COC800 - Introdução à Ciência de Dados/ Primeira lista de exercícios
Felipe R. Oliveira

1) Há quem diga que muitos projetos de mineração de dados demandam mais tempo que o necessário e que os resultados obtidos não são animadores. Quais as possíveis explicações para tais afirmações?

R:

Antes de desenvolver uma resposta mais sistemática para essa pergunta cabe uma rápida contextualização. As técnicas de mineração de dados se popularizaram em indústrias de diversos ramos, mas isso não significa que necessariamente seus usuários possuem o conhecimento formal e/ou experiência para aplica-las. Dessa forma, expectativas irreais ou mesmo a não compreensão dos resultados, podem causar insatisfação.

A mineração de dados é um processo de várias etapas. Cada fase apresenta os próprios desafios conceituais e técnicos. Quando realizadas de maneira incorreta, essas etapas podem desencadear um gasto excessivo de tempo e/ou resultados insatisfatórios. Baseado nos trabalhos Shapiro *et al.* [1] e Munson [2] as etapas da mineração de dados, o tempo necessário pra executa-las e seus impactos nos resultados são:

i. **Compreensão das regras de negócios (22% do tempo total):** estabelecer o motivo pelo qual a mineração de dados está sendo empregada e qual objetivo final da tarefa. Essa etapa é determinante na duração das etapas seguintes e na qualidade dos resultados produzidos;
ii. **Coleta e preparação de dados (40% do tempo total):** dependendo da natureza do problema e/ou dos métodos de aquisição, coleta e tratamento de dados, essa etapa pode dispender muito tempo (frequentemente recursos financeiros).Além de ser o gargalo de todo processo, quando mal executada essa etapa produz dados de baixa qualidade (informações faltantes, imprecisões , incoerências, etc.) que, consequentemente , conduzem a resultados insatisfatórios. Vale ressaltar que essa etapa costuma ser a mais problemática;
iii. **Geração de modelos (19% do tempo total):** a criação de modelos costuma ser um processo relativamente rápido quando comparado com as etapas anteriores. A escolha equivocada de um modelo e/ou seus parâmetros podem gerar resultados insatisfatórios. Vale citar o sobreajuste (*overfitting*) como um exemplo clássico de problema provocado nessa etapa da mineração de dados;
iv. **Interpretação dos resultados (19% do tempo total):** esse é um processo altamente dependente do fator humano. Dessa forma, mesmo possuindo resultados de alta qualidade a intepretação pode variar de acordo com o conhecimento/experiência do analista.

2) O conceito "*big data*" traz à tona muitos desafios ao lidar com as características dos dados atuais. Discuta esses desafios trazendo suas reflexões para o ambiente da pesquisa científica.

R:

Baseado no trabalho de Amalina *et al.* [3] , que utiliza o princípio dos V's do *big data*, listam-se os seguintes desafios :

- **Volume:** o volume massivo de dados produzido de maneira ininterrupta exige uma infraestrutura (*hardware*) cada vez mais sofisticada, para o armazenamento e processamento da informação;
- **Variedade:** a diversificação e o aumento da complexidade dos dados coletados (textos, imagens, áudio, informação georreferenciada, por exemplo) demanda o uso de técnicas e tecnologias adequadas. Selecionar as ferramentas e/ou *softwares* apropriados é um desafio para analistas, devido as limitações de tempo e orçamento;
- **Velocidade:** a velocidade é crucial na transmissão de dados em tempo real. Dessa forma, o processamento ágil e em tempo real é um tópico desafiador considerando a infraestrutura atual de *big data*.
- **Veracidade:** um problema costumeiro na aquisição em tempo real de dados é a corrupção de informação. Portanto, a preparação de dados é necessária para eliminação de informações de baixa qualidade. A preparação de dados torna-se um desafio quando os dados são volumosos e/ou complexos;
- **Variabilidade:** a produção continua de informações não implica em um fluxo constante de dados. Picos no fluxo de informação costumam ser desafiadores para infraestrutura atual de *big data*.

