Projeto Final

Lista de Projetos

A avaliação da disciplina de Projeto Final (INF-0619) será constituída por dois projetos, um básico e um avançado, que serão alocados para cada grupo, considerando a lista de preferência de projetos que cada grupo poderá fornecer através de seu coordenador. Lembrando que não há garantia de que o projeto com maior preferência escolhido por um grupo será alocado para o mesmo.

Os projetos estão disponíveis em plataformas de modelagem preditiva e de competições analíticas, como Kaggle (`www.kaggle.com`), DrivenData (`www.drivendata.org`) e CrowdAnalytix (`www.crowdanalytix.com/community`). Além disso, iremos disponibilizar os dados de cada projeto no Google Drive (`http://bit.ly/projetos-mdc`). Os projetos estão divididos nas seguintes duas categorias:

Projetos Básicos:

- Characters from Product Images
- Default of Credit Card Clients
- Dogs vs. Cats
- Fruits-360
- Medical Appointments
- PetFinder Adoption Prediction

Projetos Avançados:

- Doodle Recognition
- Landmark Recognition
- Pump it Up
- Quora Questions
- Richter's Predictor
- Toxic Comment Classification

Observações

- A Apresentação Parcial será feita de forma oral (4 a 5 minutos), com auxílio de slides (entre 5 e 10).
- A Apresentação Final será feita por meio de um vídeo de 4 a 5 minutos, preparado por cada grupo.
- Pontos importantes a serem abordados nas apresentações:
  - Técnicas utilizadas.
  - Resultados obtidos.
  - Dificuldades encontradas.
- Após as apresentações, os membros da banca avaliadora podem realizar perguntas ao grupo sobre o trabalho realizado.
- Complementando a Apresentação Final de cada projeto, um relatório de até 10 páginas (em formato `PDF`) deverá ser submetido no Moodle, pelo coordenador do grupo, junto com todos os arquivos fontes utilizados na realização do projeto.
- Os vídeos das Apresentações Finais serão publicadas no canal YouTube do curso (`http://bit.ly/youtube-mdc`).
- Qualquer tentativa de fraude implicará em nota zero na disciplina para todos os membros do grupo, sem prejuízo de outras sanções.
- O projeto final não pode ser compartilhado entre os grupos, o que caracteriza tentativa de fraude.
- Projeto final realizado com ajuda externa será considerado como tentativa de fraude.

Characters from Product Images

Descrição

Esse conjunto de dados é composto de imagens de produtos (como camisetas, chaveiros, bolsas) com desenhos de personagens da cultura pop.

Objetivo Principal

Classificar qual o personagem retratado no produto.

Técnicas Envolvidas

- Classificação multi-classe.
- Técnicas para análise de imagem.

Desafios

Para esse projeto, alguns desafios são:

- Classificação de imagens de 42 personagens.
- Lidar com imagens com variações de iluminação, alinhamento, ruído, oclusão.
- Montar uma rede capaz de identificar características nas imagens para identificar personagens de diferentes tipos.

Conjunto de Dados

- `http://bit.ly/mdc-character`
- `https://www.crowdanalytix.com/contests/identify-characters-from-product-images`

Default of Credit Card Clients

Descrição

Esse conjunto de dados é composto de dados de pagamentos efetuados por clientes de cartão de crédito de Taiwan.

Objetivo Principal

Predizer se um determinado cliente irá ou não pagar faturas do cartão de crédito.

Técnicas Envolvidas

- Análise descritiva de dados.
- Classificação binária.
- Clustering.

Desafios

Para esse projeto, alguns desafios são:

- Análise descritiva para identificar qual é a probabilidade de ocorrer falta de pagamento por diferentes variáveis demográficas.
- Identificar as variáveis mais relevantes que determinam o não pagamento das faturas do cartão de crédito.

Conjunto de Dados

- `http://bit.ly/mdc-defaultcredit`
- `https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset`

Dogs vs. Cats

Descrição

Esse conjunto de dados é composto de imagens coletadas da internet, exibindo cachorros e gatos em diferentes ambientes, poses e com diferentes condições de captura.

