

PROJETO 2

PREDIÇÃO

Este documento apresenta as premissas do Projeto 2 de Ciência dos Dados.

Objetivo

O principal objetivo do Projeto 2 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	<i>Target</i>	Variável resposta ou dependente
Demais variáveis	<i>Features</i>	Variáveis explicativas ou independentes

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

O tema deve deixar claro uma pergunta e se o objetivo contempla:

- **Prever um rótulo** (nesse caso, o *target* é qualitativo e trata-se de uma classificação). Por exemplo, considerando uma *playlist* de uma pessoa, *Spotify* deve ou não recomendar uma nova música a essa pessoa.
- **Prever uma informação numérica** (nesse caso, o *target* é quantitativo). Por exemplo, considerando as algumas características de imóveis de uma determinada região, uma corretora de imóveis deve prever o valor de um novo imóvel que será lançado nessa região.

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: regressão linear; *regression tree*; *random forest regression*; *multinomial naive bayes*; regressão logística; *decision tree* e *random forest*. 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 da análise exploratória.

Grupos

O projeto pode ser realizado em grupos de no máximo 4 alunos (inclusive individual).

Possíveis técnicas a serem aplicadas

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

1. Regressão

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

Exemplos de *datasets* (Estes exemplos não devem ser utilizados):

Predição de preços de casas em King County, Seattle

Predição de por quanto uma casa vai ser vendida

Predição de qual *rating* alguém vai dar para um filme no Netflix

2. Classificadores – extensão do Naïve-Bayes

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

Exemplos de *datasets* (Estes exemplos não devem ser utilizados):

[Porto Seguro - cliente vai acionar o seguro?](#)

[Deteção de fraude no cartão de crédito](#)

[Deteção de fraude financeira](#)

[Predição de se funcionário vai deixar empresa ou não](#)

[Predição de sucesso de um filme](#)

Datasets interessantes para trabalhar no seu projeto

[Lista de todos os datasets do Kaggle](#)

[Alguns datasets disponíveis publicamente](#)

[INEP](#)

Estrutura do Projeto

É esperado que o seu projeto seja autocontido, ou seja, um leitor que não sabe sobre o que ele se trata deve ser capaz de entender a sua linha de raciocínio. Escreva para um leitor que não possui 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 seu professor.

A proposta do Projeto 2 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. 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 <i>normalize</i> adequado ao problema); Gráficos de barras (empilhados ou <i>stacked</i>); entre outras
Duas variáveis quantitativas	Medidas de associação; Gráfico 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 **DOIS** modelos 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.
 - 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 medidas de *performance* para os modelos de predição feitos em seu projeto e compare-as após calcular tanto prever a variável usando os dados de treinamento como para a parte dos dados teste (o mais importante).
 - Discuta se essas duas medidas 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 R2 e RMS 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.

IMPORTANTE:

Neste projeto, pode utilizar bibliotecas prontas disponíveis no Python que façam as modelagens de predição aqui exigidas. Entretanto, **é necessário explicar o que cada modelo de predição faz e também explicar como funciona a biblioteca escolhida.**

Referências

Além dos materiais da disciplina e dos livros-texto, sugerimos as seguintes obras para uma visão geral de:

[Machine Learning / Classificação em Python]:

[DANTAS, D. Comparação Entre Técnicas de Regressão Logística, Árvore de Decisão, Bagging e Random Forest Aplicadas a um Estudo de Concessão de Crédito - Trabalho de Conclusão de Curso. UFPR, Curitiba, 2013 - Capítulo 2](#)

[Introduction to Statistical Learning - capítulos 4 e 10](#)

[Hands-on Machine Learning - notebooks Python. Temos o livro na biblioteca](#)

[Python Data Science Handbook - Capítulo 5](#)

[Python Machine Learning](#)

[Features]:

<https://paulovasconcellos.com.br/como-selecionar-as-melhores-features-para-seu-modelo-de-machine-learning-2e9df83d062a>

<https://dataml.com.br/feature-engineering-para-variaveis-categoricas-target-encoding/>

DICA:

Encontre um *dataset* primeiro de um assunto do seu agrado, depois formule uma pergunta, e daí busque uma técnica condizente.