3) Frequentemente o aprendizado de máquina é formulado como um problema de otimização. Explique esta formulação.
R:

É usual expressar o processo de otimização na forma de minimização da função objetivo, de acordo com a seguinte expressão genérica:

$$minimize: \qquad f(x_1, x_2, x_3 \dots x_d)$$
$$satisfazendo: \quad c_i(x) \leq 0 \;\; para \;\; i = 1,2,3 \dots N$$

onde x correspondem as variáveis de projeto, elas podem ser contínuas, discretas ou uma mistura de ambas. N é o número de restrições, expressas pelas funções de restrição c_i, elas podem ser restrições de igualdade ou desigualdades. É comum resolver este tipo de problema através de algoritmos iterativos determinísticos ou estocásticos.

De maneira genérica também é possível interpretar a etapa de ajuste de um modelo de aprendizado de máquina como um problema de otimização, cuja função objetivo é minimização entre erro da predição e o valor real. Praticamente todos os modelos de aprendizado de máquina possuem pelo menos uma etapa de otimização, geralmente solucionada através do método dos gradientes. Por exemplo a maximização (minimizar $-\boldsymbol{f(x)}$) da margem de separação em modelos de classificação SVM e a minimização da entropia cruzada em modelos de Regressão Logísticas.

Também é comum a otimização dos parâmetros que constituem o modelo a fim de maximizar algumas das suas métricas de avaliação (a taxa F1, por exemplo), processo conhecido como otimização de hiperparâmetros.

4) OLAP e mineração de dados envolvem manipulação de dados com objetivos distintos. Qual a recomendação de uso de cada um? Suponha que exista um data *warehouse* na COPPE com dados sobre os alunos. Exemplifique as operações comuns de OLAP e como elas poderiam ser úteis a um gestor acadêmico.
R:

A diferença básica entre o OLAP e o *data mining* está na maneira como a exploração dos dados é abordada. O OLAP é uma metodologia de verificação de dados, isto é, o analista conhece a questão, elabora uma hipótese e utiliza a ferramenta para confirmá-la. Com data mining, a questão é total ou parcialmente desconhecida e a ferramenta é utilizada para extração de informação.

A fim de simplificar a exemplificação vamos assumir o cubo OLAP hipotético ilustrado na Figura 1.

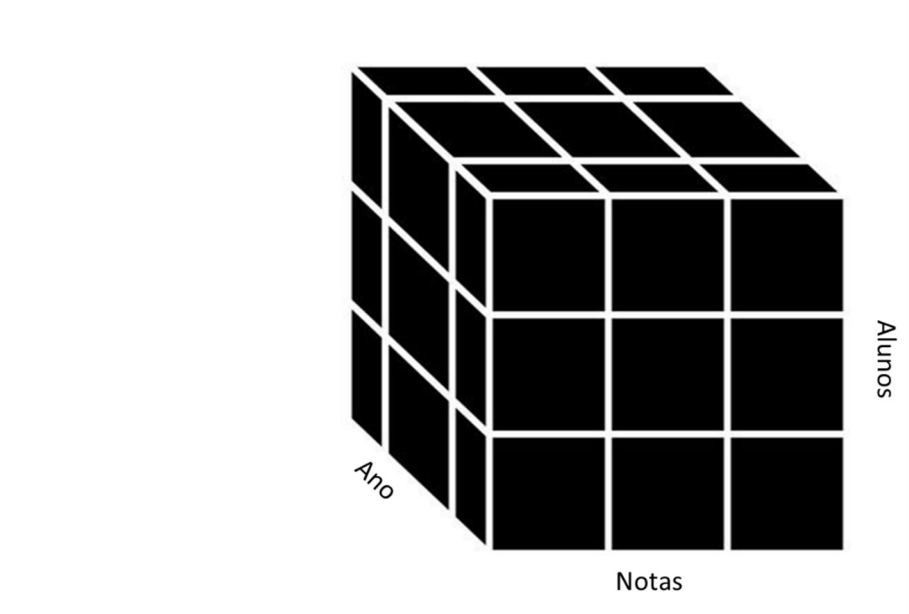

Figura 1: Cubo OLAP hipotético.