Objetivo Principal

Classificar corretamente se a imagem corresponde a um cachorro ou a um gato.

Técnicas Envolvidas

- Classificação binária.
- Técnicas para análise de imagem.

Desafios

Para esse projeto, alguns desafios são:

- Classificar corretamente o animal a partir da imagem.
- Lidar com imagens com variações de iluminação, pose, ruído e oclusão.
- Montar uma rede capaz de identificar características nas imagens para identificar o animal.

Conjunto de Dados

- `http://bit.ly/mdc-dogsvscats`
- `https://www.kaggle.com/c/dogs-vs-cats-redux-kernels-edition/`

Fruits-360

Descrição

Esse conjunto de dados é composto de imagens de diversas frutas para classificação.

Objetivo Principal

Classificar corretamente as frutas contidas nas imagens.

Técnicas Envolvidas

- Classificação multi-classe.
- Técnicas para análise de imagem.

Desafios

Para esse projeto, alguns desafios são:

- Classificação de 64 tipos de frutas no total.
- Montar uma rede capaz de identificar características nas imagens para identificar frutas de diferentes tipos.

Conjunto de Dados

- `http://bit.ly/mdc-fruits360`
- `https://www.kaggle.com/moltean/fruits`

Medical Appointments

Descrição

Esse conjunto de dados é composto de informações sobre diversos paciente.

Objetivo Principal

Predizer se um determinado paciente irá faltar ou não a uma consulta.

Técnicas Envolvidas

- Análise descritiva de dados.
- Classificação binária.

Desafios

Para esse projeto, alguns desafios são:

- Análise descritiva de dados e classificação para identificar se um paciente irá ou não faltar a uma consulta.
- Estratificar características que mais determinam a ausência de pacientes em consultas.

Conjunto de Dados

- `http://bit.ly/mdc-medical`
- `https://www.kaggle.com/somrikbanerjee/predicting-show-up-no-show`

PetFinder Adoption Prediction

Descrição

Esse conjunto de dados é composto de informações sobre adoções de animais na plataforma PetFinder.

Objetivo Principal

Predizer quão rápido um animal será adotado a partir de suas informações.

Técnicas Envolvidas

- Análise descritiva de dados.
- Classificação multi-classe.

Desafios

Para esse projeto, alguns desafios são:

- Análise descritiva de dados.
- Classificar a velocidade de adoção em cinco categorias.
- Identificar as características mais importantes para determinar a velocidade de adoção de um animal.

Conjunto de Dados

- `http://bit.ly/mdc-petfinder`
- `https://www.kaggle.com/c/petfinder-adoption-prediction`

Doodle Recognition

Descrição

Esse conjunto de dados é composto de imagens desenhadas com o mouse ou trackpad (*doodle*) no jogo *Quick, Draw!* (`https://quickdraw.withgoogle.com/`).

Objetivo Principal

Predizer qual objeto corresponde à imagem desenhada.

Técnicas Envolvidas

- Classificação multi-classe.
- Clusterização.
- Técnicas para análise de imagem.

Desafios

Para esse projeto, alguns desafios são:

- Converter a representação dos doodles (pontos e vetores) em imagens.
- Lidar com imagens de diferentes qualidades e com poucas informações visuais.
- Classificar o doodle entre 100 possíveis categorias.
- Agrupar doodles visualmente semelhantes.

Conjunto de Dados

- `http://bit.ly/mdc-doodle`
- `https://www.kaggle.com/c/quickdraw-doodle-recognition`

Landmark Recognition

Descrição

Esse conjunto de dados é composto de imagens de pontos de referência (*landmarks*) naturais ou feitos pelo homem espalhados pelo mundo.

Objetivo Principal

Identificar corretamente o landmark retratado na imagem.

