Pontifícia Universidade Católica - PUCRS

Disciplina: Inteligência Artificial
Prof. Lucas Rafael Costella Pessutto

Trabalho 2

Aprendizado de Máquina

Objetivo: O objetivo deste trabalho é aplicar algoritmos de aprendizado de máquina em um conjunto de dados pré-definido.

Administrativo: O trabalho deve obedecer as seguintes regras:

- Grupos: o trabalho pode ser feito em **duplas ou trios**, não sendo possível realizar individualmente.
- Data de Entrega: 25/11/2024 até às 23:59, via moodle. Apenas um dos integrantes deve submeter os arquivos do trabalho no moodle.
- Apresentações: Aulas dos dias 26/11/2024 e 28/11/2024. Todos os integrantes devem estar presentes e devem participar da apresentação.
- Fórum: Há um fórum no moodle onde devem ser indicados os grupos juntamente com o dia preferido para a apresentação. A escolha do horário de apresentação é por ordem de indicação no fórum.
- Avaliação: A nota deste trabalho será dividida em: implementação (70%) e apresentação (30%).

O problema
Nesse trabalho estudaremos um conjunto de dados que contém uma pesquisa de opinião com mais de 50 mil pessoas (a pesquisa não foi real, o dataset é sintético) a fim de entender onde elas preferem tirar férias: nas montanhas ou na praia[1].
Cada grupo deve estudar o dataset, ou seja, entender o que cada feature representa, e aplicar pelo menos **três** algoritmos de aprendizagem de máquina sobre esse dataset, de modo a obter um bom classificador que receba as informações de uma pessoa e retorne qual a sua preferência.
O dataset possui 13 features, listadas abaixo:

- **Age**: Idade do indivíduo (numérico).

[1] O dataset pode ser encontrado em: `https://www.kaggle.com/datasets/jahnavipaliwal/mountains-vs-beaches-preference/data`

- **Gender**: Identidade de gênero do indivíduo (categórico: male, female, non-binary).
- **Income**: Renda anual do indivíduo (numérico).
- **Education Level**: Nível mais alto de escolaridade alcançado (categórico: high school, bachelor, master, doctorate).
- **Travel Frequency**: Número de férias tiradas por ano (numérico).
- **Preferred Activities**: Atividades preferidas pelos indivíduos durante as férias (categórico: hiking, swimming, skiing, sunbathing).
- **Vacation Budget**: Orçamento destinado para férias (numérico).
- **Location**: Tipo de residência (categórico: urban, suburban, rural).
- **Proximity to Mountains**: Distância da montanha mais próxima (numérico, em milhas).
- **Proximity to Beaches**: Distância da praia mais próxima (numérico, em milhas).
- **Favorite Season**: Estação preferida para férias (categórico: summer, winter, spring, fall).
- **Pets**: Indica se o indivíduo possui animais de estimação (binário: 0 = Não, 1 = Sim).
- **Environmental Concerns**: Indica se o indivíduo possui preocupações ambientais (binário: 0 = Não, 1 = Sim).

A última coluna do dataset (Preference) contém a variável objetivo deste dataset. Ela é uma variável binária que indica se a preferência daquele indivíduo é praia (0) ou montanhas (1). Além da aplicação dos algoritmos, cada grupo será avaliado pelo tratamento dado ao dataset, o que inclui:

- Configuração dos hiperparâmetros dos algoritmos de aprendizagem
- Análise exploratória dos dados: gráfica / estatística
- Tratamento de valores faltantes
- Seleção de Features
- Transformação de Atributos
- Normalização
- Agregação
- Criação de Features
- etc.

Todos os resultados devem ser reportados usando o precisão, revocação e F_1 score para as duas classes e para todo o experimento (média das classes). Devem ser utilizadas as partições de treino/teste/validação fornecidos pelo professor para fazer o treinamento do modelo e realização dos testes.

Apresentação

A apresentação do trabalho será feita para toda a turma nas aulas previstas no cronograma. Cada grupo terá 10 minutos para se apresentar. Deverão ser confeccionados alguns slides descrevendo o pré-processamento, quais algoritmos foram utilizados e resultados obtidos.

Observações

O grupo é livre para decidir a linguagem de programação a ser utilizada na implementação. Além disso, o grupo pode utilizar outras bibliotecas que implementem algoritmos de aprendizagem de máquina, desde que o grupo estude a documentação destas bibliotecas e saiba descrever os algoritmos que estão sendo utilizados no trabalho.

É permitido o uso de ferramentas de IA para gerar código-fonte, desde que: (i) elas sejam usadas somente para tarefas auxiliares e (ii) o grupo indique com comentários no código quais os trechos gerados por IA. O mesmo vale para código que não foi produzido pelo grupo. Se for identificado plágio de qualquer natureza a nota do trabalho será zerada.