Atividade 2 - Visualização de informação

Leonardo Cesar Silva dos Santos Fernando Augusto Cardoso Candalaft

1 Análise do dataset Wine

Nesta atividade exploramos o conjunto de dados *Wine* com o objetivo de identificar o relacionamento de cada classe de vinho com as respectivas *features disponíveis*.

Para tal, analisamos o comportamento das *features* com relação a cada classe disponível usando um gráfico do tipo *heatmap* com ordenação *optimal leaf ordering* e a métrica de avaliação *correlation* (Figura 1). A partir disso podemos observar que:

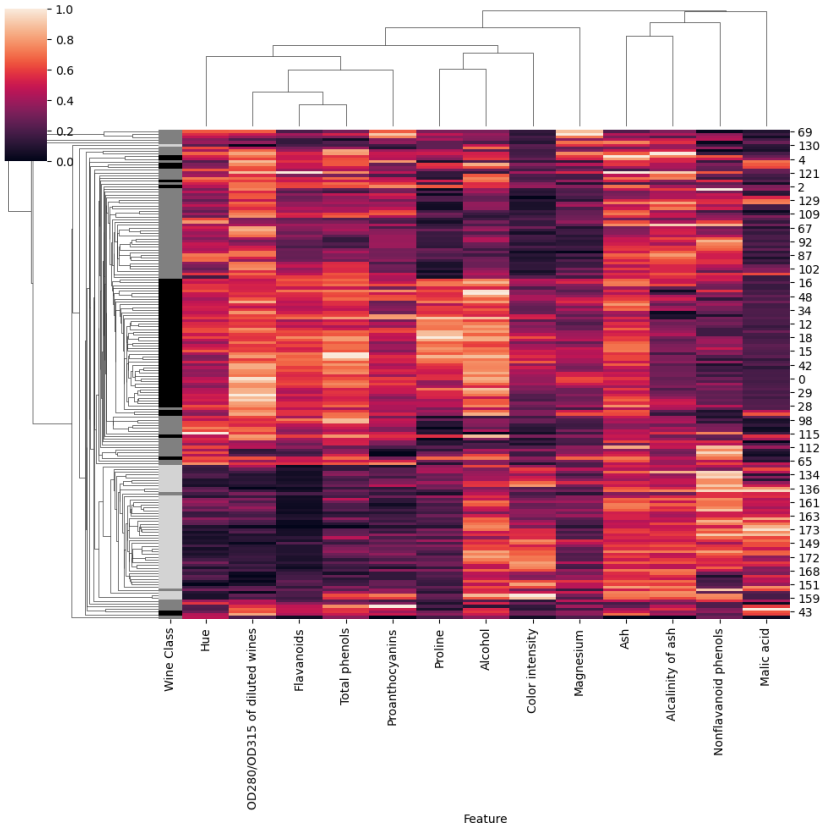

Figura 1: Mapa de calor para as *features* do dataset *Wine*. Classe 1 está representada pela cor preta, a classe 2 pela cor cinza e a classe 3 pela cor cinza claro no eixo y.

- para a **classe** 1 (cor preta) as variáveis *OD*280*, Hue, Proanthocyanins* e *Flavanoids* possuem correlação mais próxima entre si, de modo que não trazer tanta diversidade em informação, diferentemente das variáveis *Malic, Alcalinity* e *Nonflavanoid*, que possuem uma correlação baixa.
- para a **classe** 2 (cor cinza) as variáveis que trazem mais variedade de informação são *Proline, Color* e *Magnesium*.
- para a **classe** 3 (cor cinza claro) as variáveis que trazem mais diversidade de informação são *OD*280*, Hue, Proanthocyanins* e *Flavanoids* (as que trazem menos informação para a classe 1).

Portanto, é possível notar que as informações de cada *feature* são complementares com relação a cada classe disponível.

2 Analisando o conteúdo textual da bíblia

Para esta tarefa utilizamos alguns capítulos específicos da bíblia. Além dos disponibilizados previamente (Gênesis, Êxodo, Mateus e Marcos) estudamos também os capítulos Romanos, Galatas, Efesios e Apocalipse.

