surge como uma importante base para experimentação e estudo, pois reúne medidas quantitativas obtidas a partir de imagens microscópicas de tecido mamário. Durante o processamento, cada variável passou por etapas de normalização e verificação de inconsistências para garantir a integridade das informações utilizadas. Essas informações, extraídas de características como forma, textura e estrutura celular, possibilitam a construção de análises capazes de apoiar o diagnóstico do câncer de mama. Segue com algumas fórmulas estatísticas usadas nesse meio:

Média $= \bar{x}$

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned} \tag{1}$$

Variância amostral

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n-1}} \tag{2}$$

Referenciando automaticamente um número de equação usando seu rótulo: Equação 2.

A aplicação desenvolvida tem como propósito realizar o tratamento completo desse banco de dados, englobando etapas como limpeza, verificação de inconsistências, normalização e preparação dos atributos para uso em modelos analíticos. Ao estruturar esses dados de maneira adequada, torna-se viável explorar padrões relevantes e compreender variáveis que exercem influência no diagnóstico clínico. Assim, o projeto demonstra como métodos de ciência de dados podem ser aplicados de forma eficiente para transformar registros biomédicos em insights úteis e potencialmente aplicáveis ao apoio à decisão médica.

Nesta aplicação, utilizamos técnicas de tratamento e preparação de dados para organizar e tornar essas informações adequadas ao uso em modelos analíticos. Esse processo envolve desde a identificação e correção de possíveis ruídos até a transformação

das variáveis em formatos que facilitem interpretações e análises subsequentes. Dessa forma, buscamos demonstrar como práticas de engenharia de dados podem contribuir significativamente para estudos preditivos em saúde e para o desenvolvimento de soluções que apoiam o diagnóstico precoce do câncer de mama.

Metodologias

A aplicação desenvolvida tem como objetivo realizar o tratamento, organização e exploração do conjunto de dados **Breast Cancer Wisconsin (Diagnostic)**, amplamente utilizado em estudos de predição e identificação de câncer de mama. As metodologias adotadas concentram-se em técnicas de pré-processamento, limpeza, padronização e análise estatística, garantindo que os atributos referentes às características celulares sejam adequadamente preparados para utilização em modelos computacionais. Cada etapa foi estruturada de forma modular, permitindo maior flexibilidade na análise e maior robustez na extração de padrões significativos.

Amostras & Processamento

O conjunto de dados é composto por amostras derivadas de imagens digitalizadas de biópsias mamárias, a partir das quais foram extraídas propriedades como raio, textura, concavidade e simetria. Durante o processamento, cada variável passou por etapas de normalização e verificação de inconsistências para garantir a integridade das informações utilizadas. Esse procedimento assegura que os modelos analíticos sejam alimentados com dados consistentes e representativos.

Esta linha mostra como usar uma nota de rodapé para explicar ou citar um texto com mais detalhes. Exemplo:[1].

Esta é uma lista com marcadores:

- Assegurar a limpeza e remoção de ruídos presentes nas medidas originais
- Aplicar normalização estatística para homogenização dos atributos
- Identificar correlações relevantes entre variáveis clínicas e morfológicas
- Preparar o conjunto final para análises preditivas e métodos de classificação

Após essas etapas, o conjunto Breast Cancer Wisconsin (Diagnostic) passa por um refinamento que garante maior confiabilidade às análises subsequentes. Esse processo permite que as informações derivadas das imagens de biópsias sejam representadas de forma clara e consistente, reduzindo interferências causadas por medições irregulares ou variações indesejadas. Além disso, a padronização aplicada facilita a comparação entre amostras e destaca padrões relevantes que poderiam passar despercebidos em seu formato original. Dessa maneira, o banco de dados ganha estrutura para suportar investigações mais profundas, contribuindo tanto para estudos exploratórios quanto para a aplicação de algoritmos de classificação capazes de auxiliar na identificação precoce de tumores mamários. Esta é uma lista numerada:

1. Coleta e estruturação inicial das amostras do dataset
2. Pré-processamento e padronização das variáveis
3. Geração de métricas analíticas para avaliação dos resultados

Subseção 1

Nesta etapa, a aplicação se concentra na estruturação e no refinamento inicial dos dados, garantindo que o conjunto Breast Cancer Wisconsin (Diagnostic) esteja devidamente preparado para análises posteriores. O fluxo de trabalho envolve a inspeção detalhada das amostras, buscando identificar inconsistências, valores ausentes ou registros potencialmente ruidosos que possam interferir nas etapas seguintes do processamento. Além disso, são construídos gráficos exploratórios que permitem visualizar padrões preliminares, como distribuições das variáveis, relações entre atributos morfológicos e diferenças visuais entre amostras benignas e malignas.

