Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

DCC057 Mineração de Dados

Trabalho Prático 1

Alexander Thomas Mol Holmquist

25 de janeiro de 2021

1 Introdução

Em tempos de calamidade que resulte em fome extrema, ao se deparar com um cogumelo, pode ser tentador comê-lo. Todavia, alguns cogumelos carregam toxinas que podem ser letais ao serem ingeridas; se acontecer que o escolhido na hora do desespero é venenoso, a situação posterior do sujeito agravará bastante.

Enciclopédias fornecem informações precisas para a distinção de um cogumelo qualquer, mas não é sábio sempre contar com elas. O presente estudo apresenta um conjunto de três características físicas de um cogumelo que revelam com convicção 11,33 que ele não deve ser ingerido.

Muitos trabalhos já foram realizados nessa área, inclusive com a mesma tabela de dados, e no mesmo tópico. O arquivo agaricus-lepiota.names, fornecido em pacote com a tabela em questão, traz essas informações. Contudo, este trabalho não tem o enfoque principal científico, mas sim o da exploração de uma técnica de mineração de dados, a saber, a mineração de conjuntos frequentes.

2 Metodologia

A princípio, a plataforma Lemonade foi explorada como ferramenta para minerar os padrões buscados. Contudo, após encontrar muitos erros e decepções, como não conseguir rodar código Python3 normalmente, e um algoritmo de mineração que nunca acaba, decidiu-se por implementar tudo localmente em Python3 dentro de um único Jupyter Notebook.

Esse caderno está disponível, juntamente com todos os arquivos associados ao projeto, em [1].

A metodologia CRISP-DM foi utilizada como padrão. Não foi seguida à risca, mas sim tomada como uma referência, consultada sempre que preciso. A principal razão de não ser tomada como critério absoluto é que o contexto presente não é empresarial. Apesar disto, e ainda que fora de ordem, as tarefas constantes no processo CRISP-DM foram realizadas em um nível rudimentar.

Vale notar que se tornou muito claro a praticidade e conexão com realidade que expressa o caráter cíclico do método CRISP. Como exemplo, constantemente durante o projeto voltou-se tanto à fonte dos dados como a fontes educativas externas para construir um entendimento sólido sobre o contexto (negócio) e sobre os dados em si (fases 1 e 2 do processo).

3 Dados Selecionados, Tratamento

```
In [167]: originalDF.sample(5)
```

Out[167]:

	Class	Cap-shape	Cap-surface	Cap-color	Bruises	Odor	Gill-attachment	Gill-spacing	Gill-size	Gill-color	...	Stall surfac below rin
7793	p	k	s	n	f	y	f	c	n	b	...	
6851	p	f	y	e	f	s	f	c	n	b	...	
3506	p	x	f	g	f	f	f	c	b	p	...	
7529	e	b	s	g	f	n	f	w	b	p	...	
1312	p	f	s	w	t	p	f	c	n	k	...	

5 rows × 23 columns

Figura 1: Amostra dos dados

```
In [171]: originalDF.nunique()
```

```
Out[171]: Class                       2
          Cap-shape                   6
          Cap-surface                 4
          Cap-color                  10
          Bruises                     2
          Odor                        9
          Gill-attachment             2
          Gill-spacing                2
          Gill-size                   2
          Gill-color                 12
          Stalk-shape                 2
          Stalk-root                  5
          Stalk-surface-above-ring    4
          Stalk-surface-below-ring    4
          Stalk-color-above-ring      9
          Stalk-color-below-ring      9
          Veil-type                   1
          Veil-color                  4
          Ring-number                 3
          Ring-type                   5
          Spore-print-color           9
          Population                  6
          Habitat                     7
```

Figura 2: Número de entradas únicas, para cada coluna

A fonte de dados escolhida está disponível publicamente[2], e foi criada inicialmente pela *Audobon Society Field Guide*, em 1981, e doada para UCI em 1987. Se refere a cogumelos dos gêneros *Agaricus* e *Lepiota*, de distribuição cosmopólita[3, 4].

