**UNIVERSIDADE FEDERAL DA BAHIA**

**COMPONENTE:** ACCS: Oficina de Projetos em Inteligência Artificial

**Professores:** Barbara Coelho Neves, Daniela Claro e Ricardo Coutinho

**Equipe 07**

**RELATÓRIO DE PRÉ-PROCESSAMENTO**

Primeiramente, tivemos dificuldades em trabalhar com uma base de dados com tantas variáveis e que por conta de sua alta dimensionalidade, nos impossibilitava de fazer qualquer análise, sendo assim, consideramos a urgência de estudar o dicionário para analisar as informações de cada coluna e definir quais delas seriam necessárias manter e quais poderíamos remover para que pudéssemos ter uma base alinhada aos nossos objetivos.

As informações que analisamos no dicionário, nos possibilitou chegar às seguintes conclusões:

* As variáveis com dados sobre números de unidades hospitalares e números de telefones dessas unidades foram descartadas por serem informações irrelevantes para a análise.
* Definimos a variável CLASSI\_FIN como alvo para os nossos modelos de predição por conter dados de classificação final para a investigação de suspeita de doença.
* As variáveis com dados de data de investigação, datas de exames, data de internação e coletas de isolamento do paciente foram descartadas, considerando que haviam outras variáveis que continham os resultados desses exames e coletas, sendo essas informações úteis para o modelo que queríamos atingir.
* As variáveis de sintomas foram mantidas e decidimos também descartar as variáveis que continham dados de acompanhamento do paciente após o diagnóstico por considerarmos que não seriam informações relevantes para a etapa do projeto.

Quando contabilizamos a quantidade de colunas que havia no arquivo da base e comparamos com a quantidade de variáveis com informações no dicionário, percebemos que havia uma diferença de 63 variáveis a mais no arquivo do que havia sido informado no dicionário. Decidimos remover essas variáveis excedentes por considerarmos impróprio trabalhar com dados dos quais não teríamos informações e consequentemente não conseguiríamos analisá-los.

Com a remoção de todas as colunas que consideramos descartáveis, nossa base de dados ficou com apenas 40 colunas e ao verificarmos a nulidade desse novo conjunto, descobrimos que cada coluna de sintomas estavam com 79.000 instâncias nulas e cada uma das colunas com informações sobre exames excediam mais de 400.000 instâncias com nulidade. Para as variáveis de sintomas consideramos preencher a nulidade com o valor da moda, considerando também que a quantidade de valores nulos em cada uma era inferior a metade do número total de instâncias e para as outras colunas, consideramos que seria mais viável fazer a imputação desses dados ausentes com o KNN Imputer, porém por problemas de desfalques na equipe e por problemas em imputar conjuntos de valores ausentes tão grandes, decidimos descartar essa ideia e trabalhar apenas com a substituição de valores ausentes por valores de moda.

Decidimos fazer testes com quatro variações da base pré-processada para rodar nos algoritmos de aprendizado de máquina e definirmos qual apresenta a melhor estratégia de pré-processamento em relação ao preenchimento da nulidade:

* Na primeira abordagem, removemos as 79.000 instâncias nulas das variáveis de sintomas e os valores nulos das variáveis HOSPITALIZ, SOROTIPO e das variáveis com informações de exames
* Na segunda abordagem removemos os valores nulos das variáveis de sintomas e preenchemos a nulidade das variáveis HOSPITALIZ, SOROTIPO e das variáveis com informações de exames com os valores de moda.
* Na terceira abordagem, preenchemos a nulidade de todas as variáveis com os valores da moda
* Na quarta abordagem, preenchemos apenas a nulidade das variáveis de sintomas e removemos a nulidade das variáveis HOSPITALIZ, SOROTIPO e das variáveis com informações de exames

Utilizamos as bibliotecas Pandas, Matplotlib e Seaborn para criar gráficos de correlações com a coluna alvo e percebemos que haviam altas correlações positivas entre essa coluna e as colunas PLAQ\_MENOR, LACO\_N, HEMARTURA e PETEQUIAS e correlações negativas com as colunas EPISTAXE e PLASMATICO, porém, só quando estávamos em uma etapa mais avançada do pré processamento, percebemos que na verdade todas essas colunas tinham altas correlações com o alvo por terem poucos dados não nulos, o que poderia significar que os índices de alta correlação dessas variáveis não eram de fato confiáveis e ao gerar gráfico boxplot, percebemos que os dados da base estavam muito concentrados e por conta disso havia poucos valores outliers

Considerando que a maioria das colunas da nossa base eram variáveis qualitativas nominais e que por isso havia a necessidade de fazer o tratamento desses dados propriamente para evitar resultados equivocados dos testes nos modelos de predição, decidimos codifica-los com o codificador LabelEncoder para que os dados do tipo float fossem transformados em números inteiros, porém, essa transformação foi apenas eficiente para as variáveis de sintomas que tinham apenas duas categorias e assim esses números se tornavam categorias binárias, mas para as demais variáveis que continham mais de duas categorias, houve a necessidade de utilizar o codificador One-Hot Encoder para transformá-las de forma que cada coluna se divide-se em espaços vetoriais contendo apenas categorias binárias e não induzisse os modelos de predição a interpretarem esses valores como ordinais.

**Membros:**

Ana Clara Almeida Moreira **- Contribuição:** Pré-processamento

Daniel Oliveira Santiago da Silva **- Contribuição:** Pré-processamento

Emily Santos Sancho **- Contribuição:** Gestão de Projeto

Felipe Carvalho Goes **- Contribuição:** Aprendizado de máquina e Pré-processamento

João Vitor Moreira de Jesus **- Contribuição:** Pré-processamento