Outro ponto essencial desta fase é a padronização das medidas numéricas, especialmente aquelas relacionadas ao raio celular, compactação e concavidade, cuja variabilidade natural entre as amostras pode comprometer a performance de modelos estatísticos e algoritmos de aprendizado de máquina. A aplicação utiliza técnicas de normalização que tornam as escalas homogêneas, facilitando interpretações e garantindo que cada característica contribua de forma equilibrada ao conjunto tratado. Esse processo fornece uma base sólida para os passos seguintes da análise.

Subseção 2

Nesta fase, o foco da aplicação volta-se para a validação estrutural do banco de dados e para a reorganização das variáveis com o objetivo de otimizar a construção de modelos preditivos. Vestibulum sodales orci a nisi interdum tristique. In dictum vehicula

[1] Este projeto utiliza dados reais, porém anonimizados, amplamente empregados em estudos acadêmicos.

dui, eget bibendum purus elementum eu. A análise de consistência é aplicada para verificar se todas as medidas extraídas das imagens microscópicas seguem padrões coerentes e se não há discrepâncias que possam distorcer o processo de classificação entre amostras benignas e malignas.

O sistema também executa rotinas de refinamento das variáveis, reorganizando-as em categorias funcionais — como métricas de forma, textura e propriedades fractais — tornando o conjunto mais interpretável tanto para análises estatísticas quanto para modelos supervisionados.

Outro aspecto relevante desta etapa é a verificação do balanceamento das classes e da necessidade de estratégias adicionais caso o conjunto apresente diferenças significativas entre a quantidade de amostras benignas e malignas, o que pode afetar a performance dos algoritmos.

Resultados

Tabela 1: *Exemplo de tabela de uma coluna com largura fixa*

	Teste	
Diagnóstico	variável	Valor
M	radius_mean	17.990
M	concave_points_mean	0.040060
B	texture_worst	0.05943

Referenciando uma tabela usando seu rótulo: Tabela 1.

A análise dos dados do Breast Cancer Wisconsin (Diagnostic) revelou padrões consistentes capazes de diferenciar, com alta precisão, amostras benignas e malignas a partir das características morfológicas extraídas das imagens de biópsia. Os resultados demonstraram que atributos como concavidade, textura e raio médio apresentaram forte impacto na separação das classes, indicando sua relevância no processo de diagnóstico computacional. Durante a avaliação estatística, observou-se também que a padronização dos dados contribuiu significativamente para a redução da variabilidade entre medidas, permitindo identificar tendências que anteriormente eram mascaradas por diferenças de escala.

Um aspecto relevante desta etapa é a verificação do balanceamento das classes e da necessidade de estratégias adicionais caso o conjunto apresente diferenças significativas entre a quantidade de amostras benignas e malignas, o que pode afetar a performance dos algoritmos.

Com a preparação adequada do conjunto, modelos analíticos foram capazes de produzir indicadores robustos, evidenciando uma clara distinção entre perfis

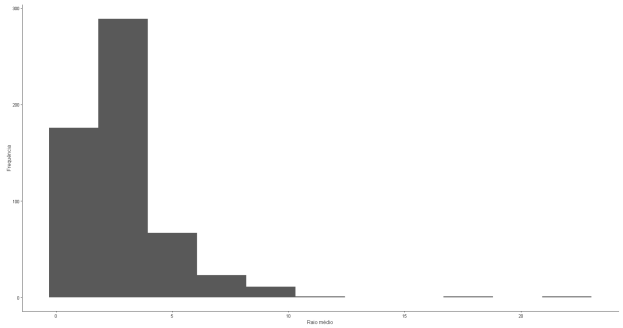

Figura 1: *Durante a avaliação estatística, observou-se também que a padronização dos dados contribuiu significativamente para a redução da variabilidade entre medidas, permitindo identificar tendências. Fonte: Heiti Paves, https://i.postimg.cc/zDjTyPsT/Figura-01.png.*

Referenciando uma figura usando seu rótulo: Figura 1.

