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

O primeiro passo do nosso pré-processamento consistiu em estudar o dicionário, pois o grande número de instâncias e alta dimensionalidade nos impossibilitou de fazer qualquer análise direta sobre os dados. A partir da análise do dicionário, verificamos as informações de cada coluna, possibilitando 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 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 falta de poder computacional, 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 e quatro variações da base pré-processada na forma balanceada 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 removemos os atributos com informações de exames, HOSPITALIZ e SOROTIPO
- 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 os atributos HOSPITALIZ, SOROTIPO e as variáveis com informações de exames
- Para as variações da base na forma balanceada, correspondeu manter apenas duas classificações nos modelo de predição: Dengue ou não Dengue.

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.

Testes das variações das bases pré-processadas nos modelos de predição:

- **Percentuais de Acurácias:**

Abordagem	MLP	KNN RS	KNN Cross	Regreg RS	Regreg Cross	Arvore RS	Arvore Cross	Random F RS	Random F Cross
Base1	0.47080	0.38077	0.36228	0.44203	0.44006	0.43817	0.43615	0.44310	0.43929
Base1_1	0.49800	0.43877	0.43668	0.47547	0.47315	0.46577	0.46871	0.47130	0.47430
Base2	0.54690	0.49217	0.39420	0.48797	0.48459	0.51480	0.51236	0.52147	0.51595
Base2_1	0.50560	0.43527	0.41312	0.45823	0.45617	0.47687	0.47944	0.48127	0.48150
Base1 Balanceada	0.58790	0.52823	0.52894	0.56647	0.56287	0.56427	0.56210	0.56407	0.56147
Base1_1 Balanceada	0.66660	0.61687	0.61728	0.66413	0.66103	0.64927	0.65362	0.65273	0.65695
Base2 Balanceada	0.71060	0.66943	0.67294	0.69443	0.69456	0.68620	0.68967	0.69007	0.69341
Base2_1 Balanceada	0.63680	0.58303	0.55742	0.60433	0.59985	0.61177	0.60711	0.61003	0.60727

Abordagem	Media RS	Media Cross	Desvio padrão
Base1	0.44203	0.43772	0.03814750538
Base1_1	0.47130	0.47093	0.01785014099
Base2	0.51480	0.49848	0.05680574413
Base2_1	0.47687	0.46781	0.03177369811
Base1 Balanceada	0.56427	0.56179	0.01661320057
Base1_1 Balanceada	0.65273	0.65529	0.02018871467
Base2 Balanceada	0.69007	0.69154	0.01002314156
Base2_1 Balanceada	0.61003	0.60348	0.0239134082

*testes com a base reduzida para 100.000 istâncias devido tempo de execução e 50 epocas para MLP

Valores de acurácia dos modelos com hiperparâmetros fixos definidos pela professora em distintos cenários de pré-processamento da base

Legenda

RS = Random Split - ou seja, método houldout usando test-train-split do SkLearn

Cross = K-Fold Cross Validation - no caso k=5

- **Variação escolhida:**

Abordagem	MLP	KNN RS	KNN Cross	Regreg RS	Regreg Cross	Arvore RS	Arvore Cross	Random F RS	Random F Cross
Base2 Balancea	0,71060	0,66943	0,67294	0,69443	0,69456	0,68620	0,68967	0,69007	0,69341

Abordagem	Media RS	Media Cross	Desvio padrão
Base2 Balanceada	0.69007	0.69154	0.01002314156

Shape: (549797, 54)

info():

