

*Aprendizagem e Mineração de Dados*

Projeto Final – A

Semestre de inverno 2023/2024 - MI1D

Grupo x   
Gonçalo Silva – 47255   
João Rocha – 47xxx

Luís Morgado -

Docente

Eng.º

Data: 1/10/2023

Índice

[Projeto A 2](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark0)

[Introdução 2](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark1)

[Trabalho Diário/ Enquadramento 2](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark2)

[Análise dos Dados 2](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark3)

[Modelos de Dados 4](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark4)

[Criação da base de dados – Scripts 6](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark5)

[Classificador 1R 7](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark6)

[Implementação 1R 7](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark7)

[Árvore de decisão – ID3 8](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark8)

[Naïve Bayes - NB 9](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark9)

[Implementação ID3 e Naïve Bayes 9](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark10)

[Classificação dos Modelos 10](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark11)

[Projeto A1 12](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark12)

[Introdução 12](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark13)

[Análise dos Dados 12](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark14)

[Converter um ficheiro csv para um ficheiro tab 12](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark15)

[Aplicação do Algoritmo 1R 12](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark16)

[Orange 14](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark17)

[Conclusão 18](file:///C:\Users\gonca\OneDrive\Ambiente%20de%20Trabalho\Trabalhos%20Amorim\TrabalhosAmd-bacana\AMD_D_01_A_A1.docx#_bookmark18)

# Projeto A

## Introdução

Com este trabalho, pretende-se explorar e compreender inter-relações entre dados, informação e conhecimento, aplicando esses conceitos em cenários práticos e explorando diversas técnicas e métodos que permitam converter dados em conhecimento, de modo a desenvolver a competência de analisar, modelar e validar um projeto de mineração de dados.

## Enquadramento

#### O centro médico especializado em oftalmologia, MedKnow, utiliza um sistema de gestão de bases de dados (SGBD) que armazena todos os dados reunidos ao longo do tempo referentes a cada consulta. A MedKnow forneceu o arquivo "d01\_lenses.xls", contendo um conjunto de dados específico relacionado à atividade de prescrição de lentes nas últimas duas semanas (conforme apresentado na Tabela 1), algumas premissas e padrões conseguem ser identificados, fornecendo informação valiosa para a identificação de tendências na atividade diária da clínica oftalmológica.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **age** | **prescription** | **astigmatic** | **tear\_rate** | **lenses** |
| young | myope | yes | normal | hard |
| young | myope | no | normal | soft |
| young | hypermetrope | yes | reduced | none |
| young | hypermetrope | no | normal | soft |
| young | hypermetrope | no | reduced | none |
| presbyopic | myope | yes | reduced | none |
| presbyopic | myope | yes | normal | hard |
| presbyopic | hypermetrope | yes | reduced | none |
| presbyopic | hypermetrope | yes | normal | none |
| presbyopic | hypermetrope | no | normal | soft |
| presbyopic | hypermetrope | no | reduced | none |
| pre-presbyopic | myope | yes | reduced | none |
| pre-presbyopic | myope | yes | normal | hard |
| pre-presbyopic | myope | no | normal | soft |
| pre-presbyopic | hypermetrope | yes | normal | none |
| pre-presbyopic | hypermetrope | no | normal | soft |

*Tabela 1 - snippet de dados relacionado com a atividade de prescrição de lentes*

Com base nos dados da Tabela 1, podemos afirmar que a *Medknow* categoriza os seus clientes com base na sua idade (jovem, pré-presbiópico, presbiópico), tipos de prescrição (miopia e hipermetropia), presença de astigmatismo (sim/não), taxa de lágrima (normal/reduzida) e para cada um deles, podem ser prescritos 3 tipos diferentes de lentes (nenhuma, moles, rígidas).

## Análise dos dados

#### Depois de uma análise completa deveremos ser capazes de saber que tipo de lentes devem ser escolhidas para cada paciente dependendo da sua idade e dos seus aspetos de saúde ocular.

#### Podemos observar que os dados da Tabela 1, são estruturados num formato tabular, com valores nominais em domínio discreto, quanto à sua semântica, o dataset apresenta 5 valores, que podem ser divididos em 4 atributos, sendo estes a sua idade (jovem, pré-presbiópico, presbiópico), tipos de prescrição (miopia e hipermetropia), presença de astigmatismo (sim/não), taxa de lágrima (normal/reduzida) e 1 classe/alvo que são as lentes, que podem ser representadas por 3 tipos diferentes de lentes (nenhuma, moles, rígidas).

## Modelo de dados

## Foi construído um modelo concetual de dados utilizando a notação Entidade-Relação (Entity-Relationship notation) para o apoio operacional (trabalho diário) e um modelo lógico derivado do mesmo. Tais modelos são cruciais para a organização dos dados e futura análise.