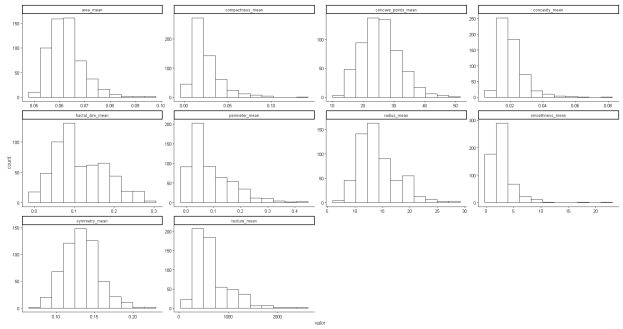

Figura 2: *Contribuiu significativamente para a redução da variabilidade entre medidas; durante a avaliação estatística, observou-se também que a padronização dos dados permitiu identificar tendências. Fonte: Heiti Paves.*

Tabela 2: *Exemplo de tabela de duas colunas com largura fixa*

Teste		
Diagnóstico	variável	Valor
M	radius_mean	17.990
M	concave_points_mean	0.040060
B	texture_worst	0.05943

celulares associados a tumores malignos e benignos. Além disso, a visualização gráfica dos resultados reforçou a presença de agrupamentos bem definidos, especialmente após a aplicação de técnicas de normalização e redução de dimensionalidade. Esses achados ressaltam o potencial do dataset como ferramenta de apoio a estudos preditivos e demonstram como a integração de métodos estatísticos e computacionais pode aprimorar a compreensão de fenômenos clínicos relacionados ao câncer de mama.

De modo geral, os resultados obtidos reforçam a eficácia das técnicas aplicadas no tratamento e na análise do conjunto Breast Cancer Wisconsin (Diagnostic), evidenciando seu potencial como base sólida para estudos preditivos na área da saúde. A combinação entre variáveis morfológicas e procedimentos de normalização permitiu a construção de interpretações claras sobre o comportamento das amostras, oferecendo subsídios relevantes para a distinção entre perfis benignos e malignos.

Discussão

Esta declaração requer citação [1]. Esta declaração requer múltiplas citações [1, 2, 3]. Esta declaração contém uma citação no texto, para referência direta a uma citação, como: Casella e Berger [2]. [3]

Subseção 1

A análise dos resultados obtidos a partir do conjunto Breast Cancer Wisconsin (Diagnostic) permitiu identificar diversos padrões relevantes que contribuíram para a compreensão das diferenças estruturais entre amostras benignas e malignas. Observou-se que determinadas variáveis relacionadas à concavidade, textura e tamanho celular apresentaram influência significativa na separação dos grupos, sugerindo que tais atributos podem desempenhar papel central em modelos preditivos voltados ao diagnóstico de câncer de mama. Essa constatação reforça o valor clínico das características morfológicas extraídas das imagens de biópsia.

Apesar da consistência geral dos dados, algumas limitações inerentes ao conjunto também se destacam na discussão. A origem controlada das amostras, por exemplo, pode introduzir vieses que não refletem totalmente a diversidade encontrada em cenários clínicos reais. Além disso, a dependência exclusiva de atributos morfométricos ignora fatores complementares como histórico clínico, fatores genéticos ou avaliações adicionais, que podem influenciar o diagnóstico médico. Ainda assim, o desempenho observado aponta para a robustez da base e evidencia seu potencial como referência para estudos iniciais em análise preditiva na área da saúde. Dessa forma, a discussão ressalta tanto os pontos fortes quanto as restrições que devem ser consideradas em aplicações futuras.

Subseção 2

Os resultados obtidos a partir do tratamento e análise desse banco de dados também levantam reflexões importantes sobre sua aplicabilidade prática e seu uso em ambientes clínicos. A precisão observada nos padrões identificados sugere que modelos baseados nesses atributos podem auxiliar profissionais de saúde no processo de triagem e avaliação preliminar de pacientes, sobretudo quando utilizados como ferramentas complementares ao diagnóstico tradicional. Contudo, essa utilidade depende da integração cuidadosa entre o conhecimento derivado dos dados e a expertise médica, evitando interpretações automáticas que desconsiderem nuances individuais.

