



Especificação do Projeto

Introdução

Este documento especifica o projeto que cada grupo deverá desenvolver como Projeto Integrado.

Descrição do Projeto

A partir de um objetivo de negócio escolhido por cada grupo, o projeto consiste em sistematizar as técnicas apresentadas durante o semestre para atingir os objetivos.

Método

Cada projeto deve seguir o seguinte método:

- Oferecer uma contextualização do contexto do negócio e da necessidade da solução que seu grupo irá propor.
- Estabelecer claramente o objetivo do trabalho
- Implementar o projeto
- Analisar como a implementação atende ao objetivo proposto
- Conclusão

Composição dos grupos

Cada grupo deve ter entre 2 e 6 alunos. Trabalhos individuais não são permitidos.

Requisitos do projeto

Os projetos devem utilizar as mesmas técnicas apresentadas durante o curso para cumprir os requisitos.

1. Utilizar um ou mais datasets (não pode ser toy) para o treinamento dos classificadores. O dataset deve conter múltiplas dimensões e ter classes desbalanceadas¹.
2. Realizar uma análise exploratória do dataset por meio de um caderno Jupyter. Utilize gráficos na análise.
3. Treinar um classificador Naive Bayes (Multinomial, Bernoulli ou Gaussian)
4. Treinar um classificador Support Vector Machine (SVM)
5. Treinar um classificador de Árvore de Decisão ou Floresta Aleatória
6. Reduzir a dimensionalidade com Principal Component Analysis (PCA) e interpretar os resultados.
7. Todos os classificadores devem ser avaliados com validação cruzada, utilizando as métricas F_β, acurácia, revocação (recall), precisão. Além disso, utilizem a matriz de confusão para a visualização do desempenho

¹ Para atender todos os requisitos do projeto, vocês podem complementar o dataset com dados realistas, assim como feito nos cadernos e no projeto- piloto.



Pontifícia Universidade Católica de São Paulo
Faculdade de Estudos Interdisciplinares
Ciência de Dados e Inteligência Artificial
CDIA4-23/2 Projeto Integrado
Professor Dr. Rooney R. A. Coelho

PUC-SP

8. Todos os classificadores devem ser persistidos (joblib, pickle) antes de serem entregues e publicados no GitHub.
9. O projeto-final de contar com um modelo por integrante do grupo, possivelmente além dos modelos exigidos nesta especificação.
10. Aplicar um frameworks de séries temporais (opcional)
11. Aplicar PySpark (opcional)

Como cada projeto será avaliado?

Para cada disciplina, a nota final do projeto (A_{proj}) será dada da seguinte forma:

R = cumprimento dos requisitos (0 = baixo; 5 = médio; 10 = alto)

M = aplicação do método (0 = baixo; 5 = médio; 10 = alto)

O = relevância do objetivo de negócio e do trabalho (0 = baixo; 5 = médio; 10 = alto)

J = qualidade da apresentação do relatório Jupyter (0 = baixo; 5 = médio; 10 = alto)

B = qualidade da apresentação (powerpoint e oral) para a banca (0 = baixo; 5 = médio; 10 = alto)

$$A_{proj} = \frac{R + M + O + J + B}{50}$$

Como cada aluno será avaliado?

Cada membro do grupo fará uma autoavaliação e uma avaliação anônima da participação dos outros membros do grupo na elaboração do projeto. Em todas as avaliações, será necessário justificar a nota atribuída a si e aos pares. Essa avaliação é conhecida como avaliação 360. A nota a_{360} de cada aluno será dada pela média entre as avaliações dos pares e a autoavaliação.

A nota final de projeto de cada aluno é dada por:

$$A_{est} = A_{proj} \cdot \frac{a_{360}}{10}$$

Apresentação parcial

Os alunos deverão fazer uma apresentação com o resultado parcial da pesquisa no dia 25/09/2023

Prazo final

O trabalho deve ser entregue até o dia 20/11/2023.