Universidade Federal de Viçosa – Campus Florestal
Bacharelado em Ciência da Computação
CCF425 Introdução à Ciência dos Dados

Projeto de Desenvolvimento

Professor: Fabrício A. Silva
Monitores: João Marcos e Letícia
Entrega final: 23/06/2025
Valor: 60 pontos
Grupo: 4 alunos (avaliação 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, discussões sobre os resultados, e código fonte. Exportem o arquivo em *html* ou *pdf*. Disponibilizar no GitHub para o professor (usuário: *fabaguiarsilva*) e o monitores (usuários: *raitocan* e *lleticiasilvaa*).

2) Apresentação do projeto em 10 minutos, **com foco nos resultados descobertos.**

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 deste projeto é aplicar os conteúdos aprendidos em sala de aula em um problema 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, temos disponíveis dois conjuntos de dados: *Dados de criminalidade de SP (SPSafe)* e *Dados demográficos dos municípios brasileiros (BrStats)*. Para cada conjunto, há um artigo descrevendo os detalhes de como os dados foram gerados, que deve ser lido criteriosamente pelos membros do grupo. A seleção do conjunto de dados será de acordo com a turma prática da turma (P1: *BrStats* e P2: *SPSafe*).

Etapas

O projeto está dividido em cinco entregas (a distribuição dos pontos pode mudar):

1. Entendimento inicial dos dados e preparação (5 pontos): Nesta etapa, o grupo irá fazer uma primeira análise dos dados, para identificar os atributos

existentes, e elaborar uma lista com pelo menos 10 perguntas que pretende responder com o trabalho. Também devem ser feitos tratamentos dos dados em termos de formatação, enriquecimento com novos atributos externos, tratamento de ausências, dentre outros. Essa etapa envolve entender os atributos e objetos dos dados, o tipo e 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 1: 31/03/2025 (criar o projeto no GitHub, e incluir o professor e os monitores como colaboradores). Criar arquivo README com integrantes do grupo (nome e matrícula) e as 10 perguntas elaboradas. Já incluir o código utilizado, e um relatório parcial com todas as decisões tomadas e suas justificativas. Será feita uma entrevista com cada integrante do grupo. Incluir na documentação o que foi feito por cada integrante do grupo.

2. Análise exploratória dos dados (5 pontos): Com os dados preparados e entendidos, nesta etapa o grupo deve 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: 22/04/2025 (entregar no GitHub relatório feito no Jupyter Notebook com Markdown com documentação, decisões, e código). Será feita uma entrevista com cada integrante do grupo. Incluir na documentação o que foi feito por cada integrante do grupo.

3. Inferência Estatística e Regras de Associação (15 pontos): Nesta etapa, o grupo deve aplicar inferência estatística para uma análise mais rigorosa estatisticamente sobre pelo menos um aspecto dos dados. Além disso, também deve ser aplicado um algoritmo de regras de associação para extrair padrões relevantes dos dados. Os resultados dessas análises devem ser interpretados e discutidos na documentação.

Entrega etapa 3: 12/05/2025 (entregar via GitHub relatório final feito no Jupyter Notebook com Markdown, incluindo todas as etapas anteriores). Será feita uma entrevista com cada integrante do grupo. Incluir na documentação o que foi feito por cada integrante do grupo.

4. Regressão (10 pontos): Nesta etapa, o grupo deve aplicar a regressão linear para fazer a estimativa de algum atributo numérico. Devem ser feitas análises detalhadas dos atributos a serem utilizados, e dos resultados. É importante interpretar os resultados corretamente e tirar conclusões a respeito dos mesmos.

Entrega etapa 4: 02/06/2025 (entregar via GitHub relatório final feito no Jupyter Notebook com Markdown, incluindo todas as etapas anteriores). Será feita uma entrevista com cada integrante do grupo. Incluir na documentação o que foi feito por cada integrante do grupo.

5. Aprendizado Supervisionado e Não-Supervisionado (15 pontos): Nesta entrega final, devem ser incluídos a aplicação de algum algoritmo de aprendizado supervisionado para previsão de um atributo. Todas as etapas de preparação dos dados e análise dos resultados devem ser bem descritas e detalhadas. Além disso, também deve ser utilizado um algoritmo de aprendizado não-supervisionado para gerar agrupamentos, que devem ser analisados e discutidos.

Entrega etapa 5: 22/06/2025 (entregar via GitHub relatório final feito no Jupyter Notebook com Markdown, incluindo todas as etapas anteriores). Será feita uma entrevista com cada integrante do grupo. Incluir na documentação o que foi feito por cada integrante do grupo.

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 não seja da área de tecnologia, e não esteja interessada em como você chegou a tais conhecimentos, mas apenas nos conhecimentos em si. As apresentações serão feitas nos **dias 23/06/2025 e 25/06/2025**.

Observações:

1) Toda a correção será feita pelo Github. Mantenham os *commits* atualizados sempre. O histórico de *commits* será considerado, e entregas de última hora serão penalizadas.

2) Apesar de o trabalho ser em grupo, a avaliação será individual. Em cada entrevista, cada integrante deverá saber responder sobre o projeto.