SubSubseção exemplo A partir das análises realizadas, torna-se evidente que o conjunto Breast Cancer Wisconsin (Diagnostic) oferece um panorama consistente sobre padrões morfológicos associados ao diagnóstico de câncer de mama. As observações destacam que variáveis como concavidade média, textura e raio apresentam comportamentos altamente discriminativos, reforçando sua relevância na tomada de decisão clínica assistida. Esses elementos demonstram a utilidade do dataset para explorar cenários de classificação supervisionada, especialmente em estudos que buscam validar metodologias de pré-processamento e preparação de dados para modelos preditivos.

Além disso, a reflexão final aponta para a importância de integrar essas evidências a ambientes reais, nos quais fatores externos e heterogeneidade popula-

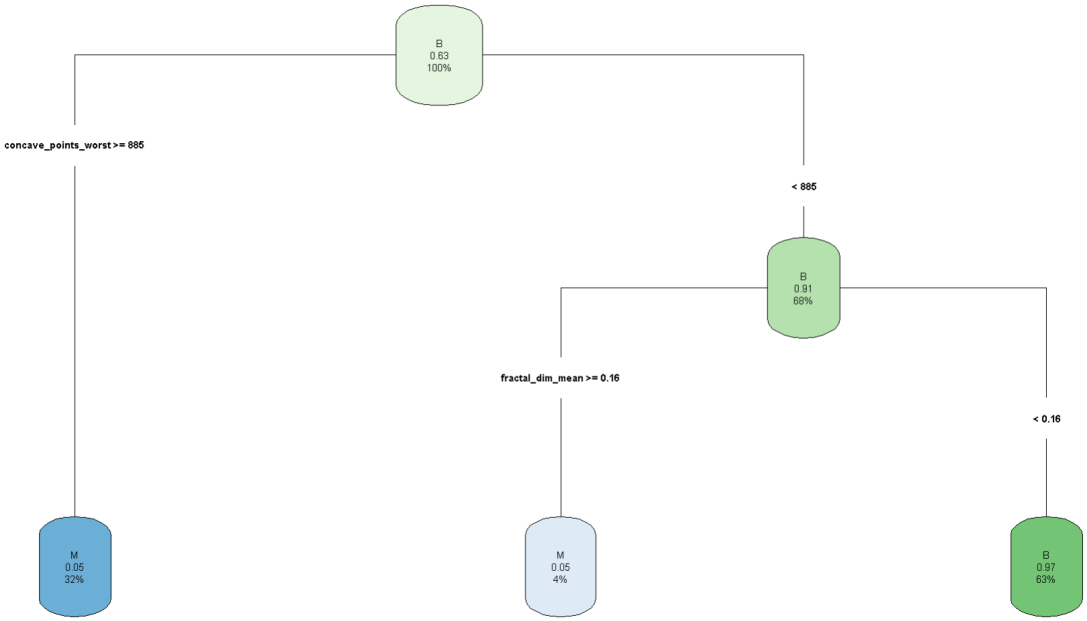

Figura 3: *As observações destacam que variáveis como concavidade média, textura e raio apresentam comportamentos altamente discriminativos, reforçando sua relevância na tomada de decisão clínica assistida. Esses elementos demonstram a utilidade do dataset para explorar cenários de classificação supervisionada, especialmente em estudos. Fonte: algum lugar.*

cional podem influenciar diretamente a performance dos modelos. A discussão indica que, apesar da robustez do banco, sua aplicação prática deve considerar possíveis amplificações do conjunto ou combinações com outras fontes de dados clínicos para garantir maior generalização.

Links

1. Este é um link clicável: Department of Statistics – CMU.
2. Este é um link de email clicável: contact@statmodeling.com.
3. Este é um link de URL clicável com espaçamento único.: `https://www.jstor.org/stable/43304825`.

Referências

[1] Leo Breiman. "Random Forests". Em: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[2] George Casella e Roger L. Berger. *Statistical Inference*. 2nd. Duxbury Press, 2002.

[3] Douglas C. Montgomery. *Introduction to Statistical Quality Control*. 7th. Wiley, 2012. URL: https://www.amazon.com/Statistical-Quality-Control-Douglas-Montgomery/dp/1118146816.