Trabalho 3
Student Performance Dataset

1 Descrição do Dataset

Nestre trabalho, você irá classificar o rendimento escolar de alunos do ensino fundamental de dois colégios (*Gabriel Pereira* e *Mousinho da Silveira*) em um ano letivo. Para isso, iremos utilizar um dataset com 30 anotações (algumas numéricas e outras categóricas), a respeito do aluno e de sua família. Alguns exemplos destas anotações são:

- **Sexo**
- **Idade**
- **Endereço**
- **Tamanho da família**
- **Se os pais moram juntos**
- **Escolaridade da mãe**
- **Escolaridade do pai**
- **Profissão da mãe**
- **Profissão do pai**
- **Duração do trajeto casa-escola**
- **Tempo de estudo semanal**
- **Número de reprovações passadas**
- **Se faz reforço escolar**
- **Se pratica atividades extra-curriculares**

O objetivo será classificar, a partir destas informações, se o aluno foi aprovado naquele ano letivo (atributo *"approved"* do dataset). Para uma explicação de cada feature e os valores que elas podem assumir, veja o arquivo *"dataset_info.txt"*.

2 Tarefas

Pedimos que você:

1. Inspecione os dados. Quantos exemplos você tem? Qual o intervalo de valores de cada feature?
2. Como baseline, treine uma árvore de decisão para o problema.
3. Treine também florestas aleatórias variando o número de árvores geradas.
4. Calcule a matrix de confusão e a acurácia normalizada para o conjunto de treino/validação e compare os modelos treinados. Houve *overfitting* em algum caso? Qual modelo obteve o melhor resultado?
5. Escreva um relatório de no máximo 3 páginas:
   (a) Descreva o que foi feito, bem como as diferenças entre o seu melhor modelo e o seu baseline;
   (b) Reporte o resultado do baseline e da melhor configuração de florestas aleatórias no conjunto de teste (será disponibilizado alguns dias antes do prazo final de submissão).
   (c) Escreva pelo menos 1 parágrafo com as conclusões tiradas na atividade;

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *student_performance_train.data*: conjunto de dados para treinamento;
- *student_performance_val.data*: conjunto de dados para validação;
- *student_performance_test.data* **(será disponibilizado na sexta-feira anterior ao prazo final da submissão)**: conjunto de dados retido pelo professor;
- *dataset_info.txt*: descrição de cada feature e os valores que elas podem assumir;

4 Avaliação

O dataset foi previamente dividido aleatoriamente em três conjuntos — treino, validação e teste — e apenas os dois primeiros serão disponibilizados para que você implemente as suas soluções.

Na sexta-feira anterior ao prazo final de submissão, iremos disponibilizar no Moodle o conjunto de teste e iremos avisá-lo pelo canal da disciplina no Slack. No relatório, você deve reportar os seus resultados no conjunto de validação e no conjunto de teste.

A avalidação consistirá da análise do relatório e do código submetidos no Moodle. Iremos avaliar se as tarefas pedidas foram realizadas, como o treinamento e validação foi feito, os resultados reportados e as conclusões feitas.

Observações sobre a avaliação:

- O trabalho poderá ser feito individualmente ou em duplas, podendo haver repetição das duplas a cada trabalho;
- O código e o relatório deverão ser submetidos no Moodle por **apenas um integrante da dupla**;
- Não se esqueçam de listar os nomes dos integrantes da dupla no início do relatório;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;