**Questão 02 -** ID3, seleciona o melhor atributo sobre o conceito de entropia para o desenvolvimento da árvore. Já o C4.5, segue o mesmo caminho do anterior, mas, com alguns melhoramentos. Como, o uso de dados perdidos ou desconhecidos, o uso de atributos com diferentes pesos, e a acurácia do algoritmo C4.5, é ligeiramente melhor do que o algoritmo ID3.

**Questão 03-** Razão de ganho é a proporção de informação gerada pela partição útil, ou, que aparenta ser útil para a classificação.

**Questão 04-** O aprendizado supervisionado consiste em, pares de entrada-saída, com essa forma, aprende-se mapeando as entradas e saídas, inferindo a função a partir de dados rotulados. Já o não-supervisionado, detecta padrões a partir de dados que não estão marcados, ou, rotulado.

**Questão 05 -** São quatro os tipos de problemas, são eles: classificação, regressão, agrupamento (clustering) e regras de associação.
**Classificação:** Utilizado para prever ou descrever uma classe, o seu atributo é nominal, exemplo, se quero saber se irei assistir 'Star Wars' ou não, vai depender se será as trilogias e qual delas, se será na sequência, o local em que irei assistir.
**Regressão:** É semelhante a **Classificação**, a diferença é que, o seu atributo é numérico. Exemplo: O preço da assinatura do serviço streaming para assistir a franquia 'Star Wars'.
**Agrupamento (clustering):** Agrupa-se as instâncias de acordo com os atributos de entrada. Exemplo: Identificar o perfil de usuário que assistem ''Star Wars''.
**Regras de Associação:** Usada para, buscar semelhança e/ou associações entre os elementos. Exemplo: Quem assistiu a trilogia original de 'Star Wars', também assistiu ‘The Mandalorian’.

**Questão 06 -** O algoritmo escolhe o melhor atributo baseado no conceito de entropia e ganho de informação, desenvolvendo assim a árvore. O que é uma boa escolha para tratar uma quantidade enorme de dados, por exemplo, dados numéricos na internet como, CPF(social secutrity).

**Questão 07-**

**A) Cross-validaton:** Usada para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. É amplamente empregada onde, o objetivo da modelagem é a predição, técnica usada para estimar a precisão do modelo na prática.

**B)** Consiste em dividir o conjunto total, geralmente na proporção 2/3 do dados para o treinamento e 1/3 dos dados para o teste. É uma abordagem indicada quando tem disponível uma grande quantidade de dados, caso o conjunto seja pequeno, o erro calculado pode sofrer muita variação.

**Questão 08 -** O problema do do desbalanceamento de classes é que, o algoritmo se torna tendencioso a favorecer a classe majoritária. Para poder (tentar) resolver esse problema, temos que tratar certos aspectos, como, a redefinição do tamanho do conjunto de dados, para diferentes classes usar diferentes custos de classificação e a classe precisa ter um modelo.

Pode-se adicionar instâncias a classe minoritária, chamado de over-sampling, ou, remover da classe majoritária, chamado de under-sampling. Ambos, assim como outras técnicas, possui seus problemas, no caso do over-sampling, o que for adicionado, instâncias, podem não ocorrer em situações reais nesse caso, pode resultar em um modelo inadequado o overfitting é uma possibilidade, fazendo com que o modelo seja superajustado aos dados do treinamento.

Já no under-sampling, dados com grande relevância, podem ser perdidos no processo de retirada dos dados levando a um outro problema, o underfitting que, o modelo escolhido não se ajuste aos dados do treinamento.