Algumas das operações que podem ser realizadas em um modelo OLAP e os exemplos de aplicações (baseados no trabalho de Domingues [4]) no cenário proposto, são:

- ***Slice*:** selecionar dados de uma única dimensão. Assumindo o cubo da Figura 1, um exemplo da aplicação do *slice* seria observar as toda notas de um mesmo aluno, como a Figura 2 ilustra;

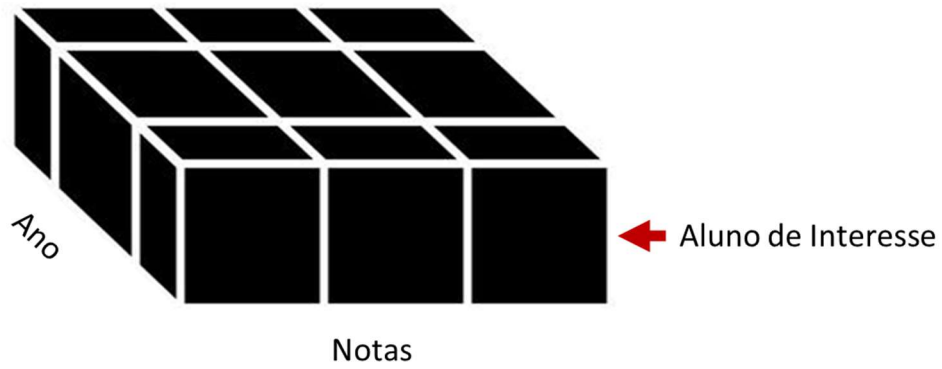

Figura 2: Exemplo de *slice*.

- ***Dice*:** extrair uma subparte do cubo de duas ou mais dimensões. Usando o exemplo anterior uma aplicação do *dice* sereia seleção de dois ou mais alunos em uma mesma consulta;

- ***Roll up*:** concatenação de células de uma ou mais dimensões para atingir um nível maior de generalização. A sumarização de todas as notas em um único valor (o CR no caso da COPPE) é um exemplo de aplicação do r*oll up*;
- ***Drill dow*:** examinar dados de forma refinada. Uma aplicação do *drill dow* sereia a consulta de alunos que possuem CR superior a 2.4 e são do gênero feminino;
- ***Pivot*:** visualizar dados por uma nova perspectiva. Partindo do princípio que o cubo OLAP apresentado na Figura 1 é direcionado à análise de desempenho dos alunos, um exemplo de aplicação do *pivot* seria a rotação do cubo direcionando a análise a avaliação das disciplinas.

5) Dados os classificadores no espaço ROC na Figura 3, descreva suas matrizes de confusão e calcule suas acurácias e precisões, sabendo que a base de dados possui XYZ registros, sendo XYZ a maior centena do DRE (ver exemplo abaixo) e que aproximadamente 40% dos registros são da classe positiva. Qual deles é o melhor? Justifique.

