

## **PARTE 1: Sobre a profissão e o universo de dados**

### **1. Descreva com suas palavras qual a diferença entre a profissão de cientista, analista e engenheiro de dados? Que tipo de tarefa você imagina que cada um desses faça?**

O engenheiro de dados é o profissional que prepara o terreno, digamos assim. O trabalho dele é parecido com o de um engenheiro de software, pois ele é responsável pelo desenvolvimento de plataformas e arquiteturas para o processamento dos dados. É sua responsabilidade lidar com a arquitetura de pipeline, testes, construir pipelines tolerantes a falhas e estáveis, além de administrar bancos de dados.

O analista de dados é o profissional que cuida da análise em si. Ele está preocupado em tirar informações valiosas dos dados, através da limpeza dos dados, do tratamento, da visualização, além de realizar a modelagem dos dados. Também precisa ter habilidades de story telling, pois precisa apresentar suas análises. O seu trabalho é realizar análises para que assim possa tomar decisões melhores sobre o negócio, e em pouco tempo.

O cientista de dados é, antes de tudo, um analista de dados, pois parte do seu trabalho é fazer análises em cima dos dados disponíveis. O diferencial do cientista de dados está relacionado com a união de vários conhecimentos, passando por matemática, estatística e programação. O seu principal trunfo é a capacidade de gerar insights sobre eventos futuros, a partir dos dados existentes no momento, utilizando algoritmos de Machine Learning.

Em resumo, o engenheiro é responsável por desenvolver uma plataforma para que tanto o analista quanto o cientista de dados possam trabalhar em cima. O analista é responsável por tomar decisões que afetam a situação atual da empresa. Por fim, o cientista é responsável por gerar insights sobre o futuro do negócio, baseado nos dados disponíveis.

### **2. Por que um cientista de dados é, apesar de tudo, uma pessoa de negócios?**

Uma das áreas de conhecimento de um cientista de dados está relacionada à área em que ele atua, seja área médica, e-commerce, financeira, etc. O seu conhecimento sobre o negócio da empresa é tão importante quanto os outros conhecimentos, pois o seu trabalho deve estar orientado ao sucesso do negócio. Ele

pode criar o melhor modelo de ML do mundo, mas se isso não fizer diferença para o negócio, não tem utilidade. O cientista de dados deve sempre se perguntar: como isso afeta o negócio?

**3. O que é Machine Learning? Que tipo de trabalho de Machine Learning poderia ser feito na área em que você atua? Mencione alguns estudos/modelos que te auxiliariam no dia a dia.**

Quando criamos algoritmos, ou mais especificamente modelos, que podem melhorar com experiência, utilizando dados, estamos falando de Machine Learning. A ideia desses algoritmos é aprender com dados existentes, para depois realizar previsões a partir de dados novos. Pode ser usado para reconhecer rostos de humanos, se um email é spam, ou até mesmo o resultado de um jogo.

Na minha área de interesse, a de educação, podemos encontrar ML sendo aplicada para realizar uma recomendação de aulas personalizada, ajustar a velocidade do aprendizado, trazer relatórios a respeito do desempenho do aluno, entre outros.

Um bom exemplo de ferramenta que faz uso de ML, e que me ajuda nos meus estudos, é o Quizlet, que possui um assistente de aprendizado responsável por criar um plano de estudo individualizado, acompanhamento do progresso do aluno, além de sessões de estudo próprias para o aluno.

**4. Qual a utilidade da estatística para alguém que trabalha com data science?**

A estatística serve para descobrir o que tem por trás dos dados, obter sentido deles, encontrar uma estrutura. Como exemplo disso, temos medidas como média, mediana, moda, desvio padrão e distribuição, que são úteis para o Cientista de Dados entender o comportamento das variáveis com as quais está trabalhando, além de poder identificar problemas. Além disso, a estatística fornece o conhecimento para que se possa escolher os melhores modelos, saber como e porquê funcionam, de forma que sejam melhor aproveitados.

**PARTE 2: Conhecendo o Python e o Jupyter Notebook**

**5. Demonstre como realizar no Python as seguintes operações: soma, subtração, multiplicação, divisão, exponenciação e raiz quadrada.**

6. Qual operador utilizamos para obter o resto de uma divisão no Python?
7. Como verificar qual o tipo de algum objeto no Python?
8. Qual a diferença entre o objeto `x = 2`, `y = 2.0` e `z = '2.0'`?
9. O que acontece se escrevermos `2++2` ao invés de `2+2`? E se fizermos `2 - - 2`? E `2==2` ou `2 2`? Verifique o output de cada situação.
10. É possível escrever `02` ao invés de apenas `2` no Python?
11. Crie uma variável chamada *bônus* (sem acento mesmo) que possui valor `50000`. Digamos que esse é o registro do bônus dos trabalhadores da sua firma. A inflação no ano foi de `11%`. Depois de criar a variável *bônus* em uma célula, altere seu valor na célula seguinte para que acompanhe a inflação. I.e., ajuste o valor da variável para que ele seja `11%` acima de `50000` (`1,11 x 50000`).

### PARTE 3: Questões éticas

12. Cite alguns problemas que a não-proteção dos dados pode trazer para as pessoas.

O vazamento de dados é um dos grandes problemas da não-proteção dos dados, principalmente quando não são apenas informações pessoais, mas o vazamento de dados bancários, podendo acarretar em transações financeiras ilegais, trazendo grandes prejuízos a curto e a longo prazo.

A total perda de dados também faz parte dos problemas enfrentados por quem não protege seus dados. Exemplos não faltam de casos em que um invasor obtém acesso aos dados da empresa/pessoa e simplesmente criptografa tudo, para depois pedir o resgate, muitas vezes sem sucesso.

13. Como o Machine Learning pode piorar a situação das pessoas?

O mau uso do ML está atrelado, muitas vezes, ao objetivo da aplicação da ferramenta, como é o caso de Deepfakes, que podem não só espalhar notícias falsas, mas prejudicar a reputação de pessoas que são alvos destas aplicações. Outro caso em que a aplicação de ML pode ser prejudicial está relacionado ao uso de reconhecimento facial, com diversos casos em que pessoas inocentes foram presas injustamente por conta desta ferramenta, além de outros problemas atrelados ao reconhecimento facial. Um outro caso que ficou bastante conhecido foi o da IA da

Amazon que rejeitou currículos de mulheres, causado por vieses humanos que o algoritmo aprendeu.

**14. Cite um exemplo de risco do mau uso de Machine Learning na sua área de atuação ou formação.**

Um exemplo que é muito conhecido no momento é o Deepfake, usado para manipulação de vídeos ou outros tipos de representações. O mau uso desta ferramenta pode trazer inúmeros problemas, como ao espalhar notícias falsas, principalmente para as pessoas que não estão familiarizadas com este tipo de tecnologia e, quem sabe no futuro, sendo difícil de reconhecer se é verdadeiro ou não até mesmo para aqueles que fazem parte do meio tecnológico.

**15. Que tipos de problemas o reconhecimento facial vem causando a certas minorias?**

Os problemas causados pelo reconhecimento facial estão principalmente relacionados à discriminação racial. Como exemplo, temos o relatório disponibilizado pelo governo federal dos Estados Unidos, em que os algoritmos de reconhecimento facial utilizados apresentam bons resultados para homens brancos de meia idade, mas resultados ruins para pessoas negras, idosas, mulheres e crianças. Entre os problemas causados por este tipo de falha, estão prisões de pessoas inocentes, aumento de pena e até mesmo morte por violência policial.