Questão 01

Utilizando suas palavras, defina Big Data e seus "4V's".

Big Data, além do que o próprio nome diz, é como grande motor estratégico de análises. Com toda a demanda existente e a necessidade de guardar dados, lidar com o grande volume é essencial para o sucesso das organizações.

Os 4V's do Big Data são: **Volume, Velocidade, Veracidade e Variedade.**

- **Volume** – Se refere a quantidade de dados que é gerado. O negócio precisa estar preparado para a quantidade esperada de informação que será armazenada.
- **Velocidade** – Se refere a o quão rápido o dado é coletado, armazenado e analisado. Dependendo do problema a velocidade é o divisor de águas para a tomada de decisão e predições rápidas.
- **Veracidade** – Se refere ao quão confiável o dado é. Toda a análise deve ser pautada em fontes confiáveis.
- **Variedade** – Dificilmente os dados virão de uma única fonte. Aqui os dados são recebidos de origens variadas e podem estar estruturados ou não estruturados.

Questão 02

A utilização de Docker no contexto de Big Data pode ser feita dedivers as maneiras a fim de ajudar a lidar com o grande volume de dados. Cite e explique resumidamente pelo menos um caso de uso de Docker e Big Data.

Questão 03

Quais pontos de abstração você julga importante para compor um projeto de Big Data?

Questão 04

Explique a diferença entre aprendizado supervisionado, não supervisionado e semi-supervisionado.

Quando trabalhamos com um 'alvo' ou uma resposta que queremos ter dado características de dados passados de um grupo ou população estamos falando de um aprendizado supervisionado, ou seja, há o fornecimento de informações e rótulos para que o algoritmo possa aprender a aplicar em dados nunca vistos.

Quando não temos essa respostas, ou alvo, e queremos descobrir relações existentes em um grupo de indivíduos usamos a abordagem não supervisionada. Geralmente problema de regressão e classificação são resolvidos com abordagens supervisionadas enquanto métodos de agrupamento ou redução de dimensionalidade são problemas não supervisionados.

A abordagem semi-supervisionada se refere a necessidade de atuação supervisionada quando se tem poucos dados rotulados e muitos dados não rotulados. Um exemplo é a clusterização que a Google faz no aplicativo Google Fotos e de tempos em tempos pede para o usuário rotular o grupo de novos rostos parecidos. É um problema misto, entre supervisionado e não supervisionado.

Questão 05

Explique a diferença entre Inteligência Artificial, Aprendizado de Máquina e Aprendizado Profundo, além disso dê exemplos de algoritmos para cada.

Inteligência artificial engloba Machine Learning que por sua vez engloba Deep learning.

- IA – É o campo de estudo que busca entender como é possível um computador pensar como um humano e resolver problemas complexos.
- Machine learning – É a aplicação da IA que permite a criação de algoritmos que permite sistemas aprenderem com ou sem a intervenção humana.
    - Problemas de Classificação/Regressão/Clusterização e etc são usados nessa abordagem. Exemplos de algoritmos são: Naive Bayes, Decision Trees, Kmeans, PCA.
- Deep learning – Tem foco na criação de sistemas que imitam uma rede neural humana. Aqui há a capacidade da máquina treinar a si mesma para aprendizado mais completo e ao longo do tempo ou input de usuários.
    - Perceptron é um dos tipos mais simples de rede neural
    - RNN – Rede neural recorrente (usada na google para completar o que o usuário pode estar querendo pesquisar)
    - LSTM usado em reconhecimento de voz e assistentes de voz.

Questão 06:

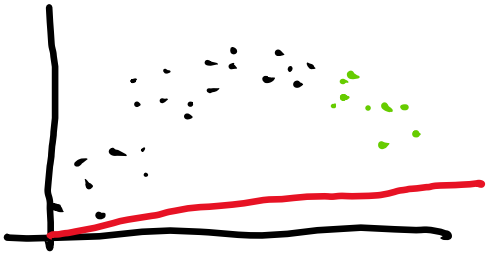
Explique o que é Overfitting e Underfitting.

Se refere ao quão bem ou mal as previsões estão ajustadas aos dados.

Overfitting se refere ao ao sobreajuste, ocorre quando o modelo "aprendeu demais" sobre as características das previsões, performa muito bem nos dados de treino porém tem um desempenho péssimo em dados nunca visto. É um modelo que perde a capacidade de generalização. O desenho abaixo ilustra a situação:

- A linha vermelha se refere ao ajuste das predições ao valor real.
- Os pontos verdes representam novos dados e os pontos pretos representam os dados de treino.
- Está ligados com modelos complexos demais

O Underfitting se refere ao modelo que ainda nos dados de treino não conseguiu aprender as relações para fornecer previsões com assertividade



- Está ligado a modelos com baixa complexidade, poucas features. É um modelo que ‘generaliza demais’