Prova 1

Aluno: Kamila de Almeida Benevides
Professor: Bruno Pimentel

1) Defina Ciência de Dados.

É uma área interdisciplinar como Big dada, estatística, aprendizagem de máquina, economia, mineração de dados, etc. É uma área voltada para o estudo e a análise de dados, busca extração de conhecimento e Insights para possíveis tomadas de decisão. Ciência de Dados lida com criação de soluções para modelagem de dados;

1) Quais desafios os cientistas de dados possuem ao trabalhar com Big Data?

Dados além de muito importantes, muitas vezes são muito complexos, com isso, o cientista de dados deve ter uma análise bem criteriosa de ferramentas e algoritmos para tratar esses dados. Um desafio do cientista de dados é utilizar um algoritmo levando em consideração a escalabilidade de tempo em relação ao tamanho do conjunto de dados, que sejam eficientes na remoção de ruídos, mesmo que em alguns casos tenham tantos dados que o algoritmo não é capaz de tratá-los por completo, além de observar como o algoritmo se comporta em ambientes distribuídos.

3) De que forma a Estatística pode auxiliar Ciência de Dados?

A estatística é importante para políticas públicas, dados demográficos e econômicos, se dedica à coleta, análise e interpretação de dados. Dessa forma, a estatística é uma ferramenta da área de ciência de dados para analisar os dados coletados para tomada de decisões. Sendo assim, a estatística é uma ferramenta de recolha, organização, resumo, apresentação, interpretação e conclusões dos dados.

4) Mostre a importância de pré-processar os dados para a extração de informação.

É importante tratar os dados antes de analisá-los, mas nem sempre os dados estão bem estruturados e organizados, eles podem conter ruídos, outliers, elementos faltantes e dados inconsistentes. Pré-processando os dados aumenta sua qualidade e melhora os resultados da mineração, melhorando assim a acurácia do modelo, fazendo com que os dados se tornem abrangentes, verifica se aqueles dados realmente estão representando de fato a realidade, confiabilidade dos dados que está associada a limpeza dos dados e a interpretabilidade dos dados.

5) Indique a importância de utilizar métodos de avaliação de modelos no processo de extração de informação.

É importante para descobrir se a complexidade do modelo está adequada com a base de dados para identificar overfitting e underfitting.

6) Com respeito à Análise de Agrupamento, cite um exemplo de aplicação onde é preferível utilizar agrupamento hierárquico.

Quando se espera criar grupos com subgrupos, onde é mais interessante formar hierarquias com os subgrupos. Agrupamento hierárquico é muito utilizado para bioinformática, em análises onde é necessário criar conjuntos de subgrupos para análises de um determinado comportamento dos dados analisados, por exemplo, análise de genes para entender o comportamento de câncer, tendo um conjunto de proteínas para cada amostra, onde pode-se agrupar as proteínas de modo a buscar quais similaridades existem entre as amostras.

7) Comparando os métodos K-Vizinhos Mais Próximos e K-Means, mostre em quais aspectos eles têm semelhanças e diferenças.

As semelhanças são que ambos utilizam métodos de distância para agrupar ou classificar, assim pode ser aplicado vários métodos para escolher essa distância que será usada. As principais diferenças é que K-Means é um algoritmo agrupamento e não supervisionada, já o K-Vizinho mais próximo é um algoritmo de classificação e supervisionado.