## Modelo Entidade-Relação

## A black and white diagram Description automatically generated

*Figura 1 – Modelo Entidade-Relação*

## Modelo Lógico

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

*Figura 2 – Modelo Lógico*

## Detalhes de Implementação

**Modelo de Dados e Associações entre Tabelas**

O modelo de dados é centrado na tabela "Appointment," que serve como o núcleo das informações clínicas. Essa tabela está relacionada com todas outras tabelas do sistema. Para associar as informações sobre as doenças a uma consulta, foi criada a tabela "DIAGNOSIS." Nessa associação, reconhecemos que uma única consulta pode estar relacionada a mais de uma doença. No entanto, no contexto clínico, esperamos que haja exatamente uma doença do tipo "myope" ou "hypermetrope" associada a cada consulta. Além disso, opcionalmente, pode estar associada uma doença do tipo "astigmatic."

**Atributos Chave e Associações Diretas:**

O valor de "TearRate," assume-se que é atribuído um único valor à consulta, refletindo uma medição específica, estando este valor diretamente descrito na tabela “APPOINTMENT”. Da mesma forma, "LensHardness" também é atribuído como um único resultado, no entanto, esses resultados foram armazenados numa tabela separada, uma vez que representam os resultados específicos obtidos durante cada consulta.

**Intervalos de Valores e Mapeamento:**

Para caracterizar os intervalos de valores possíveis para as medidas de idade, TearRate e LensHardness, foram criadas três tabelas de mapeamento. Essas tabelas servem como referências para definir faixas de valores aceitáveis. Cada intervalo é mapeado por meio das tabelas "OcularAge," "TearRate," e "LensHardness," proporcionando uma maneira padronizada de representar e consultar os resultados clínicos dentro dessas faixas.

Essas associações e representações foram projetadas para garantir uma estrutura de dados coerente e eficaz. O uso de tabelas de mapeamento ajuda a manter a consistência e a integridade dos dados, fornecendo uma estrutura clara para as informações clínicas em questão. Os valores foram organizados de acordo com a seguinte tabela:

|  |  |  |
| --- | --- | --- |
| Mínimo | Máximo | *Valor* |
| *OcularAge* | | |
| 0 | 30 | *“young”* |
| 30 | 45 | *“pre-presbyopic”* |
| 45 | 99 | *“presbyopic”* |
| *LensHardness* | | |
| 0 | 0.1 | *“none”* |
| 0.1 | 0.3 | *“soft”* |
| 0.3 | 0.99 | *“hard”* |
| *TearRate* | | |
| 0 | 0.5 | *“reduced”* |
| 0.5 | 1.0 | *“normal”* |

Discretização

Com base no conjunto de dadosapresentado na *tabela 1,* podemos concluir que todos os valores dos atributos são valores nominais, com um intervalo de valores bem definido.

Para chegar a estes valores, é necessário um processo de discretização, pois normalmente dados como a idade e a taxa de lagrima costumam ser valores numéricos.

Como mencionado anteriormente, precisamos converter todos os dados numéricos em valores nominais e podemos fazer isso atribuindo nomes a determinados intervalos de valores. Por exemplo, podemos dizer que pessoas de 0 a 30 anos são "jovens", de 31 a 60 anos são "pré-presbiópicas" e com mais de 60 anos são "presbiópicas". Ao fazer este processo, estamos a converter um potencial intervalo infinito de valores em apenas 3 valores: "jovem", "pré-presbiópico" e "presbiópico". Este processo é especialmente importante neste projeto, pois o Classificador One Rule, que é um dos classificadores obrigatórios, só pode trabalhar com valores discretos.

Classificador 1R - One Rule

O classificador "One Rule" tem como ideia principal encontrar a regra que melhor discrimina as classes alvo com base em um único atributo.

De forma a descobrir qual o atributo que melhor classifica os dados, é construída uma regra que relacione cada seu valor (desse atributo) com um dos valores do conceito (classe).

Processo para descobrir a “One Rule”:

1. Escolher um atributo para analise (o processo deve ser repetido para cada atributo).
2. Calcular a frequência de cada valor do atributo em relação as classes alvo.
3. Calcular o erro para para cada valor do atributo.
4. Escolher os pares (valor-atributo, valor-classe) com menor erro, aleatório se igual.
5. Calcular o erro do atributo somando os erros dos pares escolhidos.
6. Depois de obter as regras e os erros de cada atributo (itens acima), é escolhido o atributo com o menor erro e considera-se as regras associadas a esse atributo como representativas do conjunto de dados.

Implementação 1R - One Rule