Para formular os valores de *td-idf* de cada capítulo novo processamos os dados textuais de modo a remover: pontuações e acentuações. Também convertemos todas as palavras para *lowercase* (o que a priori não estava feito nos capítulos com os cálculos já disponíveis).

Após isso obtimos os respectivos valores para a tabela *td-idf* (Tabela 1).

Tabela 1: Tabela *TD-IDF*

id	**Espírito**	**Deus**	**Jeová**	**batismo**	**Senhor**	**Jesus**	**Cristo**
genesis.txt	$8,91 \times 10^{-5}$	0	0	0	0	0	0
exodo.txt	$5,25 \times 10^{-5}$	0	$3,78 \times 10^{-4}$	0	0	0	0
mateus.txt	$2,402 \times 10^{-4}$	0	0	$4,25 \times 10^{-5}$	0	$1,1463 \times 10^{-3}$	$1,059 \times 10^{-4}$
marcos.txt	$4,574 \times 10^{-4}$	0	0	$1,363 \times 10^{-4}$	0	$1,3189 \times 10^{-3}$	$6,99 \times 10^{-5}$
romanos.txt	0	0	0	0	0	$6,727 \times 10^{-4}$	$7,175 \times 10^{-4}$
galatas.txt	0	0	0	0	0	$6,727 \times 10^{-4}$	$1,6593 \times 10^{-3}$
efesios.txt	0	0	0	$1,529 \times 10^{-4}$	0	$8,969 \times 10^{-4}$	$2,0629 \times 10^{-3}$
apocalipse.txt	0	0	0	0	0	$2,242 \times 10^{-4}$	$1,345 \times 10^{-4}$

Analisamos o relacionamento entre os capítulos por meio da distribuição de algumas palavras pré-selecionadas: Espírito, Deus, Jeová, batismo, Senhor, Jesus e Cristo. Para tal análise utilizamos o auxílio do gráfico *MDS* (Figura 2).

É possível notar pela Figura 2 que os capítulos da biblía que mais se distanciam um do outro com relação ao conteúdo são: Mateus, Efesios, Marcos e Exodo. Por outro lado, os que são mais parecidos são: Genesis, Romanos, Galatas e Apocalipse.

Porém, vale lembrar que algumas etapas de pré-processamento dos capítulos adicionais estudados podem ter sido diferentes em relação aos capítulos já pré-selecionados de modo que a frequência de algumas palavras podem estar diferentes. Vale lembrar também que o espaço analisado de palavras está restrito, de modo que um espaço vetorial com uma dimensão mais alta pode trazer resultados um pouco diferentes.

2.1 Análise do dataset Sleep Study

Para este conjunto de dados buscamos entender o relacionamento entre as variáveis e qualidade de sono dos indivíduos analisados. Deste modo, focamos em responder às seguintes perguntas:

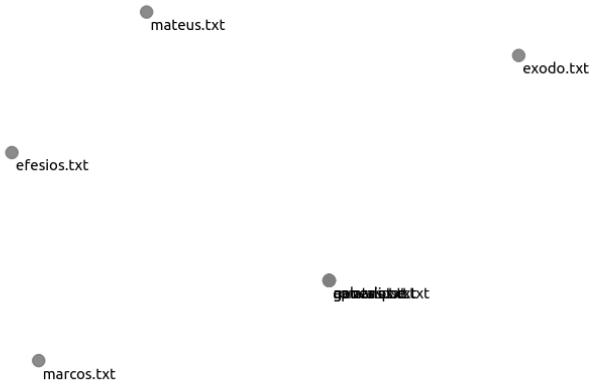

Figura 2: Relacionamento entre cada capítulo por meio das palavras pré-selecionadas.

- Qual a distribuição da quantidade de horas que os indivíduos dormem e qual a relação entre a quantidade de horas dormidas e qualidade do sono (sufiente ou não).
- Existe alguma relação entre o nível de cansaço com outras variáveis disponíveis?
- Existe algum indicativo de benefício em tomar café da manhã para a qualidade do sono?
- Alguma variável afeta mais o sono do que outra?

