

Instituto Politécnico de Setúbal

Escola Superior de Tecnologia do Barreiro

Licenciatura em Bioinformática

Data Mining

Docentes: Ana Mendes e Luís Pereira

Estudo de casos de Cancro da Mama

Autores:

Joana Fonseca - 202100981

Pedro Pacheco - 202100957

Tiago Beja – 202000197

Índice

[Índice de figuras 3](#_Toc138185394)

[Estrutura do relatório 4](#_Toc138185395)

[Introdução 4](#_Toc138185396)

[Apresentação do caso de estudo e Análise dos dados 4](#_Toc138185397)

[Objetivos 11](#_Toc138185398)

[Desenvolvimento 12](#_Toc138185399)

[Dificuldades sentidas 13](#_Toc138185400)

[Conclusão 14](#_Toc138185401)

[Referências 14](#_Toc138185402)

# Índice de figuras

Figura 1: análise das idades que apresentam maior mortalidade de cancro da mama.

Figura 2: Análise do estado de cancro que afeta mais o tempo restante de vida.

Figura 3: Análise no número de nódulos e a sua influência no tamanho do tumor.

Figura 4: Análise do estado da menopausa e a sua influência na idade de diagnostico.

Figura 5: Análise do número de células cancerígenas e a sua influência na causa de morte.

Figura 6: Taxa de sucesso no tratamento do cancro nos diferentes data sets.

# Estrutura do relatório

Neste relatório, abordaremos diversos aspetos relacionados ao cancro de mama. Iniciaremos com uma introdução, na qual discutiremos a natureza dessa doença e a prevalência de casos diagnosticados em mulheres. Em seguida, apresentaremos os casos de estudo, destacando a análise dos dados obtidos por meio de um estudo que conduzimos com as informações disponíveis.

Nos objetivos, descreveremos as metas que buscamos alcançar ao formular perguntas específicas e analisar os resultados. No desenvolvimento, exploraremos a parte prática deste trabalho, incluindo metodologias e abordagens utilizadas para a colheita e análise de dados.

Abordaremos também as dificuldades enfrentadas ao longo do processo, destacando desafios metodológicos, limitações de dados e quaisquer obstáculos encontrados durante a pesquisa.

Por fim, na conclusão, apresentaremos os principais insights e conclusões obtidos por meio deste estudo.

# Introdução

Este relatório apresenta um estudo de casos sobre o cancro da mama, uma doença que afeta milhões de mulheres em todo o mundo.

Este trabalho foi desenvolvido em Power BI e o objetivo deste estudo é analisar diferentes casos de mulheres diagnosticadas com cancro da mama, investigando alguns fatores de risco. Ao examinar esses casos individualmente e em conjunto, busca-se obter uma compreensão mais abrangente dessa doença complexa, identificar padrões comuns e destacar a importância da deteção precoce, do tratamento adequado e do apoio multidisciplinar para o sucesso no combate ao cancro da mama.

Através da análise dos casos, esperamos contribuir para o avanço contínuo na área do cancro da mama e melhorar a qualidade de vida das mulheres afetadas por essa doença.

# Apresentação do caso de estudo e Análise dos dados

Cada figura a seguir apresentada representa um caso de estudo, uma questão que nos intrigou sobre a informação que continha nos data sets utilizados que na qual podíamos estudar e chegar a algum tipo de conclusão.

Na nossa primeira questão pensámos em encontrar as idades com maior mortalidade de cancro da mama, muitas mulheres sofrem de cancro da mama, todos os anos o número de casos aumenta e queríamos perceber em que idades é que existiam mais mortes desta mesma doença. Na figura 1 conseguimos observar que as idades nas quais existem mais mortos são 69 anos, 62 anos e 66 anos, o que conseguimos concluir que entre os 62 anos e os 69 anos é onde existe um maior número de mortes de mulheres com cancro da mama.

