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

A UFV em Dados

Trabalho Prático – Introdução à Ciência dos Dados

Professor: Fabrício A. Silva
Entrega final: 24/06/2024 (ver cronograma abaixo)
Grupo: 4 alunos (avaliação será individual de acordo com entrevistas e ações no github)
Forma de Entrega:
1) arquivo de relatório feito no *Jupyter Notebook* (**usando Markdown**), com documentação sobre decisões, resultados, gráficos, discussão sobre os resultados, e código fonte. Exportem o arquivo em *html* ou *pdf*. Disponibilizar no GitHub para o professor (usuário: fabaguiarsilva) e o monitor (usuário: gegen07). 2) apresentação do projeto em 10 minutos, **com foco nos resultados descobertos**, 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 são referentes aos alunos da UFV. O objetivo principal é identificar características que estão associadas com a evasão dos alunos, com a nota do ENEM e com o CRA (dentre outros aspectos), e criar modelos para tentar prever algum desses elementos ou agrupar alunos com características similares.

Etapas

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

1. Entendimento inicial dos dados e planejamento (5 pontos): Nesta etapa, o grupo irá fazer uma primeira análise dos dados, para identificar os atributos existentes, e elaborar uma lista de pelo menos 10 perguntas que pretende responder com o trabalho. O principal foco deve ser na análise da evasão, nota do ENEM e CRA. Porém, outras possibilidades podem ser incluídas.

Entrega etapa 1: 03/04/2024 (criar o projeto no GitHub, e incluir o professor e o monitor como colaboradores). Criar arquivo README com integrantes do grupo e as 10 perguntas elaboradas. Já incluir um código inicial com a leitura dos dados e uma primeira passada pelos atributos, olhando para o nome, tipos, valores, dentre outros aspectos.

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 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 das perguntas elaboradas. Lembrem-se que novos questionamentos podem surgir.

Entrega etapa 2: 08/05/2024 (entregar no GitHub relatório feito no Jupyter Notebook com Markdown 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, com foco em evasão, nota do ENEM e CRA.

Entrega etapa 3: 24/06/2024 (entregar via GitHub relatório final feito no Jupyter Notebook com Markdown, incluindo todas as etapas anteriores).

4. Apresentação Final (10 pontos): Cada grupo deverá fazer uma apresentação com duração aproximada de 10 minutos, contendo as principais descobertas do trabalho. Foque mais na extração de informação baseado nos dados, e não nas técnicas. Imagine que a plateia seja da área de educação, e não da computação, e não esteja interessada em como você chegou a tais conhecimentos, mas apenas nos conhecimentos em si. As apresentações serão feitas nos **dias 24/06/2024 e 26/06/2024** de acordo com sorteio dos grupos.