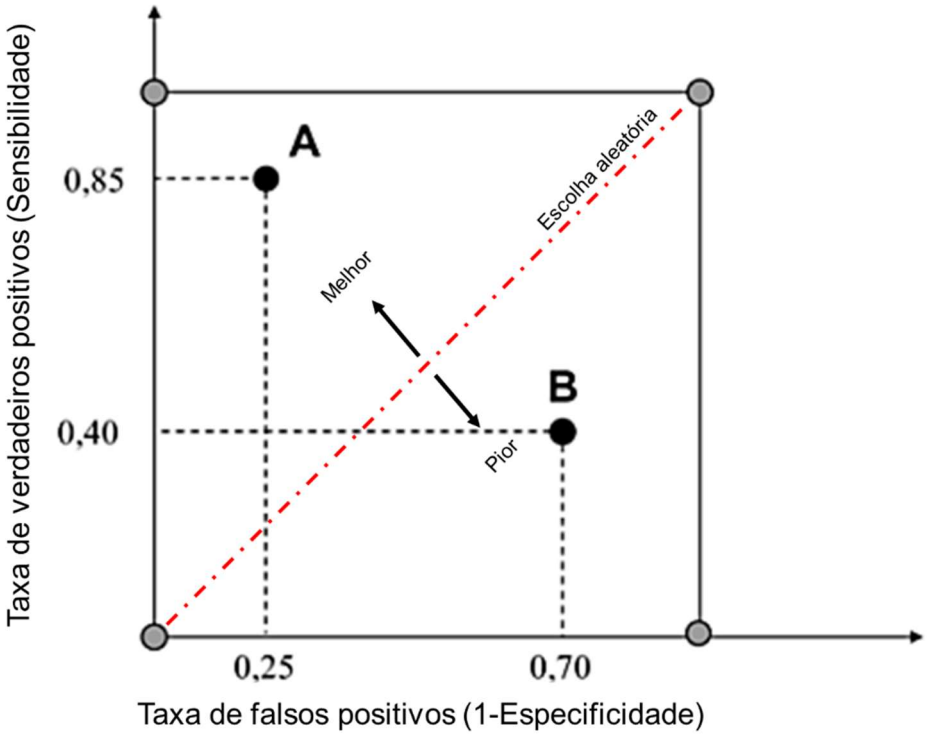

Figura 3: Espaço ROC da questão.

R-1:

$DRE = 120117488$; $N° \, de \, Registros = NR = 748$

$Registros \, Positivos \;\; = NP = 748 \cdot 0{,}4 \cong 300$

$Registros \, Negativos = NN = 748 - 300 = 448$

Classificador A:

$Sensibilidade = S = 0{,}85$

$Especificidade = E = 1 - 0{,}25 = 0{,}75$

$$Verdadeiros\ Positivos = VP = NP.S = 300\ .0{,}85 = 255$$

$$Falsos\ Negativos = FN = NP - VP = 300 - 255 = 45$$

$$Verdadeiros\ Negativos = VN = NN.E = 448\ .0{,}75 = 336$$

$$Falsos\ Positvos = FP = NN - VN = 448 - 336 = 112$$

$$Acurácia = AC = \frac{VP + VN}{NR} = \frac{255 + 336}{748} = 79{,}01\%$$

$$Precisão = P = \frac{VP}{VP + FP} = \frac{255}{255 + 112} = 69{,}48\%$$

Classificador B:

$$S = 0{,}40\ ;\ E = 1 - 0{,}70 = 0{,}30$$

$$VP = 300\ .0{,}40 = 120\ ;\ FN = 300 - 120 = 180$$

$$VN = 448\ .0{,}30 \cong 135\ ;\ FP = 448 - 135 = 313$$

$$AC = \frac{120+135}{748} = 34{,}09\%\ ;\ P = \frac{120}{120+313} = 27{,}71\%$$

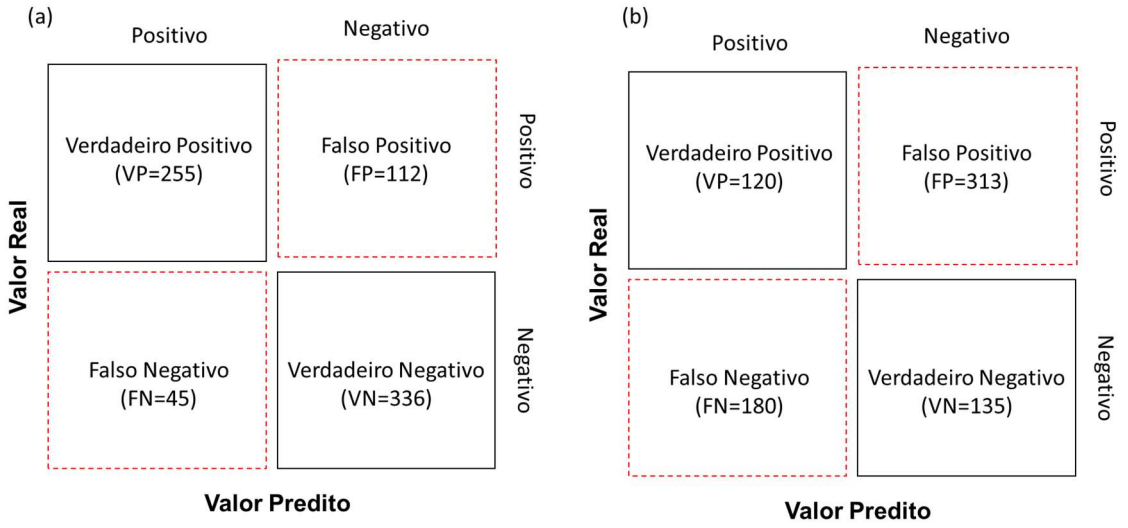

Figura 4 Matriz de confusão. (a) Classificador A. (b) Classificador B

R-2:

O Classificador A é melhor pois apresenta a maior precisão e acurácia, o que reflete em um menor número de erros de classificação. Vale ressaltar que o Classificador B não é só pior que o Classificador A, mas também é pior que um classificador randômico qualquer pois está abaixo da linha de desempenho de uma escolha aleatória.

6) A tabela abaixo apresenta os resultados obtidos nas etapas de dois experimentos realizados com o mesmo conjunto de dados original. Explique o funcionamento e a avaliação final do comportamento de cada um tendo por base os resultados. Que cenários podem justificar o uso de cada técnica de amostragem?

Tabela 1: Referência da questão 6.

Experimento A	Validação cruzada	Etapa	1	2	3	4
		Acurácia	0,95	0,90	0,85	0,80

Experimento B	Treino-teste-validação	Treino	Teste	Validação
		0,90	0,80	0,87

R:

O Experimento B utiliza a técnica de *holdout* , que consiste em separar o banco de dados em dados (**i**) treino, utilizados no ajuste do modelo; (**ii**) dados de teste, utilizados para avaliar uma, ou mais, métricas de desempenho do modelo (acurácia, no exemplo apresentado) ; (**iii**) dados de validação, utilizados para simular o desempenho do modelo em uma situação real em dados nunca antes observados. Vale ressaltar que no *holdout* a bipartição treino-teste é mais comum que a tripartição treino-teste-validação. O *holdout* é ideal para bancos de dados muito volumosos e /ou com dados muito complexos, que inviabilizam a aplicação de métodos interativos de ajuste e avaliação de modelo.

O Experimento A utiliza a Validação Cruzada (CV) *4-fold,* para avaliar a capacidade de generalização do modelo adotado. Através da CV é possível avaliar a média e o desvio padrão de uma, ou mais, métricas do modelo (acurácia 87.5% ± 6.5%, no exemplo apresentado) em *k* partições de treino-teste, como a Figura 5 ilustra. A CV é ideal para bancos de dados mais compactos e/ou quando é desejado conhecer a sensibilidade do modelo em diferentes cenários.

Figura 5 : Validação Cruzada.

Os resultados da acurácia média do Experimento A e a acurácia nos dados de validação do Experimento B são extremamente próximos, sugerindo o desempenho equivalente entre os modelos adotados, diferindo um pouco nos dados de teste, provavelmente, devido a uma partição desfavorável. Vale ressaltar que a Validação Cruzada e o *holdout* não são mutuamente excludentes, ou seja, as técnicas podem ser combinadas. Contudo, alguns especialistas julgam pouco interessante aplicar a CV em uma subparte do banco de dados ao invés de sua totalidade.

