Introdução à Modelagem de Big Data - Projeto III

Marcus Nunes

18 de Outubro de 2021

Resumo

Esta é a última avaliação da turma de Introdução à Modelagem de Big Data (EST0133) do semestre 2021.2. É um projeto de análise de dados, com assunto e escopo definido pelo próprio aluno, que deverá vir acompanhado de um relatório e ser apresentado em sala de aula em 09/02/2022.

Instruções

- Este projeto será avaliado de duas maneiras: virtualmente, numa apresentação de slides, e por escrito, através de um relatório
- A apresentação de slides deverá durar de 10 a 15 minutos
- Não há limite mínimo ou máximo para o número de páginas do relatório escrito
- Uma das seções do relatório deverá conter o código utilizado pelos alunos para resolver seu problema
- Os arquivos com a apresentação de slides, relatório e conjunto de dados devem ser enviados em um arquivo .zip pelo SIGAA, até às 23:59 do dia 09/02/2022

Introdução

Este projeto visa avaliar aquilo que os alunos aprenderam durante o curso. Serão formados grupos de até três alunos, que irão realizar a análise de um conjunto de dados à sua escolha. Esta escolha é de inteira responsabilidade dos alunos. Algumas sugestões de áreas para o projeto são

- Dados biológicos
- Webscraping
- Mineração de textos
- Genética
- Dados de localização
- Marketing
- Finanças
- Dados de saúde

Cada grupo deverá obter um banco de dados em algum destes assuntos ou outros, de modo a responder perguntas a seu respeito. É possível usar dados de suas próprias pesquisas, caso o aluno queira propor uma nova maneira de abordar seu trabalho. As perguntas a serem respondidas serão formuladas pelos próprios alunos. A única obrigatoriedade é a aplicação de ferramentas de análise de dados vistas no curso. Os grupos

que quiserem propor a utilização de uma nova ferramenta que tenha similaridade com aquelas que vimos em aula sintam-se encorajados a fazê-lo. Algumas boas fontes para procura de dados:

- Brasil.IO - https://brasil.io/
- Kaggle - https://www.kaggle.com/datasets
- UCI Machine Learning Repository - http://archive.ics.uci.edu/ml/
- Slides da aula *Obtenção de Dados*

Exigências a Respeito dos Bancos de Dados a Serem Escolhidos

Há duas restrições a respeito dos dados a serem analisados:

1. Eles não podem ter sido analisados previamente em alguma aula da disciplina
2. Não podem haver dois ou mais grupos com bancos de dados que sejam subconjuntos um do outro ou com bancos de dados que sejam parte um um conjunto maior

Para evitar que hajam problemas respeito do item 2, solicito aos grupos que enviem um email para marcus .nunes@ufrn.br descrevendo o banco de dados que utilizarão. Assim, os primeiros grupos a escolherem o conjunto de análise terão prioridade na escolha.

Conteúdo do Trabalho

Alguns assuntos que podem ser tratados no projeto são

- Comparação entre dois ou mais métodos de classificação ou regressão
- Construção de um aplicativo no shiny
- Demonstração de uma alguma ferramenta de machine learning não vista durante a disciplina (Deep Learning, eXtreme Gradient Boosting etc.)
- Paralelização de código utilizando GPU
- Reprodução de algoritmos vistos em aula em outras linguagens, como Python ou Julia, e como o desempenho deles se compara ao R

Portanto, os alunos não precisam se limitar aos conteúdos vistos em aula.

Datas Importantes

- 15/12/2021: sorteio da ordem de apresentação oral dos trabalhos e entrega do pré-projeto, onde vão constar
  - nomes dos membros dos grupos
  - descrição do conjunto de dados a ser analisado
  - as perguntas a serem respondidas no projeto
- 09/02/2022: apresentação dos projetos em aula
  - os slides da apresentação devem ser submetidos via SIGAA
  - o relatório do trabalho, bem como o código utilizado em sua resolução, deve ser submetido também

Recomendações Gerais

- Escolha um assunto do seu interesse. Mesmo que nenhuma das opções apresentadas lhe agrade, avalie quais são os seus gostos pessoais e discuta comigo se é possível apresentar um trabalho a respeito destas preferências.
- Comece a procurar seus dados e a realizar as suas análises o quanto antes. Quanto mais cedo vocês começarem, mais tempo vocês terão para realizarem o trabalho.
- Vocês estão limitados apenas pela imaginação. Sejam curiosos e criativos.

Ficha de Avaliação

Abaixo estão descritos os critérios de avaliação do projeto, com seus respectivos pesos.

Item	Caráter	Peso	Descrição
Pré-Projeto	Classificatório	5	Documento sucinto descrevendo o conjunto de dados e os objetivos da análise
Conjunto de Dados	Eliminatório	10	Conjunto de dados **inédito** na disciplina, isto é, não será permitido reanalisar algum conjunto de dados com o qual já tenhamos trabalhado
Análise Exploratória	Eliminatório	10	Realizar a análise exploratória dos dados, com gráficos e tabelas coerentes com o problema
Modelagem	Eliminatório	25	Ajustar ao menos um modelo aos dados, seja clusterização, classificação ou regressão
Apresentação	Eliminatório	20	Preparação dos slides e habilidade de comunicar os resultados para a audiência
Preparo Individual	Classificatório	10	Capacidade de responder perguntas antes e depois da apresentação
Relatório	Eliminatório	20	Qualidade geral do relatório em relação à organização, escrita e resultados obtidos