Insper

Ciência dos dados

2024/1º semestre - ENGENHARIAS

Professora: Maria Kelly Venezuela

PROJETO 3: PREDIÇÃO

Este documento apresenta as premissas do Projeto 3 de Ciência dos Dados.

Conjunto de dados

1) Neste projeto, será permitido uso de microdados, os quais podem ser nacional ou internacional, desde que relacionados a pesquisas governamentais (estaduais ou federais).

2) Será permitido uso de base de dados da plataforma Kaggle, desde que seja recente (a partir de 2020) e seja um conjunto de dados real. Precisa ter fonte. Não pode ser dados simulados.

3) PNADC: A PNAD Contínua foi implantada, experimentalmente, em outubro de 2011 e, a partir de janeiro de 2012, em caráter definitivo, em todo o Território Nacional.
   - Temas e tópicos suplementares pesquisados em trimestres específicos do ano:
     - Educação (**2o trimestre**); e
     - Acesso à televisão e à Internet e posse de telefone móvel celular para uso pessoal (**4o trimestre**).

   Microdados: Disponibilizado no Blackboard um arquivo Jupyter Notebook ensinando como fazer a leitura dos dados da PNAD Contínua dos quatro trimestres de 2023.

4) ENEM: O Exame Nacional do Ensino Médio (Enem) foi instituído em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. Em 2009, o exame aperfeiçoou sua metodologia e passou a ser utilizado como mecanismo de acesso à educação superior.
   Microdados: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

5) Qualquer mínima parte do código de um grupo que seja parecida de um outro grupo, será considerado caso de plágio. Não importa qual seja a motivação que deixou os códigos iguais (ou quase iguais), e nem mesmo se todos entre os grupos sejam muito amigos.

Insper

Ciência dos dados

2024/1º semestre - ENGENHARIAS

Professora: Maria Kelly Venezuela

Objetivo

O principal objetivo do Projeto 3 é **prever uma variável principal em função de demais outras variáveis que podem influenciar em seu comportamento**. Para seu conhecimento, a tabela abaixo mostra como essas variáveis são nomeadas nas áreas de ciência dos dados e estatística.

	Ciência dos dados	**Estatística**
Variável principal	*Target*	Variável resposta ou dependente
Demais variáveis	*Features*	Variáveis explicativas ou independentes

O **tema deverá ser proposto pelo grupo**, assim como a **busca por uma base de dados (obrigatoriamente microdados)** que permita responder alguns interesses levantados no tema escolhido.

O **tema deve deixar claro uma pergunta** e o objetivo deve contemplar obrigatoriamente um dos casos abaixo:

- **Previsão de um rótulo** (nesse caso, o *target* é qualitativo e trata-se de uma classificação). Por exemplo, considerando uma *playlist* de uma pessoa, o *Spotify* deve recomendar uma nova música a essa pessoa.
- **Previsão de uma informação numérica** (nesse caso, o *target* é quantitativo). Por exemplo, considerando o lançamento de um empreendimento imobiliário em uma determinada região, qual o preço ideal de venda desse imóvel a partir de suas características e localização?

Habilidades a serem desenvolvidas no projeto

A condução da análise de dados desse projeto deve mostrar elevado grau de: autonomia dos integrantes do grupo; de liberdade de escolha do tema; e de aprendizado das técnicas mais adequadas.

Algumas técnicas que podem ser utilizadas são: regressão linear; regression tree; random forest regression; *multinomial naive bayes*; regressão logística; *decision tree* e *random forest,* ***entre outras***. Para que este fim possa ser alcançado, os estudantes deverão se aprofundar nas técnicas escolhidas enquanto realizam o projeto.

É importante que o trabalho produza uma conclusão de previsão do *target* escolhido e vá muito além de uma análise exploratória de dados apenas.

Insper

Ciência dos dados

2024/1º semestre - ENGENHARIAS

Professora: Maria Kelly Venezuela

Grupos

O projeto pode ser realizado em grupos de **até quatro alunos** (pode ser individual, dupla ou trio) ou **cinco alunos com rubrica diferenciada** (disponível no Blackboard).