7) Geralmente não se possui dados e conhecimentos suficientes para se produzir um classificador perfeito. Assim, as capacidades de generalização e de representação de um modelo são características importantes durante o seu aprendizado, mas nem sempre é possível obter máximos desempenhos simultaneamente. Há uma percepção de que em aprendizado de máquina, em muitas situações, é melhor produzir um modelo mais simples do que um mais complexo para se resolver um problema. Expanda essas afirmações trazendo mais elementos (conceitos, exemplos etc.).

R:

Modelos, sejam eles de aprendizado de máquina ou mesmo de elementos finitos, são simplificações de fenômenos reais. O desempenho de um modelo está relacionado, entre outras coisas, com a sofisticação das hipóteses que o baseiam. Contudo, mesmo o modelo mais bem fundamentado carrega suas imprecisões, devido à complexidade de representar perfeitamente fenômenos reais (particularmente os naturais). Dessa forma, quanto mais complexo é o evento, maior é o custo (humano e computacional) para modela-lo e mais improvável torna-se o desempenho simultâneo máximo em todas as suas métricas de avaliação [5].

O dilema da escolha de um modelo dada sua complexidade, está longe de ser exclusivo de problemas de aprendizado de máquina. Um exemplo clássico da engenharia civil é a modelagem de vigas. Existem inúmeras formas de modelar uma viga, de uma simples barra unidimensional até o sofisticado Método dos Elementos Finitos (MEF), cada uma adequada para uma situação diferente. Via de regra, modelos complexos tendem a ser mais representativos (entendendo por representatividade como a capacidade de descrever o fenômeno real), menos generalistas (entendendo por generalização a capacidade de reprodutibilidade em situações variadas) e custosos computacionalmente (economicamente) [6].Em Problemas corriqueiros , que toleram certos graus de imprecisão, o custo associado ao desenvolvimento de modelos complexos muitas vezes não trazem um retorno financeiro que justifique sua utilização.

Dessa maneira antes escolher quão sofisticado será o modelo adotado é necessário primeiro determinar a complexidade problema do evolvido e a familiaridade do analista com o mesmo. Existem inúmeros trabalhos que comparam o desempenho entre classificadores simples e sofisticados. Vale citar o trabalho de Byvatov [7], que comparou um modelo SVM e uma Rede Neural Artificial (ANN) , em um problema de classificação de presença de drogas em amostras de sangue. Mesmo sendo um modelo mais complexo a ANN alcançou um desempenho inferior ao SVM.

8) Fale sobre que cenários são favoráveis à adoção do *Spark* ou à do *Hadoop*. Em sua linha de pesquisa, encontre estudos com essas tecnologias e descreva os benefícios e limitações alcançados. Além disso, caso existam, verifique se há alternativas a eles e aponte seus diferenciais.

R:

Hadoop *Distributed File System* (HDFS), é um ambiente projetado para o armazenamento (*storage*) distribuído de grandes volumes de dados em *hardwares* comuns. O *Spark*, por outro lado, é uma plataforma de processamento de *Big Data* projetada para o ganho de desempenho de velocidade de uso e análise, dentro do ambiente

Hadoop. O *Spark* oferece recursos nas principais linguagens de programação (*Python*, por exemplo) e um conjunto de bibliotecas que permite a integração entre SQL, processamento em tempo real (*stream*) e análises complexas.

Na linha de pesquisa de Tomada de Decisão, e particularmente na minha tese que envolve a aplicação de NLP em tweets, os recursos de programação e a capacidade de análise em tempo real do *Spark* são grandes atrativos. Behera [8] , uma das referências da minha tese, pontua como características positivas do *Spark* o alto desempenho em *streaming* e a facilidade de implementação (quando comparado ao HDFS) , e como característica negativa a instabilidade no *pipeline* de aprendizado de máquina .

Um exemplo recente de aplicação do *Haddop*, dentro da minha temática de pesquisa, é o trabalho de Jenhani *et al.* [9] .Nesse artigo os autores pontuam como principal ponto positivo do HDFS sua alta capacidade de armazenamento distribuído em *hardwares* de baixo custo. Como pontos negativos, o HDFS apresenta um tempo de implementação elevado e uma curva de aprendizado lenta para novos usurários.

