

Análise de Dados - Exercício 1

Aluno: Cláudio Mendes Ramos

Número: 2022121538

Características do Dataset

Fonte: <https://www.kaggle.com/datasets/ak0212/uae-cancer-patient-dataset>

O dataset contém 10.000 registos e 20 colunas, com dados relacionados ao diagnóstico, tratamento e características demográficas de pacientes com cancro nos Emirados Árabes Unidos.

Dados em Falta

Os atributos death_date e cause_of_death têm aproximadamente 90% de valores em falta, o que é esperado, visto que apenas uma pequena parte dos pacientes faleceu. Os restantes campos não apresentam valores ausentes.

Tipos de Cancro Mais Frequentes

Os cinco tipos de cancro mais comuns são:

- Leucemia (1314 casos)
- Fígado (1263 casos)
- Ovário (1259 casos)
- Pâncreas (1243 casos)
- Mama (1241 casos)
- Distribuição por Género
- Mulheres: 4985 pacientes
- Homens: 4827 pacientes
- Outro: 188 pacientes

A distribuição é bastante equilibrada entre homens e mulheres.

Estatísticas de Idade

- Idade média: 53,5 anos
- Idade mínima: 18 anos
- Idade máxima: 89 anos
- 25% mais jovens: até 36 anos
- 25% mais velhos: acima de 72 anos

Distribuição de Resultados Clínicos

- Recuperados: 4931 pacientes
- Em tratamento: 4077 pacientes
- Faleceu: 992 pacientes

A maioria dos pacientes recuperou, com uma parte ainda em tratamento e cerca de 10% que faleceram.

Emirados com Mais Registos

Os cinco emirados com maior número de casos são:

- Ajman (1497)
- Abu Dhabi (1471)
- Fujairah (1471)
- Umm Al Quwain (1432)
- Dubai (1407)

Estes dados podem refletir diferenças populacionais ou na qualidade dos registos hospitalares.

Atributos principais:

- patient_id: identificador único do paciente
- age, gender, nationality, ethnicity: dados demográficos
- emirate: localização geográfica
- cancer_type, cancer_stage: diagnóstico clínico
- treatment_type, treatment_start_date: tipo de tratamento
- outcome: resultado clínico (recuperado, em tratamento, falecido)
- death_date, cause_of_death: apenas se aplicável
- weight, height: dados físicos
- smoking_status, comorbidities: fatores de risco

Este dataset é ideal para análises de saúde pública, eficácia de tratamentos e identificação de grupos de risco com base em variáveis demográficas e clínicas.