Possíveis técnicas a serem aplicadas

Se escolher um tema cujo objetivo seja prever *target* quantitativa, poderá utilizar técnicas apresentadas em 1. *Regressão*; caso seja prever *target* qualitativa, então poderá utilizar técnicas apresentadas em 2. *Classificadores*. As técnicas a seguir são alguns exemplos, mas outras podem ser encontradas muito bem definidas em bibliotecas do Python.

1. Regressão: quando target é QUANTITATIVO

As técnicas que se prestam a este tipo de análise, por exemplo: regressão linear, *regression tree*, *random forest regression.*

Exemplos:

House Price Prediction using Machine Learning in Python

House Price Prediction With Machine Learning in Python

Create a model to predict house prices using Python

Mastering Predicting House Prices with Python: A Comprehensive Guide

2. Classificadores: quando target é QUALITATIVO

Baseado em todos os dados existentes, classificar em categorias. Técnicas que fazem classificação: *multinomial naive bayes*, regressão logística, *decision tree* e *random forest*.

Exemplos:

Credit Card Fraud Detection Using Machine Learning & Python

Credit Card Fraud Detection in Python

ML | Credit Card Fraud Detection

Estrutura do Projeto

É esperado que o seu projeto seja **autocontido**, ou seja, um leitor que não saiba sobre o que esse projeto se trata deve ser capaz de entender a sua linha de raciocínio. Escreva para um leitor que não possua os mesmos conhecimentos técnicos que você (por exemplo: um aluno do primeiro semestre, que ainda não cursou Ciência dos Dados). Abaixo, apresentamos uma sugestão de estrutura para organizar o seu documento. Se quiser seguir uma estrutura diferente, valide-a primeiro com sua professora.

A proposta do Projeto 3 foi inspirada em um trabalho que constrói alguns modelos preditivos de notas de redação do ENEM 2015 baseados em diversos fatores acerca de um candidato. Acesse-o aqui.

IMPORTANTE: Independente da estrutura adotada, a qualidade do texto produzido é tão importante quanto a análise em si e também será avaliada. Não adianta obter resultados excelentes se eles não forem comunicados de maneira clara. Veja este link para estudar mais a importância de modelos preditivos na área de Machine Learning.

A. Introdução

- Detalhar objetivo escolhido para trabalhar neste projeto juntamente com descrição da base de dados (obrigatoriamente microdados como dito na página 1). Pesquise trabalhos na literatura que discutam o tema escolhido.
- Para trabalhos acadêmicos, acesse https://scholar.google.com.br/. Guarde as referências estudadas para citá-las no seu projeto.

B. Minerando Dados e Características do Dataset

- Se necessário, faça filtro na base de dados tanto de linhas como de colunas em prol do objetivo traçado anteriormente.
- Descreva as variáveis finais que serão utilizadas a partir deste ponto.
- Faça análise descritiva detalhada das variáveis, norteado pelo objetivo do problema. Aqui, é interessante entender como sua variável *target* se comporta cruzada com cada *feature*. Note que ao cruzar duas variáveis, pode obter o cruzamento entre: duas variáveis quantitativas; duas variáveis qualitativas; ou uma de cada tipo. Cada cruzamento irá exigir ferramentas descritivas distintas. A tabela a seguir apresenta algumas ferramentas descritivas vistas no curso:

Ferramentas estatísticas

Duas variáveis qualitativas	Tabela cruzadas (com uso de *normalize* adequado ao problema); Gráficos de barras (empilhados ou *stacked*); entre outras
Duas variáveis quantitativas	Medidas de associação; Gráficos de dispersão; entre outras
Uma variável de cada	Medidas-resumo da variável quantitativa segmentando por rótulo da variável qualitativa; Histograma (ou boxplot) da variável quantitativa segmentando por rótulo da variável qualitativa; entre outras

- *Storytelling* com dados: encontre uma representação gráfica que descreva bem os seus dados e que também favoreça no *storytelling* que pretende fazer ao explicar sua linha de raciocínio às outras pessoas (seja em formato escrito ou em apresentação). Caso tenham interesse em estudar sobre o assunto, vejam neste link a parte Data Visualization. Um trecho com os links dessa seção:

- "**O que estudar**: aprenda sobre **Teoria das Cores** (tem esse vídeo sensacional que explica um pouco em 2 minutos); **Storytelling with Data**, da Cole Nussbaumer (aproveita pra seguir o blog); recomendo também seguir o **blog Nightingale** e participar da comunidade **Dataviz Society**."