Existem inúmeras alternativas tanto de *storage* quando *stream* além das tecnologias já citadas que poderiam ser adotadas. Um número tão grande que tornaria discussão e comparação um tópico excessivamente extenso, considerando o que na minha linha de pesquisa o *Haddop* e o *Spark* já são as escolhas mais interessantes. Vale citar as seguintes alternativas de s*torage* : *Databricks*, *Google BigQuery*, *Cloudera*, *Snowflake* e o *Qubole*. Também vale citar seguintes alternativas de *stream*: *Storm*, *Kafka*, *Google Cloud Dataflow* e o *Amazon Kinesis*

9) Qual a importância da "engenharia de atributos" no processo de aprendizado? Na metodologia CRISP-DM está prevista essa atividade? Justifique. Exemplifique as etapas do CRISP-DM em um problema de sua área de atuação/pesquisa/interesse.

R:

Para He *et al.* [10], os fundamentos da engenharia de atributos são (**i**) a seleção (**ii**) e transformação de dados, e (**iii**) a agregação de informação. Como discutido na questão 1, a qualidade dos dados utilizados influencia diretamente na qualidade dos resultados obtidos pelos sistemas de aprendizado. Dessa forma, além de ajudar na compreensão do analista, a engenharia de dados costuma melhorar o desempenho de modelos de aprendizado de máquina.

Para Huber [11], dentre as fases da metodologia CRISP-DM a engenharia de dados é a mais extensa ,equivalendo as etapas de entendimento e preparação de dados. As etapas do CRISP-DM e os exemplos de aplicações na pesquisa da minha tese são:

i. **Entendimento do negócio:** no âmbito acadêmico essa etapa equivale a definição do objeto de estudo, motivações e objetivos do trabalho. Até o presente momento só o objeto de estudo foi definido, tweets (texto e geolocalização);
ii. **Entendimento dos dados:** no estudo pretendido essa etapa equivale ao desenvolvimento do *pipeline* de aquisição de dados, exploração parcial das informações (a exploração profunda do corpus requer a preparação dos dados);
iii. **Preparação dos dados:** equivale ao tratamento de informações faltantes ou incorretas, normalização das informações georreferenciadas e a transformação do

conjunto de textos (remoção de pontos, remoção de acentos, conversão de emojis em texto, stemização, vetorização etc.);

iv. **Geração de modelo:** equivale a etapa futura de particionamento dos dados (treino-teste), seleção de um modelo de classificação e ajuste do modelo;

v. **Avaliação:** equivale a etapa futura de avaliar o desempenho do modelo escolhido e se necessário reexecutar as etapas anteriores;

vi. **Implementação:** uma vez definido o modelo e suas limitações, é pretendido a implantação de um sistema de classificação georreferenciado em tempo real.

10) Alguns projetos de aprendizado de máquina não atingem o sucesso esperado. Que fatores podem explicar esse fenômeno? Critique um artigo publicado em revista científica em que, apesar de um eventual sucesso promovido pelo texto, podem ter sofrido dos problemas apontados.

R:

Apesar de *data mining* e *machine learning* serem metodologias distintas (não mutuamente excludentes), a primeira focada na extração de informação e a segunda focada na aplicação prática e automatizada da informação, suas etapas e seus eventuais problemas são extremamente semelhantes. Dessa forma, a resposta da Questão 1 também é válida para essa questão.

O *machine learning,* entretanto, é muito menos dependente do fator humano que o *data mining*. Dessa forma, mesmo sendo mais ágil, o aprendizado de máquina está sujeito a cometer falhas (racismos e machismo, por exemplo) que um ser humano dificilmente cometeria. Existem inúmeros casos de, falhas causados pela "desumanização" do processo de aprendizado (como sugestão https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020).

