**A P2 é continuação do trabalho da P1. Como a idéia é incrementar e melhorar o que foi feito na P1, TODAS AS REGRAS DESCRITAS PARA A P1 VALEM PARA A P2. A seguir, seguem as regras adicionais ou regras enfatizadas:**

Descrever a base de dados:

* Fazer uma descrição coluna por coluna da base (Não apenas escrever em linhas gerais)
* Fazer plotagens em relação aos dados (não basta dizer só que fez. Tem que estar no documento/código)

Trabalhos Relacionados

* Vem depois da Introdução e antes da Metodologia
* Trabalhos relacionados (assim como a base de dados) devem conter suas respectivas referências completas em formatação ABNT

Metodologia

* Projetar e colocar o fluxograma das etapas realizadas em seu trabalho
* Especificar ítem por ítem de cada subetapa realizada das etapas de pré-processamento
* A descrição das ferramentas é focada apenas nas ferramentas, falando o que elas são e as técnicas usadas delas. Nada de explicar o método (passo a passo) em si aqui nesse ítem.
* Explicar porque realizou cada subetapa (exemplo, explicar porque removeu dados faltantes, baseado em que análise foi escolhido o processo de remoção, qual foi o novo resultado através de números e plotagens)
* Se fez integração de bases, explicar porque foi necessário e o processo.
* Ao normalizar (se normalizou) ou aplicar transformações estatísticas, explicar porque utilizou as técnicas escolhidas comparando o antes com o depois dos dados transformados através de gráficos (histogramas, boxplots, mapas de calor, etc). Se não aplicar, justificar porque não achou necessário.
* Se reduzir colunas, as colunas não devem ser reduzidas aleatoriamente. Deve-se escolher critérios baseados na análise dos dados (exemplo, matriz de correlação, técnicas foward ou backward ou quaisquer algoritmo de seleção de características). Deve-se justificar as suas escolhas. Se não aplicar, justificar porque não achou necessário.
* Ao fazer a divisão da base de dados, explicar como fez e justificar porque fez da forma escolhida
* Utilizar modelos de machine learning sobre diferentes pipelines de pré-processamento

Desenvolvimento

* QUERO TODO O CÓDIGO DESENVOLVIDO (COMENTADO), explicando o que cada código faz. O código deve estar em ipynb, puxando o dataset da url de download e com as saídas dos passos já executados no próprio arquivo
* Frisando, as saídas devem estar geradas nas células.

Resultados

* Comparar os resultados antes e depois das diferentes abordagens de preprocessamento aplicadas (exemplo, antes de tirar tal coluna, depois, antes de normalizar, depois e opor ai vai) através de métricas de avaliação. Justicar a escolha de cada uma das métricas usadas (tem que fazer sentido com o seu problema).
* Gerar além das explicações, tabelas comparativas (nas bases de treino, validação e testes) e gráficos (curvas de precisão-recall, AUC) sobre os resultados encontrados.
* Comparar a eficiencia do modelo nas bases de treino, validação e teste para verificar a ocorrência de underffiting e overffiting.

Conclusão

Mesmas da P1

* REFERENCIAS
* MESMO PROCESSO DA P1. TRABALHOS NAO REFERENCIADOS SERÃO PENALIZADOS