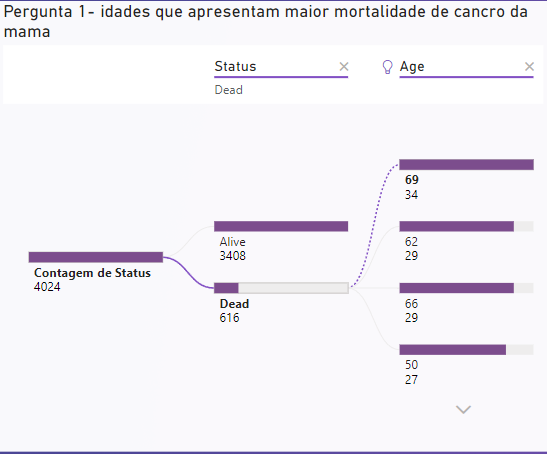


Figura1: análise das idades que apresentam maior mortalidade de cancro da mama.

Nesta nossa segunda questão pretendíamos saber como é que o tempo restante de vida seria afetado pelo quão avançado estivesse o cancro. Normalmente quanto mais avançado está o cancro menos tempo de vida se tem, mas o quão menos? Queríamos perceber se realmente uma instância afetava a outra e o impacto que tinha ao afetar. Na figura 2 conseguimos analisar que quando o cancro está no estado IIIB é quando existe o menor tempo restante de vida então conseguimos concluir que o estado de cancro não afeta o tempo restante de vida pois o pior estado de cancro (IIIC) apresenta maior tempo restante de vida do que o estado IIIB.

Uma imagem com texto, captura de ecrã, Gráfico, file

Descrição gerada automaticamente

Figura 2: Análise do estado de cancro que afeta mais o tempo restante de vida.

Com a pergunta 3 tentámos perceber se o número de nódulos influenciava o tamanho do tumor. Os nódulos que são crescimentos anormais no nosso tecido ou órgão, chamados de benignos ou malignos, sendo os malignos cancro. Ao olharmos para a figura 3 percebemos que existem pessoas com 1 nódulo, 2 nódulos e 3 nódulos, e sim… podemos concluir que influencia o tamanho, pois neste gráfico de pizza as mulheres com 3 nódulos apresentam um tumor com maior tamanho que as mulheres com 2 nódulos e 1 nódulo sendo os valores com 3 nódulos de 40,24%, com 2 nódulos de 34,63% e com um nódulo de 25,14%.

Uma imagem com texto, captura de ecrã, Tipo de letra, diagrama

Descrição gerada automaticamente

Figura 3: Análise no número de nódulos e a sua influência no tamanho do tumor.

Na questão 4 verificámos que o estado da menopausa influencia a idade de diagnóstico, isto é, se a idade de diagnóstico irá aumentar caso as pacientes já tenham passado pela menopausa. Através da figura 4 é possível tirar essa conclusão, pois a contagem de idades nas pacientes que já passaram pela menopausa é significativamente superior em comparação às pacientes que ainda não passaram por essa fase.

Uma imagem com texto, captura de ecrã, Tipo de letra, Retângulo

Descrição gerada automaticamente

Figura 4: Análise do estado da menopausa e a sua influência na idade de diagnostico.

Na pergunta 5 comparámos se o nível de células cancerígenas influência a causa de morte. O nível de células cancerígenas é avaliado por três classes: alta, media e baixa, tendo também um espaço para os resultados nulos. Como esperado, podemos concluir que quanto mais alto o nível de células cancerígenas maior o resultado do somatório de idades.

Uma imagem com texto, captura de ecrã, diagrama, roxo

Descrição gerada automaticamente

Figura 5: Análise do número de células cancerígenas e a sua influência na causa de morte.

Finalmente, na última questão averiguámos a eficácia de tratamento para as duas fontes de dados. O principal objetivo era chegar à conclusão de que ambos têm uma taxa de pacientes vivos após o tratamento superior a 50%, apesar de não ser um resultado brilhante assegurava minimamente a capacidade do tratamento da doença. Como podemos verificar através da figura 6, o data set da Metabric apresenta 56,25% muito mais fraca em relação ao segundo data set, o SEER que apresenta uma taxa de sucesso superior com 84,69% de sucesso. Embora não seja uma percentagem de sucesso muito elevada, podemos considerar que a percentagem obtida até o momento é satisfatória, considerando as dificuldades enfrentadas e o contexto em que o estudo do cancro da mama está inserido.