Para responder tais perguntas utilizamos o auxílio dos gráficos das Figuras 3 a 7. Pela Figura 3 podemos notar que as pessoas satisfeitas com a quantidade de horas dormidas se concentram entre 7 e 8 horas de sono. Já as pessoas insatisfeitas com a quantidade de horas dormidas tendem a dormir 6 horas ou menos.

A Figura 4 exibe a relação entre a qualidade do sono (coloração representando os valores da coluna *Enough*) com outras variáveis e o nível de cansaço (tamanho dos pontos: *Tired*). Por meio desta figura é possível notar que pessoas com *PhoneTime* (pessoas que mexem no celular com até 30 minutos ao pegar no sono) e pessoas com *PhoneReach* (pessoas com telefone ao alcance do braço) tendem a ser as que estão mais cansadas e também as que estão insatisfeitas com a quantidade de horas que dormem.

Na Figura 5 analisamos a relação entre nível de cansaço (size: *Tired*), a satisfação com o sono (color: *Enough*) e o fato de a pessoa ter tomado café ou não (*label: Yes/No*). É possível ver que não há uma relação clara entre essas variáveis. A Figura 6 também indica que tomar café não necessariamente ajuda na satisfação com a quantidade de horas dormidas, uma vez que da quantidade de pessoas insatisfeitas (eixo x) a maioria toma café da manhã.

Na Figura 7 analisamos o relacionamento entre as pessoas que estão satisfeitas ou não com a quantidade de horas dormidas e a relação disso com o tempo de tela ao dormir (*color: PhoneTime*. É possível notar que das pessoas insatisfeitas, grande parte usa o telefone quando está entrando em estado de sono, o que indica que tal tempo de tela é maléfico à satisfação com as horas dormidas.

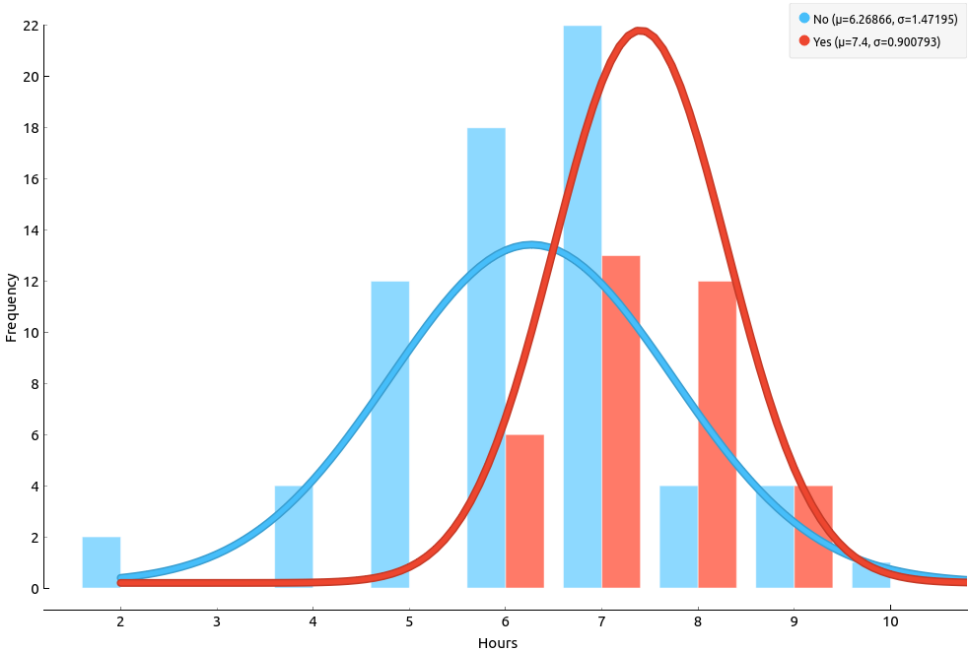

Figura 3: Distribuição de horas dormidas por grupo satisfeito com o sono.

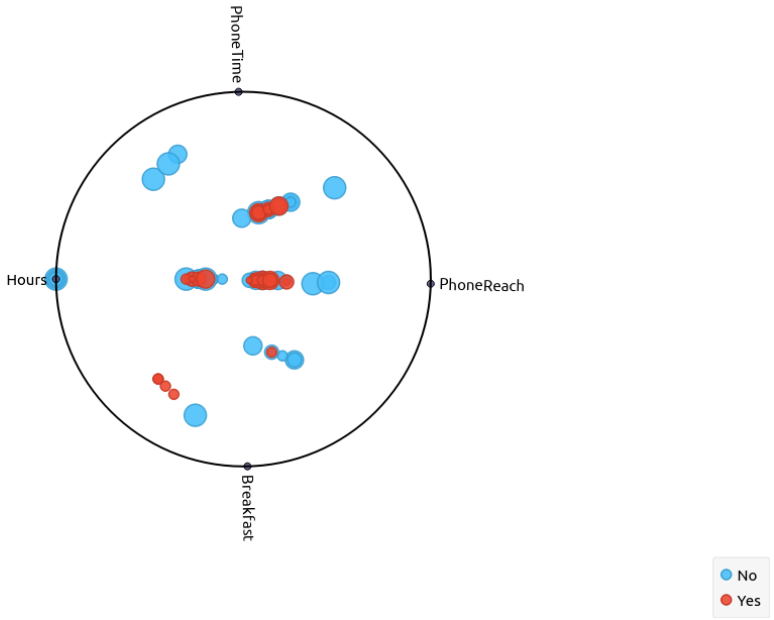

Figura 4: Relação entre a satisfação do nível de sono com outras variáveis.

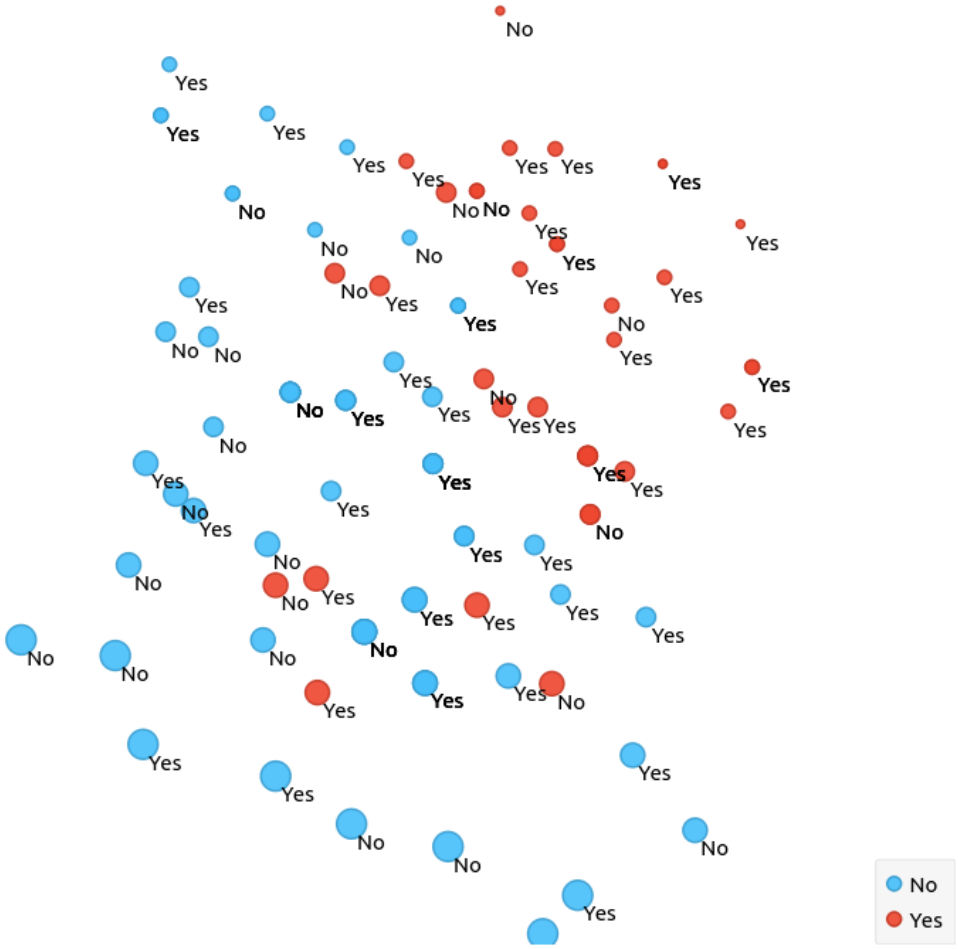

Figura 5: Relacionamento entre tomar café da manhã e estar satisfeito com a quantidade de horas dormidas.

