

TRABALHO DE CONCLUSÃO DE CURSO - TURMA 2020

ORIENTAÇÕES

Prezados alunos e alunas,

Nesse documento estão apresentadas as informações necessárias para o desenvolvimento do seu Trabalho de Conclusão de Curso. O TCC nos cursos de especialização da PUC Minas Virtual é um trabalho interdisciplinar. Nosso propósito é consolidar os conhecimentos aprendidos no curso, dando ao aluno mais uma oportunidade de colocá-los em prática em um contexto de trabalho.

No TCC você deverá desenvolver um projeto de Ciência de Dados passando por várias etapas, desde a definição do problema até a comunicação dos resultados.

O contexto e problema a ser abordado no TCC deve ser escolhido pelo aluno. Dessa forma, espera-se que os conhecimentos possam ser aplicados em um projeto alinhado com os interesses do aluno, mas que seja um problema REAL e RELEVANTE para sua organização ou para a sociedade. Nesse ponto, é importante ressaltar que, mesmo tendo a possibilidade de escolha do tema, o aluno deverá observar cuidadosamente e cumprir o conjunto de requisitos e as restrições técnicas que fazem parte deste trabalho, descritos no item **Escopo do Trabalho**. Uma observação importante: o TCC que você está desenvolvendo é um trabalho de Ciência de Dados, e não um trabalho de BI ou de Inteligência Artificial. Tenha isso em mente ao pensar em seu tema.

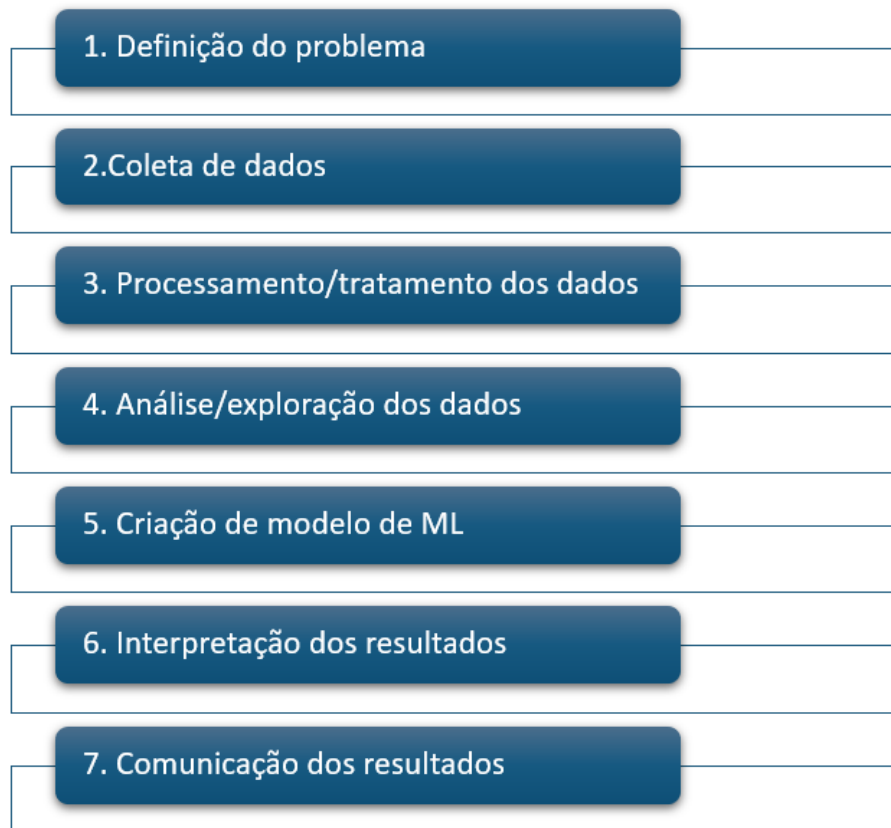
Os itens que devem ser produzidos para o fechamento da disciplina são descritos no item **O que deve ser entregue?** que fica ao final deste documento.

O trabalho deve ser feito individualmente.

Em nosso modelo de especialização, não há a orientação acadêmica/metodológica através de um professor orientador. Você irá elaborar um relatório técnico (conforme template disponibilizado no Canvas). Em caso de dúvidas sobre as regras ou procedimentos do TCC, o aluno deve enviar uma mensagem no Fórum de Discussões da disciplina. Desta forma, espera-se que este espaço sirva de orientação para todos os alunos, uma vez que todos terão acesso às perguntas realizadas pelos colegas e os esclarecimentos sobre elas.

ESCOPO DO TRABALHO

O diagrama a seguir mostra as etapas do desenvolvimento do seu trabalho.



