República Federativa do Brasil
Ministério da Educação
Universidade Federal do Amazonas
Instituto de Computação

Aprendizado de Máquina e Mineração de Dados

Lista Prática de Análise de Dados

Introdução: Neste trabalho, vamos analisar dados do INEP sobre o ENADE 2017. Estes dados estão organizados em uma relação com as informações do candidato necessárias para realizar a prova, além das suas notas e as respostas que os candidatos deram a questionários sobre a prova, seus dados pessoais e curso realizado. Em anexo, está notebook com as perguntas deste trabalho, no qual o aluno deverá também respondê-las, e um notebook que contém o **dicionário de dados** com uma descrição detalhada das informações obtidas.

1. Análise de valores faltantes

    1.1. Que colunas possuem campos nulos em **egeral**?
    1.2. Substitua dados faltantes por valores razoáveis. No caso dos questionários, as colunas até QE_I26 podem ser substituídas por um 'Z'. As colunas após a QE_I26 podem ser substituídas por 7 (7 = _Não sei responder_)

2. Consultando os dados e comparando distribuições

    2.1. Novamente considere as distribuições de notas. Como se comparam os alunos do turno noturno com os dos demais turnos? Há mais alunos estudando ao dia ou à noite?
    2.2. Em geral, o MEC acredita que **não** há importantes diferenças de desempenho entre alunos quotistas e não quotistas. Isto é o que você observa, considerando as distribuições de notas dos dois grupos? _Dica_: não são quotistas os alunos que responderam A à pergunta QE_I15.
    2.3. Como se comparam os desempenhos dos alunos de instituições públicas, privadas sem fim lucrativo e privadas com fim lucrativo?

3. Cruzando dados

    3.1. Cruze a informação sobre o turno do curso do aluno com a resposta dada a esta pergunta (QE_I25), de forma a obter, para os motivos dados, o percentual de alunos que os escolheram, de acordo com os cursos realizados.

3.2. É verdade que, quando comparado às instituições privadas, estão nos cursos das universidades federais tanto os estudantes mais pobres quanto os mais ricos do país?
3.3. Qual a nota geral média dos alunos, de acordo com a forma como ele cursou ensino médio?
3.4. Qual a nota geral média dos alunos, de acordo com sua renda, por região? Caso sua base se refira a uma única região, considere a renda por estado daquela região.

4. Agrupando e ordenando dados

4.1. Qual o ranking dos estados de acordo com a nota média obtida por alunos, considerando apenas instituições públicas?

5. Comportamentos anômalos

5.1. Como se comparam as variáveis **idade**, **nota geral**, **ano de início da graduação** e **ano de fim do ensino médio** em termos de anomalias?

6. Engenharia de atributos

6.1. Seguindo a definição dada, crie a coluna NT_INST que reflete a nota que cada estudante daria para a instituição.

7. Preparando dados para classificação

7.1. Prepare os dados para aprendizado, convertendo strings para dados categóricos, dados categóricos não binários para hot-vectors e padronizando dados numéricos usando Z-score.

8. Questão bônus

8.1. Que atributos a RandomForest julgou mais relevantes para determinar se o aluno vai ou não passar?