Erros em uma ou mais etapas do processo de aprendizado de máquina são comuns (muitos até toleráveis), mas raramente todos são cometidos ao mesmo tempo como no projeto *Watson for Oncology*,da IBM. Por etapa os erros cometidos foram:

i. **Compreensão das regras de negócio:** o *Watson for Oncology* tinha como objetivo a erradicação do câncer através da identificação prematura em pacientes. Além de ser pouco realista, essa meta não delimita um objeto de estudo claro pois existem inúmeros tipos de câncer;

ii. **Coleta e preparação de dados:** o número de dados coletados era muito pequeno para os objetivos pretendidos. Os bancos de dados continham mais informações de pacientes hipotéticos (utilizados no ensino de medicina) que pacientes reais. Devido à falta de especialistas envolvidos no projeto, muitas recomendações absurdas de tratamento não foram filtradas;

iii. **Geração de modelos:** graças aos equívocos da etapa anterior, os modelos gerados simplesmente não possuem validade;

iv. **Intepretação dos resultados:** provavelmente o erro mais grave cometido. O sistema de recomendação de tratamentos produziu resultados perigosamente incorretos (consequência dos erros da etapa **ii**). Devido à falta de especialistas para identificar tais erros, o projeto foi lançado de forma precipitada.

Particularmente, optei por pontuar os erros de uma empresa privada, pois julgo que os cientistas que utilizaram o *Watson for Oncology* foram vítimas do projeto. Muitos artigos publicados foram totalmente ou parcialmente invalidados pelos problemas metodológicos da IBM. Por exemplo, o trabalho de Somashekhar *et al.* [12] , publicado 4 meses antes da descoberta das falhas do projeto

Referências

[1] G. Piatetsky-Shapiro, R. Brachman, T. Khabaza, W. Kloesgen, e E. Simoudis, "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications", *Proc. Second Int. Conf. Knowl. Discov. Data Min.*, p. 89–95, 1996.

[2] M. A. Munson, "A study on the importance of and time spent on different modeling steps", *ACM SIGKDD Explor. Newsl.*, vol. 13, nº 2, p. 65–71, 2012.

[3] F. Amalina *et al.*, "Blending Big Data Analytics: Review on Challenges and a Recent Study", *IEEE Access*, vol. 8, p. 3629–3645, 2020.

[4] V. Domingues, "Desenvolvimento de sistema OLAP para análise de informação de gestão académica da UC", Universidade de Coimbra, 2014.

[5] R. Nisbet, J. Elder, e G. Miner, "Model Complexity (and How Ensembles Help)", in *Handbook of Statistical Analysis and Data Mining Applications*, First., Amasterdan: Elsevier Inc., 2009.

[6] A. Mujeeb, W. Dai, M. Erdt, e A. Sourin, "One class based feature learning approach for defect detection using deep autoencoders", *Adv. Eng. Informatics*, vol. 42, nº February, p. 100933, 2019.

[7] E. Byvatov, U. Fechner, J. Sadowski, e G. Schneider, "Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification", *J. Chem. Inf. Comput. Sci.*, vol. 43, nº 6, p. 1882–1889, 2003.

[8] S. Das, R. K. Behera, M. Kumar, e S. K. Rath, "Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction", *Procedia Comput. Sci.*, vol. 132, nº Iccids, p. 956–964, 2018.

[9] F. Jenhani, M. S. Gouider, e L. Ben Said, "Streaming social media data analysis for events extraction and warehousing using hadoop and storm: Drug abuse case study", *Procedia Comput. Sci.*, vol. 159, p. 1459–1467, 2019.

[10] D. Shah, J. Wang, e Q. P. He, "Feature engineering in big data analytics for IoT-enabled smart manufacturing – Comparison between deep learning and statistical learning", *Comput. Chem. Eng.*, vol. 141, p. 106970, 2020.

[11] S. Huber, H. Wiemer, D. Schneider, e S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model", *Procedia CIRP*, vol. 79, p. 403–408, 2019.

[12] S. P. Somashekhar *et al.*, "Watson for Oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board", *Ann. Oncol.*, vol. 29, nº 2, p. 418–423, 2018.