Não considerar projetos de semestres anteriores como referência para desenvolver este projeto. Isso também é constituído como plágio.

Dimensões de trabalho em grupo

Serão realizadas reuniões de acompanhamento curtas a cada aula. **A presença nas aulas estúdios de cada aluno do grupo (canais por grupo serão criados no Teams) será contada como dedicação ao projeto (veja Figura 1).**

Para ter a nota dada no projeto, é preciso ter contribuições relevantes no Github do grupo.

Importante: se não houver contribuições relevantes de algum membro do grupo no Github, o grupo pode incluir um texto no repositório (no arquivo README ou em outro arquivo texto), explicando o que cada um fez.

Atenção: A nota de trabalho em equipe nunca aumenta a nota geral do projeto. Em outras palavras, não adianta ter A em trabalho em equipe e D em projeto. A nota final ainda será D.

Participação	Um conceito a menos	Mantem nota do projeto
Grupo no Teams	UMA ou NENHUMA presença nas aulas estúdios	DUAS ou TRÊS presenças nas aulas estúdios

Datas de aulas estúdios
17/11; 19/11; 24/11

Figura 1. Avaliação em participação na construção do projeto nas aulas estúdios.

Cronograma

Na tabela a seguir apresentamos uma lista do que deve ser entregue em cada data. **A partir do dia 10/11, todas as entregas** serão feitas via git. Basta que o git esteja atualizado com entregável definido na data abaixo.

Data	Entregável	Meio de entrega
02/11	Kickoff do projeto Usar atendimentos para discutir: proposta de um tema (técnica e <i>dataset</i>) e deixando claro em cada proposta qual o tipo de variável a ser predita (variável <i>target</i>).	Sem entrega.
06/11	Entregável: preencher o formulário com os membros do grupo, o único tema escolhido e o link do github (não precisa ter nenhum <i>commit</i> na hora que for enviar o formulário, só criar o repositório).	https://forms.gle/GprYD5ZoPL2KfwnJ8
10/11	<i>Dataset</i> lido Mínimo esperado: um arquivo do jupyter notebook (.ipynb) com o código que lê o <i>dataset</i> (que também deve ser enviado no git) e realiza limpeza e manipulações necessárias no <i>dataset</i> .	Commit no git até 23:59 do dia 10/11.
13/11	<i>Dataset</i> lido e análise exploratória concluída Mínimo esperado: um arquivo do jupyter notebook (.ipynb) com o código que lê o <i>dataset</i> (que também deve ser enviado no git) e realiza uma análise exploratória inicial.	Commit no git até 23:59 do dia 13/11.
17/11 Aula studio 1	Algoritmo gera alguma resposta (<i>check</i> em aula + <i>commit</i> no git). Mínimo esperado: aplica a técnica escolhida e obtém algum resultado, mesmo que ruim.	Commit no git até 23:59 do dia 17/11.
19/11 Aula studio 2	Entrega dos resultados. Mínimo esperado: resultados prontos (ou no máximo faltando algum ajuste).	Commit no git até 23:59 do dia 19/11.
24/11 Aula studio 3	Entrega final: <ul style="list-style-type: none"> Relatório (é a versão final do projeto, ou seja, o próprio notebook) com explicação detalhada da análise, conclusões e referências para fundamentação teórica. Um arquivo README explicando o que são os arquivos contidos no repositório (especialmente qual é o arquivo contendo o relatório final, caso exista mais de um arquivo). 	Commit no git até 23:59 do dia 24/11.

Rubricas

Veja a tabela com a rubrica geral para o projeto e para o trabalho em grupo.

Postada no Blackboard e também no Github.