Os atributos são a classe do cogumelo - comestível ou venenoso, e 22 características físicas descritivas (todas variáveis categóricas). Há um total de 8124 registros. Destes atributos, somente o nomeado ”stalk-root”possui entradas faltantes. A decisão foi tomada de ignorar essas entradas, pois não influencia a obtenção dos conjuntos frequentes, e não está incluido no conjunto escolhido ao final.

Inicialmente, o plano era filtrar somente as entradas referentes a cogumelos venenosos, mas essa decisão foi posteriormente retraída, pois é claramente uma grande e desnecessária perda de informações valiosas para nosso objetivo.

A coluna ”Veil-type”foi removida por só conter o valor ”p”. Além disso, única transformação aplicada nos dados foi a codificação one-hot, através da biblioteca do sklearn[5]. Nada além disso se mostrou necessário.

4 Resultados Experimentais e Análise

O experimento realizado é muito simples, porque os objetivos do projeto são muito simples. Foi utilizado o pacote Python mlxtend[6].

Após a codificação one-hot, aplicou-se o algoritmo apriori para mineração de conjuntos frequentes, e extraiu-se as regras de associação no formato ({A, B, C} para {Venenoso}).

Vale notar que ao aplicar o algoritmo apriori, deparou-se com problemas de memória, por conta do volume relativamente grande de dados. Para consertar esse problema, a biblioteca utilizada disponibiliza um parâmetro que, quando ativado, resolve esta questão.

```
In [11]: filteredRules.sort_values(by="lift", ascending=False).head(10)
```

Out[11]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	levera
2161	(34, 109, 22)	(1)	0.322994	0.482029	0.308223	0.954268	1.979692	0.1525
2243	(86, 109, 22)	(1)	0.322994	0.482029	0.308223	0.954268	1.979692	0.1525
2181	(86, 35, 22)	(1)	0.412112	0.482029	0.392418	0.952210	1.975423	0.1937
2120	(34, 35, 22)	(1)	0.409897	0.482029	0.390202	0.951952	1.974887	0.1926
2208	(86, 59, 22)	(1)	0.291974	0.482029	0.274249	0.939292	1.948623	0.1335
2361	(34, 59, 86)	(1)	0.289759	0.482029	0.272033	0.938828	1.947660	0.1323
2133	(34, 59, 22)	(1)	0.289759	0.482029	0.272033	0.938828	1.947660	0.1323
2140	(34, 22, 63)	(1)	0.283604	0.482029	0.265879	0.937500	1.944906	0.1291
2222	(63, 86, 22)	(1)	0.283604	0.482029	0.265879	0.937500	1.944906	0.1291
2375	(34, 86, 63)	(1)	0.283604	0.482029	0.265879	0.937500	1.944906	0.1291

Figura 3: Regras de asssociação

A partir dessa tabela gerada, foi escolhido o conjunto {86, 109, 22}, por ter alto lift, e corresponder a atributos de fácil identificação no dia-a-dia. Outros conjuntos apresentavam maior *leverage*, como o {86, 35, 22}, mas foram rejeitados por duas razões. Um, que nesse contexto, maior *leverage* ou suporte por si só tendem a ser uma desvantagem, pois por se limitar a dois gêneros de cogumelos, a representatividade da população é menor, e conjuntos com suporte muito alto terão dificuldade em extrapolar a amostra. Dois, que incluem características de difícil asserção, ”Gill-attachment”e ”Gill-spacing”.

O conjunto escolhido é composto dos atributos ”Population”com o valor ”v”(vasto), ”Veil-color”com o valor ”w”(branco) e ”Bruises”com o valor ”f”(falso). Esta entrada tem *lift* 1,98 e convicção de 11,33.

5 Conclusão

O estudo conclui que as três características sintetizantes buscadas são:

1. População: vasta
2. Cor do véu: branca
3. *Bruises*: falso (área não muda de cor ao ser esfregada com o dedo)

Isso representa um sucesso, com base no critério estabelecido no plano inicial, que conseguir um conjunto com poucos elementos que aparecesse no banco de dados, entre os cogumelos venenosos, com probabilidade maior ou igual a 50 por cento. Apesar disso, através da exploração realizada, descobriu-se que é amplamente conhecido o fato de que não é possível identificar com certeza cogumelos como comestíveis ou não, a não ser por meio de equipamento especializado e material de referência[7, 8, 9].

