Universidade Federal de Viçosa – Campus Florestal
Bacharelado em Ciência da Computação

O Brasil em Dados

Trabalho Prático – Introdução à Ciência dos Dados

Professor: Fabrício A. Silva
Entrega final: 26/07/2022 (ver cronograma abaixo)
Grupo: 3 ou 4 alunos (avaliação será individual de acordo com entrevista e ações no github)
Forma de Entrega:
1) arquivo de relatório feito no *Jupyter Notebook* (usando Markdown, por exemplo), com documentação sobre decisões, resultados, gráficos, discussão sobre os resultados, e código fonte. Disponibilizar no Github para o professor (usuário: fabaguiarsilva) e o monitor (usuário: gfviegas).
2) apresentação do projeto em 10 minutos, com foco nos resultados descobertas, e não nas técnicas utilizadas.

Introdução

Na maioria das vezes, os dados utilizados em um problema real para a extração de conhecimento e predição de acontecimentos são desorganizados, com ruído, erros ou campos vazios. Além disso, resultados que são aparentemente muito prováveis e esperados, muitas vezes não são observados nos dados.

O objetivo do trabalho prático é aplicar os conteúdos aprendidos em sala de aula em um projeto real, com dados reais disponíveis publicamente. Com isso, os alunos irão enfrentar muitas das dificuldades que um cientista de dados deve estar preparado para lidar.

Em particular, os dados a serem utilizados devem ser referentes ao Brasil, em relação a educação, saúde, violência, eleições, bolsa-família, distribuição de renda, justiça, previdência, transporte/trânsito, bolsa de valores, turismo, dentre outros aspectos.

Etapas

Para que esse objetivo seja alcançado, o trabalho está dividido em três entregas:

1. Escolha dos dados e planejamento (5 pontos): Nesta etapa, o grupo irá escolher o(s) conjunto(s) de dados que será(ão) utilizado(s) no trabalho. Escolha um conjunto de dados que esteja relacionado a algum tema de interesse do grupo. No final deste documento são indicadas algumas fontes de dados, mas não fiquem restritos a elas. Após escolher os dados, o grupo deverá indicar a escolha do tema via fórum do Moodle (a escolha do assunto é por ordem de envio da mensagem). Por fim, o grupo deve elaborar uma lista de pelo menos 20 questões que pretende responder com o trabalho.

Entrega etapa 1: 26/05/2022 (criar o projeto no GitHub, e incluir o professor (usuário: fabaguiarsilva) e o monitor (usuário: gfviegas). Criar arquivo README com integrantes do grupo, assunto a ser tratado no projeto, links para os conjuntos de dados, e as 20 questões elaboradas.

2. Preparação e análise exploratória dos dados (10 pontos): Com os dados em mãos, a próxima etapa é preparar o ambiente para que a análise dos mesmos seja realizada. Essa etapa envolve entender os atributos e objetos dos dados, os tipos de cada atributo, o domínio de cada atributo, verificar e identificar possíveis ruídos ou informações ausentes, criar novos atributos se necessário, formatar valores, juntar conjuntos de dados, dentre outras atividades. Nesta etapa, o grupo também irá gerar estatísticas descritivas, gráficos e tabelas para conhecer os dados. Todo conhecimento

importante extraído deverá ser documentado e discutido. Pensem fora da caixa e tentem extrair correlações não óbvias entre os atributos e objetos. Nesta etapa, o objetivo é responder parte dos questionamentos elaborados. Lembrem-se que novos questionamentos podem surgir.

Entrega etapa 2: 21/06/2022 (entregar no github relatório com documentação, decisões, e código).

3. Análise preditiva (15 pontos): Nesta etapa, o grupo irá aplicar algum algoritmo de aprendizagem (regras de associação, regressão, aprendizado supervisionado ou aprendizado não-supervisionado) para classificar ou agrupar os dados e, assim, tentar prever algum acontecimento desconhecido.

Entrega etapa 3: 26/07/2022 (entregar via github relatório final, incluindo todas as etapas anteriores).

4. Apresentação Final (10 pontos): Cada grupo deverá gravar e disponibilizar no Youtube uma apresentação de 10 minutos, contendo as principais descobertas do trabalho. Foque mais nos interesses do negócio, e não nas técnicas. Imagine que a platéia não esteja interessada em como você chegou a tais conhecimentos, mas apenas nos conhecimentos em si. As gravações das apresentações serão transmitidas para todos nos dias 02/08/22 e 04/08/22 de acordo com sorteio dos grupos.

Lista de Sugestões de Fontes de Dados

https://colaboradados.github.io

https://www.ibge.gov.br

https://downloads.ibge.gov.br

http://inep.gov.br/dados

http://portalms.saude.gov.br/dados-e-indicadores-da-saude

http://datasus.saude.gov.br/informacoes-de-saude

http://www.previdencia.gov.br/dados-abertos/dados-abertos-previdencia-social/

http://www.ipea.gov.br/atlasviolencia/

http://portal.inep.gov.br/provas-e-gabaritos

http://dados.gov.br

http://www.curitiba.pr.gov.br/dadosabertos/

https://prefeitura.pbh.gov.br/transparencia

http://www.bmfbovespa.com.br/pt_br/servicos/market-data/historico/

https://developer.twitter.com/en.html

https://covid.saude.gov.br