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 549797 entries, 0 to 549796
Data columns (total 54 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   FEBRE            549797 non-null  int64
 1   MIALGIA          549797 non-null  int64
 2   CEFALEIA         549797 non-null  int64
 3   EXANTEMA         549797 non-null  int64
 4   VOMITO           549797 non-null  int64
 5   NAUSEA           549797 non-null  int64
 6   DOR_COSTAS       549797 non-null  int64
 7   CONJUNTVIT       549797 non-null  int64
 8   ARTRITE          549797 non-null  int64
 9   ARTRALGIA        549797 non-null  int64
 10  PETEQUIA_N       549797 non-null  int64
 11  LEUCOPENIA       549797 non-null  int64
 12  LACO             549797 non-null  int64
 13  DOR_RETRO        549797 non-null  int64
 14  DIABETES         549797 non-null  int64
 15  HEMATOLOG        549797 non-null  int64
 16  HEPATOPAT        549797 non-null  int64
 17  RENAL            549797 non-null  int64
 18  HIPERTENSA       549797 non-null  int64
 19  ACIDO_PEPT       549797 non-null  int64
 20  AUTO_IMUNE       549797 non-null  int64
 21  CLASSI_FIN       549797 non-null  float64
 22  HOSPITALIZ       549797 non-null  int64
 23  RES_CHIKS1_1.0   549797 non-null  float64
 24  RES_CHIKS1_2.0   549797 non-null  float64
 25  RES_CHIKS1_3.0   549797 non-null  float64
 26  RES_CHIKS2_1.0   549797 non-null  float64
 27  RES_CHIKS2_2.0   549797 non-null  float64
 28  RES_CHIKS2_3.0   549797 non-null  float64
 29  RESUL_PRNT_1.0   549797 non-null  float64
 30  RESUL_PRNT_2.0   549797 non-null  float64
 31  RESUL_PRNT_3.0   549797 non-null  float64
 32  RESUL_SORO_1.0   549797 non-null  float64
 33  RESUL_SORO_2.0   549797 non-null  float64
 34  RESUL_SORO_3.0   549797 non-null  float64
 35  RESUL_NS1_1.0    549797 non-null  float64
 36  RESUL_NS1_2.0    549797 non-null  float64
 37  RESUL_NS1_3.0    549797 non-null  float64
 38  RESUL_VI_N_1.0   549797 non-null  float64
 39  RESUL_VI_N_2.0   549797 non-null  float64
 40  RESUL_VI_N_3.0   549797 non-null  float64
 41  RESUL_PCR__1.0   549797 non-null  float64
 42  RESUL_PCR__2.0   549797 non-null  float64
 43  RESUL_PCR__3.0   549797 non-null  float64
 44  HISTOPA_N_1.0    549797 non-null  float64
 45  HISTOPA_N_2.0    549797 non-null  float64
 46  HISTOPA_N_3.0    549797 non-null  float64
```

```
 47  IMUNOH_N_1.0    549797 non-null  float64
 48  IMUNOH_N_2.0    549797 non-null  float64
 49  IMUNOH_N_3.0    549797 non-null  float64
 50  SOROTIPO_1.0    549797 non-null  float64
 51  SOROTIPO_2.0    549797 non-null  float64
 52  SOROTIPO_3.0    549797 non-null  float64
 53  SOROTIPO_4.0    549797 non-null  float64
dtypes: float64(32), int64(22)
```

Nunique():

```
FEBRE              2
MIALGIA            2
CEFALEIA           2
EXANTEMA           2
VOMITO             2
NAUSEA             2
DOR_COSTAS         2
CONJUNTVIT         2
ARTRITE            2
ARTRALGIA          2
PETEQUIA_N         2
LEUCOPENIA         2
LACO               2
DOR_RETRO          2
DIABETES           2
HEMATOLOG          2
HEPATOPAT          2
RENAL              2
HIPERTENSA         2
ACIDO_PEPT         2
AUTO_IMUNE         2
CLASSI_FIN         2
HOSPITALIZ         2
RES_CHIKS1_1.0     2
RES_CHIKS1_2.0     2
RES_CHIKS1_3.0     2
RES_CHIKS2_1.0     2
RES_CHIKS2_2.0     2
RES_CHIKS2_3.0     2
RESUL_PRNT_1.0     2
RESUL_PRNT_2.0     2
RESUL_PRNT_3.0     2
RESUL_SORO_1.0     2
RESUL_SORO_2.0     2
RESUL_SORO_3.0     2
RESUL_NS1_1.0      2
RESUL_NS1_2.0      2
RESUL_NS1_3.0      2
RESUL_VI_N_1.0     2
RESUL_VI_N_2.0     2
RESUL_VI_N_3.0     2
```

```
RESUL_PCR__1.0    2
RESUL_PCR__2.0    2
RESUL_PCR__3.0    2
HISTOPA_N_1.0     2
HISTOPA_N_2.0     2
HISTOPA_N_3.0     2
IMUNOH_N_1.0      2
IMUNOH_N_2.0      2
IMUNOH_N_3.0      2
SOROTIPO_1.0      2
SOROTIPO_2.0      2
SOROTIPO_3.0      2
SOROTIPO_4.0      2
```

Balanceamento:

```
CLASSI_FIN
0.0    328645
1.0    221152
```

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