Cientista de Dados

Desafio:

# Cientista de Dados

O objetivo do teste é entender como você pensa e como será sua execução no dia.

Em anexo está uma base de dado (*DOT\_teste\_data\_science.csv*) simulando notas de provas de alunos em uma instituição de ensino a distância. A base contém os seguintes dados

* Prova Realizada em : estado onde a pessoa realizou a prova
* Dia da Prova Realizada : data da prova
* Nota : nota de 0 a 10 da prova
* Nota da Redação : nota de 0 a 2,5
* Horário da Prova : período de realização da prova
* Estado civil : estado civil do aluno
* Escolaridade : escolaridade máxima do aluno
* Cidade de nascimento : cidade de nascimento do aluno
* Bairro onde reside : bairro de residência do aluno

A para um aluno ser aprovado na disciplina ele deve tirar, no **mínimo**, 5 na nota da prova. A nota da redação como nota de desempate (entre dois alunos que tiveram nota 5 na prova, o aluno que tiver maior nota na redação tem desempenho melhor). Caso o aluno tire 0 na redação ele é **reprovado** automaticamente. Os alunos são divididos por estado.

A partir destes dados, responda as seguintes perguntas:

1. O time de marketing precisa saber qual é o **público-alvo ideal** para essa instituição de ensino. Quais as caraterísticas principais desse público? (DICA: O público ideal é o público de perfil ‘vencedor’ nas provas)
   1. As maiores frequências de vencedores são solteiros, com nível de educação Ensino Médio, e casados com educação de nível ensino médio.
   2. O segundo público seria as pessoas casadas e solteiras de nível superior completo.
2. O time de monitoria de educação precisa tomar **ações preventivas** para evitar que alunos reprovem na matéria. Como você pode indicar quais são os alunos que necessitam de mais ajuda? (DICA: Existem alguma forma de prever o desempenho desses alunos?)
3. **Seja criativo**! Qual outro insight ou informação você consegue extrair desses dados? Descreva uma forma que alguém poderia utilizar essa informação. (DICA: Lembre que toda análise deve trazer uma informação que pode ser transformada em ação ou servir de tomada de decisão.)
4. Usando um algoritmo de arvore de decisão, quais correlações foram encontradas ? ( disponibilizar o código que vai até o banco, coleta os dados e analisa)
5. Usando um algoritmo bayesiano, quais correlações foram encontradas ? ( disponibilizar o código que vai até o banco, coleta os dados e analisa)
6. Usando um algoritmo apriori, quais correlações foram encontradas ? ( disponibilizar o código que vai até o banco, coleta os dados e analisa)

O resultado deve ser entregue em um **Jupyter Notebook** (pode ser o link para o repositório no GitHub) com o código fonte, análises e respostas das questões. Outros tipos de entrega (relatório, apresentação, dashboard, etc) podem complementar a entrega inicial, mas não são obrigatórios.

Prazo: 4 dias úteis.

Algumas regras:

- O código tem que seguir o padrão PEP-8 (Python)

- Disponibilizar o repositório GitHub e instruções para validarmos o código.

Boa Sorte!