Lista Arvore e Random Forest

1) **A tabela CREDITO contém informações sobre pagamentos de empréstimos concedidos pelo banco IFM. As variáveis da tabela são:**

| Nome | Tipo de variável | Nível de mensuração | Descrição |
|---|---|---|---|
| Id_client | ID | Nominal | Identificação |
| Idade | Input | Intervalar | Idade |
| Est_civ | Input | Nominal | Estado Civil (1=casado 2=solteiro 3=divorciado 4=viúvo 5=não inf. ) |
| Sexo | Input | Binária | Sexo (0=M 1=F) |
| qtd_com | Input | Binária | Primeira Aquisição (sim/Não) |
| tempo_e | Input | Intervalar | Tempo no emprego atual em meses |
| sal_cli | Input | Intervalar | Salário do Cliente |
| qtd_parc | Input | Intervalar | Qtd de Parcelas a serem quitadas |
| vlr_cpr | Input | Intervalar | Valor total do empréstimo |
| vlr_prt | Input | Intervalar | Valor da Parcela |
| tipo_cre | Input | Binária | Tipo de Crédito - Carnê ou Débito em Conta (0=Carnê 1=Débito) |
| Sal_conj | Input | Binária | Cônjuge tem salário |
| Tipo | Target | Binária | Tipo de Cliente (Adimplente ou Inadimplente) |

a) Construa a árvore de classificação considerando os critérios de gini e entropia faça a simulação considerando a base de validação com 30% e depois 20%. Analise os resultados

b) Faça um modelo de Random Forest considerando a base de validação com 30% e depois 20%. Analise os resultados.

c) Qual melhor opção para o banco?

2) **O arquivo campanha_mkt contém as seguintes variáveis:**

**Idade Sexo Cidade Email Opened Email Clicked Visitas _site Discount offered Compra**

**Onde:**

**Email Opened (E-mail Aberto):** Esta coluna binária (0 ou 1) indica se um cliente abriu um e-mail como parte de uma campanha de marketing. Um valor **1** normalmente significa que o cliente abriu o e-mail, enquanto **0** indica que não abriu.

**Email Clicked (Clique no E-mail):** Esta coluna binária (0 ou 1) representa se um cliente clicou em um link dentro de um e-mail da campanha de marketing. Um valor **1** sugere que o cliente clicou no link, enquanto **0** sugere que não clicou.

**Discount Offered (Desconto Oferecido):** Esta coluna binária (0 ou 1) indica se um desconto foi oferecido ao cliente como parte da campanha de marketing. Um valor **1** significa que um desconto foi oferecido, enquanto **0** indica que nenhum desconto foi fornecido.

Desenvolva um modelo de previsão de compra considerando

a) Árvore de classificação considerando os critérios de gini e entropia faça a simulação considerando a base de validação com 0%. Analise os resultados

b) Faça um modelo de Random Forest considerando a base de validação com 0%. Analise os resultados.

c) Qual melhor opção?

3) Considere o arquivo prcancer referente a uma amostra de 53 homens. O tratamento e prognóstico de câncer depende de quanto a doença se espalhou. Uma das regiões em que o câncer pode se espalhar refere-se aos nódulos linfáticos. Se os mesmos forem atingidos, o prognóstico é geralmente mais pobre do que em caso negativo. Por isso é desejável estabelecer o quanto antes se os nódulos são cancerosos. Para certos tipos de câncer, cirurgia exploratória é feita só para determinar se os nódulos são cancerosos, uma vez que isso determinará qual o tratamento necessário. Se for possível prever se os nódulos são afetados ou não com base nos dados sem a realização de cirurgia, considerável desconforto e gasto poderão ser evitados. Os dados referem-se a 53 homens com câncer de próstata. Para cada paciente, temos: idade (**age**), serum acid phosphatase (**acid** – um valor de laboratório que é elevado se o tumor se espalhou para certas regiões), o estágio da doença (**stage** – uma indicação do avanço da doença), o grau do tumor (**grade** – uma indicação da agressividade) e os resultados do raio x (**xray**), assim como se o câncer se espalhou para os nódulos da região linfática na fase da cirurgia. O problema é prever se os nódulos são positivos para câncer com base nos valores das variáveis que podem ser medidas sem cirurgia (variável dependente: **node**). As variáveis xray, stage e grade são categóricas, codificadas como 0 e 1. O valor 1 sempre indica a pior situação (raio x positivo, estágio avançado e agressividade).

Repita os mesmos procedimentos do exercício anterior.