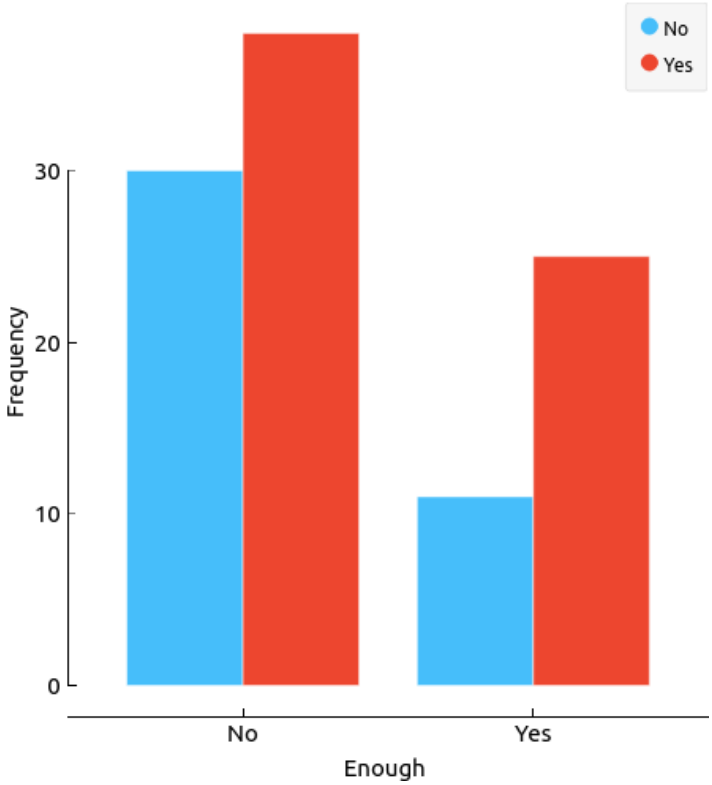

Figura 6: Quantidade de pessoas que tomam café por grupo de satisfação de horas dormidas.

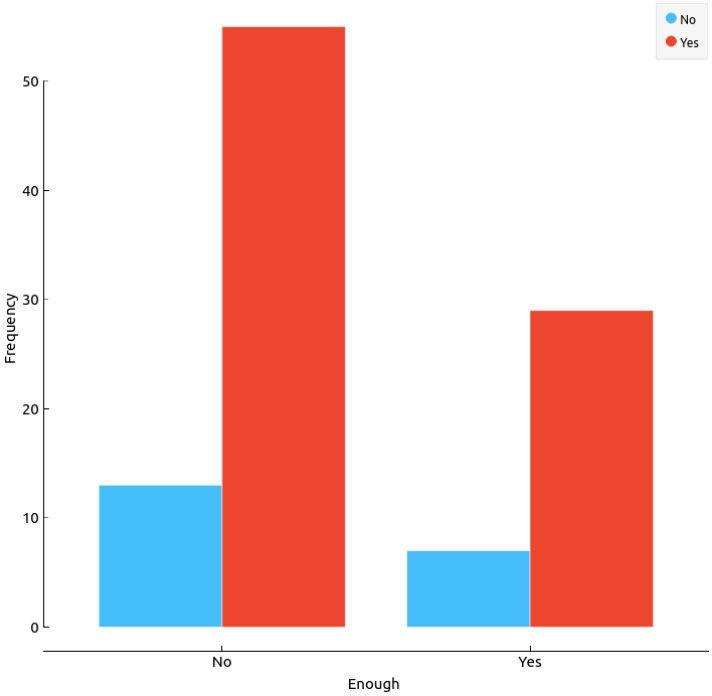

Figura 7: Análise do relacionamento entre os elementos via redução de dimensionalidade.