***Uma imagem com texto, captura de ecrã, círculo, Tipo de letra

Descrição gerada automaticamente***

Figura 6: Taxa de sucesso no tratamento do cancro nos diferentes data sets

# Objetivos

Pergunta 1- Que idades apresentam a maior mortalidade de cancro da mama?

Procurar saber que idades apresentam a maior mortalidade com cancro da mama, isto pode levar-nos a perceber se existe alguma influência com a fase da vida em que estão as mulheres em estudo e se essa mesma fase influencia o facto de morrerem de cancro na mama.

Pergunta 2- Qual o estado de cancro que afeta mais o tempo restante de vida?

Um estado mais avançado de cancro está associado a um pior prognóstico e um tempo de vida reduzido. Com esta pergunta, pretendíamos entender se realmente quanto mais avançado é o cancro, menor tempo de vida a pessoa tem ou se pode haver exceções.

Pergunta 3- Será que o número de nódulos do tumor influencia o seu tamanho?

O estudo da relação entre o número de nódulos do tumor e seu tamanho é importante para o prognóstico, planejamento do tratamento, monitoramento da resposta ao tratamento e avanços na pesquisa do cancro. Essas informações podem ajudar a melhorar a compreensão da doença e aprimorar as estratégias de tratamento para os pacientes. O objetivo de procurarmos responder a esta pergunta era percebermos se o número de nódulos ajuda no crescimento do tamanho do tumor.

Pergunta 4- O estado da menopausa influencia a idade de diagnostico?

Esta pergunta vai-nos ajudar a entender como o estado da menopausa influencia o diagnóstico do cancro de mama, pois este é crucial para melhorar as estratégias de prevenção, diagnóstico precoce e tratamento da doença, levando a melhores resultados para as mulheres afetadas, visto que por norma uma mulher na pós-menopausa tem um risco ligeiramente maior de ser diagnosticada com cancro da mama em comparação com mulheres na pré-menopausa.

Pergunta 5- O nível de células cancerígenas influência a causa de morte?

Quanto maior a carga tumoral ou a quantidade de células cancerígenas presentes no organismo, maior é o impacto na saúde geral e na sobrevivência do paciente. Este estudo ajudou-nos a compreender como a carga tumoral afeta a causa de morte e com isso ajudar a identificar pacientes de alto risco, permitindo intervenções precoces e personalizadas.

Pergunta 6- Taxa de sucesso do tratamento contra o cancro?

A taxa de sucesso do tratamento contra o cancro é um indicador crucial que avalia a eficácia dos diferentes métodos e abordagens terapêuticas utilizadas no combate à doença. Essa taxa representa a proporção de pacientes que respondem positivamente ao tratamento, seja alcançando a cura, remissão completa ou controle da doença.

Esta foi a parte mais importante do nosso estudo, pois o cancro é uma doença muito comum e depois de todas as nossas análises era relevante sabermos qual é a taxa de sucesso no tratamento da doença.

# Desenvolvimento

Com os data sets obtidos deparamo-nos com dados que eram numéricos mas eram reconhecidos como texto, tivemos então de os trocar para dados numéricos para conseguirmos utilizá-los no estudo. Procedemos então à criação das perguntas das quais nos iriamos basear para o nosso estudo, ao analisar os dados conseguimos chegar a 6 (seis) perguntas que considerámos mais relevantes, sendo elas: Que idades apresentam a maior mortalidade de cancro da mama?, qual o estado de cancro que afeta mais o tempo restante de vida?, será que o número de nódulos do tumor influencia o seu tamanho?, o estado da menopausa influencia a idade de diagnóstico?, o nível de células cancerígenas influencia a causa de morte? e por fim, a taxa de sucesso do tratamento contra o cancro.

Com a definição das nossas perguntas e os dados que queríamos utilizar para o nosso estudo procedemos para a criação dos gráficos apresentados no índice de gráficos, na criação destes gráficos contruímos 5 medidas, as quais utilizámos para a criação de alguns dos gráficos

Para a primeira questão decidimos utilizar uma árvore de decomposição em que relaciona a contagem de mortes com a idade dos pacientes. O principal objetivo da abordagem do gráfico é mostrar que idades apresentam mais mortes, para entender que faixas etárias estariam em maior risco.