A recomendação continua sendo não comer cogumelos, mesmo em situações extremas. Isso é porque as variedades tóxicas são amplamente distribuídas por todo o Globo, e esse erro, provavelmente, causará a morte do sujeito.

Segue além do presente trabalho a tarefa de reunir bases de dados mais representativas do universo dos cogumelos, para que possa ser feita uma análise mais ampla. Pelas pesquisas que fiz (pouco abrangentes), não existe tal estudo.

Além deste, um outro ramo seria realizar, nessas bases mais extensas, classificação por redes neurais, passando primeiro pela camada de decisão do gênero a que o cogumelo pertence. Por exemplo, quase todo o gênero Lapiota parece não ser recomendado para ingestão. Isso tem o potencial de gerar classificadores satisfatoriamente precisos.

6 Apêndice

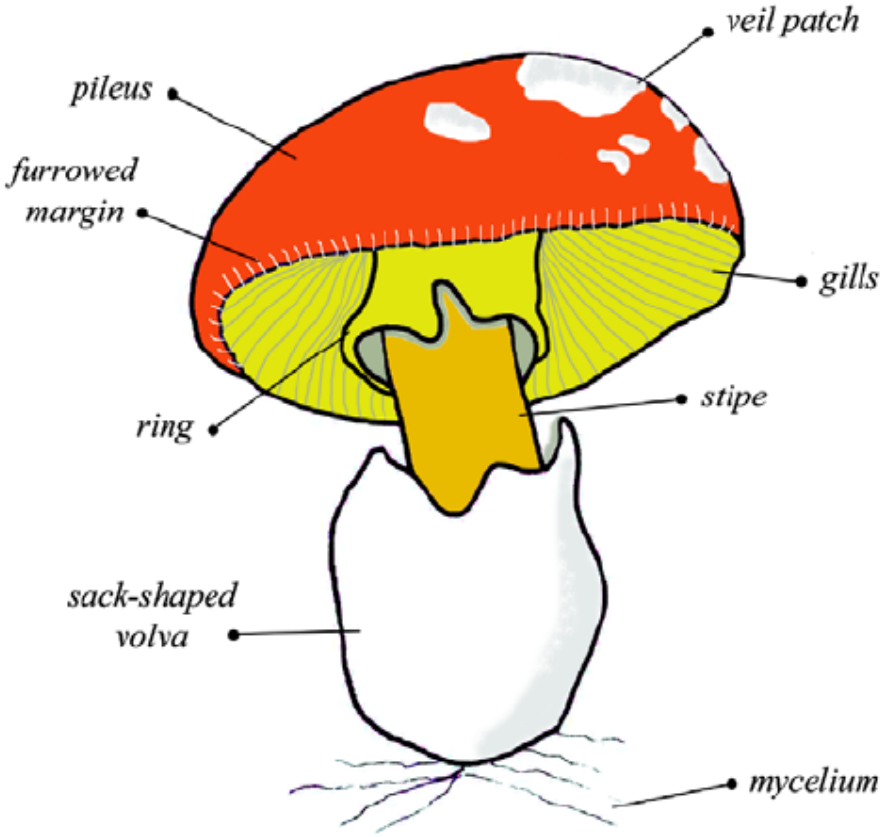

Figura 4: Esquema de um cogumelo[10]

7 Referências

1. https://github.com/Yowgf/MD-TP1
2. https://archive.ics.uci.edu/ml/datasets/mushroom
3. https://micolab.paginas.ufsc.br/files/2020/05/Drewinski-Menolli-Neves-2017-Agaricus-globocystidiatus.pdf
4. https://en.wikipedia.org/wiki/Lepiota
5. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing
6. http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
7. https://www.quora.com/How-can-I-tell-if-a-mushroom-is-poisonous-without-eating-it
8. https://www.quora.com/How-hard-is-it-to-identify-a-poisonous-mushroom
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342754/
10. https://www.researchgate.net/figure/Basidioma-of-a-gilled-mushroom-Amanita-caesarea-note-the-veil-patches-remnants-of_fig1_281625151