1. Definição do Problema

No seu TCC, primeiramente você deverá escolher um **PROBLEMA REAL** de seu interesse em qualquer área: transporte, economia, consumo, educação, saúde, etc. Você tem liberdade para escolher o tema que achar mais interessante, mas é importante que esse seja um problema **RELEVANTE** para sua organização ou para a sociedade. É muito importante que nessa primeira etapa você defina muito bem a pergunta que o seu trabalho vai responder. O problema/pergunta que você deseja resolver através de dados precisa ficar muito claro para as pessoas que lerão o seu relatório. Uma dica é utilizar o método dos 5 W's¹: why (por que), who (quem), what (o que), when (quando) e where (onde).

2. Coleta de Dados

A partir da clara definição do seu problema, você deve agora buscar os dados que ajudem a responder a sua pergunta. Como sugestão para obtenção de *datasets*, você pode buscar em sites

¹ Para mais informações sobre o método dos 5 Ws leia o artigo do Medium "[Afimal, como se desenvolve um projeto de Data Science?](#)" e o "[Remember de 5 W's](#)". Vale também dar uma olhada no artigo "[How to define your Machine Learning Problem](#)"

como o Kaggle, o famoso repositório de dados da Universidade da Califórnia em Irvine ([UCI](#)), os dados abertos do governo federal ([dados.gov.br](#)) ou de governos de outros países (EUA – [data.gov](#), Reino Unido – [data.gov.uk](#)). O artigo “[Best Public Datasets for Machine Learning and Data Science](#)” dá excelentes dicas de repositórios de dados.

Para o seu TCC, você **deverá utilizar múltiplas fontes de dados**. Ou seja, **você deve usar mais de um dataset de fontes diferentes, integrando-os, se possível, em uma única fonte de dados**. Esse dataset resultante deve ter no mínimo mil registros, sem utilização de oversampling. Recomendamos fortemente que você utilize técnicas de Recuperação de Informação para obter dados na Web e assim enriquecer sua coleta. Essa recuperação pode ser feita utilizando ferramentas como o KNIME, por exemplo, ou através da biblioteca [Beautiful Soup](#) do Python.

Observação 1: caso você queira utilizar dados da empresa onde você trabalha, sugerimos que você remova ou camufle qualquer dado que identifique pessoas ou organizações.

Observação 2: você pode coletar tweets para criar seu próprio dataset, mas apenas se estes tweets forem enriquecidos com outros dados ou se forem utilizados para enriquecer outros datasets.

Observação 3: alguns provedores de informação, podem prover datasets divididos de forma temporal (por ano, meses, etc...). Ao avaliar o seu trabalho, esses datasets serão considerados como um único dataset.

Observação 4: datasets já exaustivamente utilizados pela comunidade de Ciência de Dados e Machine Learning (ex.: IMDB, MNIST, Goodreads, Titanic, Iris, etc...), não poderão ser utilizados de forma simples (sem enriquecimento).

3. Processamento/Tratamento de Dados

Após a obtenção dos seus dados, é momento de tratá-los para que eles possam ser analisados. Essa etapa é importantíssima e você vai gastar boa parte do seu tempo aqui. Aqui você deve extrair, limpar e transformar os dados que serão utilizados. Trate dados ausentes ou duplicados, inconsistências de dados, etc. Lembre-se que você deverá utilizar múltiplas fontes de dados. É importante que você justifique suas decisões e escolhas (exclusão de registros, imputação de valores médios em *missing values*, etc...).

4. Análise e Exploração dos Dados

Nessa etapa você começará a explorar seus dados de uma forma mais analítica, tentando elaborar ideias, levantar hipóteses e começando a identificar padrões em seus dados. Talvez você sinta a necessidade de voltar em passos anteriores, obter mais dados e tratá-los para conseguir responder ao problema proposto. Use e abuse de ferramentas estatísticas consistentes como testes de hipóteses e intervalos de confiança. Plote gráficos que te ajudem a obter insights interessantes: desde os mais simples até gráficos mais sofisticados como boxplots, mapas de calor, etc. Aqui o uso do Python e/ou R e suas poderosas bibliotecas gráficas (Matplotlib, Seaborn, ggPlot2, etc).