Na segunda questão decidimos que a abordagem mais correta seria um gráfico de linhas e colunas empilhadas, em que nas colunas relacionamos o número de mortos por cada fase do cancro e a linha varia com o número de pacientes em cada fase do cancro. Utilizámos este gráfico para conseguir mostrar simultaneamente que apesar de uma fase ter mais mortes também pode implicar que esse fenómeno pode estar a acontecer devido ao maior número de pacientes.

Com a terceira questão queríamos relacionar o número de nódulos linfáticos afetados com o tamanho médio do tumor por nódulo. Na nossa opinião o gráfico que melhor representaria a dispersão de dados seria um gráfico circular para mostrar a variação de tamanho por nódulo, isto é, relacionar que nódulos em média são maiores.

O objetivo principal da quarta questão é relacionar a idade de diagnóstico do cancro e se a paciente já passou pela menopausa, utilizámos um gráfico de barras que cresce com as idades de diagnóstico, mostrando assim que existe uma relação com o diagnóstico da menopausa e a idade de diagnóstico dos pacientes, logo, os médicos que acompanham os pacientes nos estado de menopausa possam estar alertas com o aparecimento de cancro da mama.

Na quinta questão, o objetivo é mostrar que quanto maior o nível de células cancerígenas maior será a taxa das mortes de pacientes e definimos que a melhor estratégia seria através de um gráfico de colunas empilhadas.

Finalmente na última questão decidimos incidir o estudo na eficácia do tratamento. Juntos decidimos que a estratégia mais adequa é através de um gráfico circular comparando a taxa de pacientes vivos com os pacientes mortos, podendo mostrar assim o objetivo da nossa última questão para ambos os data set.

# Dificuldades sentidas

Durante o desenvolvimento deste relatório enfrentámos algumas dificuldades significativas. Uma das principais questões que enfrentámos foi a criação de medidas adequadas para os gráficos, de forma a representar com precisão os dados que procurávamos. Encontrámos desafios na transformação de dados numéricos, uma vez que alguns deles foram reconhecidos como texto.

Outro obstáculo encontrado foi a escassez de conjuntos de dados relevantes e interessantes para o nosso estudo. Encontrar fontes de dados com informações pertinentes sobre o cancro da mama foi uma tarefa desafiadora. No entanto, com uma pesquisa aprofundada e uma análise criteriosa, conseguimos localizar conjuntos de dados adequados para atender às nossas necessidades de pesquisa.

Apesar das dificuldades enfrentadas, conseguimos superá-las com sucesso. Essas superações permitiram-nos concluir este trabalho de maneira satisfatória, fornecendo um relatório completo sobre o estudo do cancro da mama.

# Conclusão

Em conclusão, o estudo abrangente do cancro da mama revelou informações valiosas sobre diferentes aspetos da doença. Ao analisar o número de nódulos tumorais, o estado do cancro, a idade e o estado da menopausa das pacientes, o nível de células cancerígenas e a taxa de sucesso do tratamento, pudemos obter resultados significativos.

A análise desses fatores permitiu uma compreensão mais aprofundada da progressão e do prognóstico do cancro da mama. Identificar a presença de múltiplos nódulos, o estado da doença e outros fatores de risco fornece informações essenciais para o planeamento e a tomada de decisões em relação ao tratamento.

Além disso, a análise do nível de células cancerígenas desempenha um papel crucial na determinação da agressividade do tumor e no ajuste do protocolo terapêutico adequado. Esses dados também podem ser úteis na avaliação do progresso do tratamento ao longo do tempo.

No geral, o estudo abrangente do cancro da mama baseado na análise de múltiplos aspetos da doença fornece uma base sólida para aprimorar as estratégias de diagnóstico, tratamento e acompanhamento dos pacientes. A compreensão aprofundada dessas informações pode levar a abordagens mais personalizadas e eficazes, melhorando os resultados e a qualidade de vida das pacientes afetadas pelo cancro da mama.

# Referências

<https://www.kaggle.com/datasets/piotrgrabo/breastcancerproteomes>

<https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric>