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: 26/10/2021 (ver cronograma abaixo)
Grupo: 3 alunos
Forma de Entrega:
a) arquivo de relatório feito no *Jupyter Notebook* (Markdown, por exemplo), com documentação sobre decisões, resultados, discussão sobre os resultados, e código fonte. Disponibilizar no Github para o professor (usuário: fabaguiarsilva)
b) 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 quatro etapas:

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 PVANet (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: 12/08/2021 (entrega de documento com integrantes do grupo, assunto, links para os conjuntos de dados, link para o Github do projeto e questões elaboradas)

2. Preparação dos dados (5 pontos): Com os dados em mãos, a primeira 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.

Entrega etapa 2: 14/09/2021 (entregar relatório com documentação, decisões, e código).

3. Análise exploratória e extração de conhecimento (10 pontos): Com os dados preparados, chegou a hora de explorar e extrair conhecimento dos mesmos. Nesta etapa, o grupo 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 3: 07/10/2021 (entregar relatório com documentação, decisões, e código).

4. Análise preditiva (10 pontos): Nesta etapa, o grupo irá aplicar algum algoritmo de aprendizagem de máquina para classificar ou agrupar os dados e, assim, tentar prever algum acontecimento desconhecido.

Entrega: 26/10/2021 (entregar relatório final, incluindo todas as etapas anteriores), e apresentação de 10 minutos do projeto de forma remota para o professor e colegas.

5. Apresentação (10 pontos): Apresentação para o professor, 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 apresentações serão nos dias 19/10, 21/10 e 26/10 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