5. Criação, treinamento, aplicação e avaliação de Modelos de Machine Learning

Em seu TCC você deve, obrigatoriamente, aplicar algum modelo de Machine Learning (aprendizado supervisionado, não-supervisionado ou por reforço) para fazer classificações, identificar padrões, fazer previsões ou agrupar dados. **ESCOLHA** e **JUSTIFIQUE** o modelo de algoritmo mais adequado para o seu problema. Embora você possa utilizar o KNIME para testar protótipos do seu modelo de ML, encorajamos você a fazer seus modelos em Python ou R. Você deve testar **no mínimo 3** algoritmos, justificando sua escolha. Explique o funcionamento de cada algoritmo e suas vantagens/desvantagens. Por exemplo, se você vai fazer uma classificação, teste algoritmos como Naive-bayes, o KNN e o Support-vector Machines. Se vai fazer uma análise de séries temporais, use o ARIMA, o LSTM e algum método de Exponential Smoothing. A palavra-chave aqui é comparar os resultados para se escolher o método que melhor se encaixa no seu problema. Utilize estratégias para otimizar o seu modelo (*model tuning*).

Observação 1: é EXTREMAMENTE recomendável que você utilize validação cruzada (cross-validation).

Observação 2: escolha métricas adequadas para avaliar o seu modelo.

6. Interpretação dos Resultados

Nessa etapa você deve interpretar os resultados obtidos na análise e exploração de dados e também interpretar os resultados da aplicação dos algoritmos de Machine Learning, descobrindo insights importantes para responder o problema proposto.

7. Apresentação dos Resultados

Até esse ponto, seu TCC foi bastante técnico. É importante que você detalhe cada etapa do seu trabalho o máximo possível, de forma que o leitor do seu trabalho consiga reproduzi-lo com certa facilidade.

Agora pense que você vai apresentar seus resultados para uma pessoa ou para um grupo de pessoas que não entende nada da parte técnica. Atribuem a Einstein a seguinte frase: *se você não consegue explicar algo de forma simples, você não entendeu suficientemente bem*. Agora é o momento de você transmitir os resultados de forma simples, mas que possibilite que sua audiência entenda o problema e possa tomar a melhor decisão. Monte um incrível *dashboard*, use e abuse de gráficos, tabelas e, principalmente, de sua criatividade, para comunicar seus resultados da forma mais efetiva e simples possível. Um Cientista de Dados também deve ser um bom contador de histórias².

FERRAMENTAS PERMITIDAS

Para o desenvolvimento do trabalho de conclusão, será permitido utilizar apenas o Python, R e o Knime para manipulação, análise e visualização de dados, com preferência pela utilização pelo Python e suas bibliotecas.

² O livro “Storytelling com Dados” de Cole Knafllic é uma excelente introdução ao tema de apresentação de resultados.

USO DE UM MODELO CANVAS PARA REGISTRAR SEU WORKFLOW

Louis Dorard e Jasmine Vasandani desenvolveram modelos Canvas, baseados no consagrado *Business Model Canvas*, para ajudá-los em seus projetos de Ciência de Dados e *Machine Learning*. Essa é uma boa maneira de programar as etapas do seu projeto e de apresentar um resumo do que foi feito. Você pode obter mais informações sobre o modelo proposto por Louis Dorard em sua página (clique [aqui](#)) e sobre o modelo desenvolvido por Jasmine Vasandani em um artigo publicado no *Towards Data Science* (clique [aqui](#)).

ENTREGA DO TRABALHO

Para realizar a entrega do trabalho, você deve postar o seu TCC conforme o template disponibilizado, em formato PDF. Os links para o vídeo no Youtube e para o repositório contendo dados e demais arquivos (scripts, etc) devem estar contidos nesse documento.

Após avaliação das entregas postadas no Canvas, os professores da banca irão indicar os alunos aptos para a apresentação final do TCC.

O QUE DEVE SER ENTREGUE?

- Relatório conforme template disponibilizado (em formato PDF). Lembre-se de detalhar o máximo possível. Informe as ferramentas utilizadas, onde e quando coletou os dados
- Link para vídeo no Youtube. Esse vídeo terá tempo máximo de 5 minutos e deverá apresentar de forma sucinta o seu projeto, desde a definição do problema, a obtenção e tratamento dos dados, até a apresentação dos resultados.
- Endereço do repositório (de preferência Github, mas também podem ser utilizados One Drive, Google Drive, Dropbox, etc...) contendo os scripts desenvolvidos e os dados utilizados e gerados em seu trabalho.
- Tanto o link para o vídeo no Youtube quanto o endereço do repositório devem constar no relatório
- A não entrega de algum desses itens, tornará o trabalho automaticamente reprovado.

APRESENTAÇÃO DO TRABALHO PARA A BANCA

- Você deve preparar um conjunto de slides no PowerPoint (ou algum software equivalente) para apresentar seu trabalho para a banca de professores. Geralmente o tempo de apresentação é de 15 a 20 minutos.

DÚVIDAS?

Nosso objetivo foi disponibilizar todos os materiais necessários para a execução do trabalho. Entretanto, entendemos que dúvidas podem surgir. Neste caso, mande uma mensagem para a gente no fórum de discussão. Bom trabalho!