C. Modelos de Predição

- Descreva e justifique sua escolha de pelo menos **DUAS** técnicas diferentes de predição. Exemplos de uso de modelos neste trabalho, mas você pode usar outros que fizerem mais sentido para o seu problema. Nesta etapa, ajuste cada modelo preditivo apenas a uma parte da base de dados chamada de treinamento. A validação do modelo está descrita no próximo subitem.

D. Processo e Estatísticas de Validação

- Para os modelos preditivos que foram desenvolvidos no item anterior, é necessário calcular medidas que informam a *performance* de cada modelo ajustado. Assim, para cada modelo preditivo, faça:

- Divida a base de dados na parte treinamento e na parte teste. Use a parte treinamento para estimar cada modelo preditivo. Use ambas as partes (treinamento e teste) para validar seus modelos preditivos.
- Estude as medidas que permitem validar que seu modelo de previsão está funcionando bem. Veja alguns exemplos nos *links* a seguir: link 1, link 2 e link 3 (este apenas se *target* for quantitativo). Escolha **DUAS medidas de *performance*** para os modelos de predição feitos em seu projeto e compare-as após calcular tanto predizer a variável usando os dados de treinamento como para a parte dos dados teste (o mais importante).
- Discuta se essas duas medidas de *performance* se comportam de forma semelhante para as duas partes de dados. Leia o texto disponível aqui para compreender *overfitting* e *underfitting* e refinar senso crítico para discutir sobre as medidas calculadas.
- **Extra: Faça o processo de Validação Cruzada utilizando também 10 ciclos e calcule a *performance* média e desvio padrão das duas medidas de *performance* tanto para a parte treinamento como para a parte teste. Discuta com riqueza de detalhes.**

E. Conclusão

- Faça conclusão final com detalhes levando em consideração todas as interpretações realizadas no decorrer do projeto.

F. Referências Bibliográficas

- Todas as pesquisas feitas e estudadas que foram relevantes para o desenvolvimento devem ser citadas no projeto.

Insper

Ciência dos dados

2024/1º semestre - ENGENHARIAS

Professora: Maria Kelly Venezuela

Cronograma

DATA	Finalização:
02/05 (quinta)	Cadastro do grupo no Blackboard (todos integrantes do grupo): ✔ Quarteto ou Quinteto formado.
05/05 (domingo)	No Blackboard (pelo menos um integrante do grupo): Ter **dados** e **tema** no escopo do projeto (Leitura das seções **Objetivo** e **Estrutura do Projeto: A-Introdução** descritos acima no enunciado do Projeto 3). **Destacando que se você não cumprir com esse deadline, já estará atrasado com o Projeto 3.**
12/05 (domingo)	No Blackboard (pelo menos um integrante do grupo): **Análise exploratória** dos dados pronta (conteúdo visto no início do semestre) (Leitura das seções **Objetivo** e **Estrutura do Projeto: B-Minerando Dados e Características do Dataset** descritos acima no enunciado do Projeto 3). **Destacando que se você não cumprir com esse deadline, já estará atrasado com o Projeto 3.**
21/05 (terça) **DEADLINE FINAL** **Até às 23h59**	No Blackboard (pelo menos um integrante do grupo): **Inferência Estatística** (Leitura das seções **Objetivo** e **Estrutura do Projeto: C, D, E e F** descritos acima no enunciado do Projeto 3). **Destacando que se você não cumprir com esse deadline, já estará atrasado com o Projeto 3.**

Insper

Ciência dos dados

2024/1º semestre - ENGENHARIAS

Professora: Maria Kelly Venezuela

Participação das duas aulas estúdios:

Dias: 16/05 e 21/05

Participação	Dois conceitos a menos	Um conceito a menos	Mantém nota do projeto
Grupo	NENHUMA presença nas aulas estúdios Precisa estar ativo (verdadeiramente presente) e participativo na construção do Projeto 3 durante a aula	UMA presença nas aulas estúdios Precisa estar ativo (verdadeiramente presente) e participativo na construção do Projeto 3 durante a aula	DUAS presenças nas aulas estúdios Precisa estar ativo (verdadeiramente presente) e participativo na construção do Projeto 3 durante a aula