Técnicas Envolvidas

- Classificação multi-classe.
- Clusterização.
- Técnicas para análise de imagem.

Desafios

Para esse projeto, alguns desafios são:

- Lidar com a grande quantidade de imagens capturadas em diferentes condições (iluminação, ruído, ângulos).
- Classificar a imagem entre 100 possíveis categorias.
- Agrupar landmarks visualmente semelhantes.

Conjunto de Dados

- `http://bit.ly/mdc-landmark`
- `https://www.kaggle.com/c/landmark-recognition-2019`

Pump it Up
Data Mining the Water Table

Descrição

Esse conjunto de dados é composto de informações de bombas de água instaladas em poços artesianos na Tanzânia.

Objetivo Principal

A partir das informações de uma bomba de água — como localização, data de instalação, quantidade disponível de água — identificar se ela está operante, se necessita de reparos ou se não está funcionando.

Técnicas Envolvidas

- Análise descritiva dos dados.
- Classificação multi-classe.
- Clustering.

Desafios

Para esse projeto, alguns desafios são:

- Analisar as informações disponíveis para poços e bombas.
- Classificar a operação da bomba a partir das informações disponíveis.
- Agrupamento em função de características das bombas.

Conjunto de Dados

- `http://bit.ly/mdc-pumpitup`
- `https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table`

Quora Question Pairs

Descrição

Esse conjunto de dados é composto por uma grande quantidade de pares de perguntas do Quora.

Objetivo Principal

Identificar pares de perguntas que possuem o mesmo significado a partir de seu texto.

Técnicas Envolvidas

- Classificação binária.
- Técnicas de processamento de linguagem natural.

Desafios

Para esse projeto, alguns desafios são:

- Processar texto das perguntas para lidar com *stop words* e sinônimos.
- Classificar as perguntas em duplicadas ou não.
- Identificar os termos e palavras que foram determinantes para classificar um par de perguntas como duplicado.

Conjunto de Dados

- `http://bit.ly/mdc-quora`
- `https://www.kaggle.com/c/quora-question-pairs`

Richter's Predictor
Modeling Earthquake Damage

Descrição

Esse conjunto de dados é composto por informações de construções que foram destruídas pelo terremoto Gorkha em 2015, no Nepal.

Objetivo Principal

Predizer o grau de destruição causado pelo terremoto nas construções, a partir de informações sobre localização e características do edifício.

Técnicas Envolvidas

- Análise descritiva dos dados.
- Classificação multi-classe.
- Clusterização.

Desafios

Para esse projeto, alguns desafios são:

- Análise descritiva de dados das construções.
- Estratificar características dos imóveis relacionadas com o grau de destruição do terremoto.
- Classificar o grau de destruição em 3 categorias, a partir das informações de cada construção.

Conjunto de Dados

- `http://bit.ly/mdc-richter`
- `https://www.drivendata.org/competitions/57/nepal-earthquake/`

Toxic Comment Classification

Descrição

Esse conjunto de dados é composto por comentários na Wikipedia, cujos conteúdos foram anotados em relação à toxicidade. Comentários tóxicos são definidos por serem *rudes, desrespeitosos ou que provavelmente fariam alguém abandonar uma discussão*. Além disso, um subconjunto dos comentários foi anotado em relação às identidades mencionadas em seu conteúdo (por exemplo, homens, mulheres, asiáticos, ateus, entre outras).

Objetivo Principal

Analisar o conteúdo de cada comentário e identificar comentários tóxicos.

Técnicas Envolvidas

- Análise descritiva dos dados.
- Classificação binária.
- Técnicas de processamento de linguagem natural.

Desafios

Para esse projeto, alguns desafios são:

- Processar texto dos comentários para lidar com *stop words* e sinônimos.
- Classificar os comentários em relação ao grau de toxicidade.
- Identificar os termos e palavras mais comuns em comentários tóxicos ou relacionados às identidades mencionadas.

Conjunto de Dados

- `http://bit.ly/mdc-toxic`
- `https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification`