Predição de Infecções Bacterianas: Estratégias de Antibioticoterapia

1st Paulo Henrique Maciel Fraga
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
paulohmacielf@gmail.com

2nd Adriano Veloso
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
adrianov@dcc.ufmg.br

***Abstract*—Bacterial resistance to antibiotics poses a major threat to global health, intensified by the improper use of antibiotics. The lack of effective and rapid diagnostic tools exacerbates the problem, resulting in less efficient treatments and increased spread of bacterial resistance. Therefore, there is an urgent need to develop models that can assist in more precise and restricted antibiotic prescriptions. Thus, the overall objective of this work is to aid in the construction of binary classification models aimed at determining whether a patient is infected by a given bacterium based on clinical and laboratory data To this end, we utilized data from the Hospital das Clínicas at UFMG. In our previous work, we focused on organizing and cleaning the data to extract relevant patient *features* and bacterial *targets* identified through bacterial culture results, enabling the creation of primary datasets that served as inputs and outputs for our models. This process was conducted in collaboration with infectious disease specialists from the hospital through weekly meetings. In this study, we evaluate the performance of three distinct modeling pipelines: a baseline *Gradient Boosting*-based model optimized using the ROC (Receiver Operating Characteristic) AUC (Area Under the Curve) metric with default parameters, a hyperparameter-tuned *Gradient Boosting*-based model using random search, and a combination of Multivariate Imputation by Chained Equations (MICE) with Z-Score Normalization and a Support Vector Machine (SVM) model. We conducted extensive data preprocessing and filtering, and compared the performance of these approaches using K-fold cross-validation. The hyperparameter-tuned *Gradient Boosting*-based model outperformed the other approaches, achieving satisfactory results given the complexity of the problem.**

I. Introdução

Nos últimos anos, a aplicação da Inteligência Artificial (IA) na área da saúde tem demonstrado um potencial significativo para melhorar o diagnóstico e o tratamento de diversas condições médicas (SHANG, 2021; WAN, 2011). Esse avanço tecnológico vem em um momento crucial, em que enfrentamos desafios crescentes relacionados à resistência antimicrobiana e à gestão eficaz de infecções hospitalares. O relatório de 2022 acerca do sistema global de resistência a antimicrobianos e de vigilância de uso da Organização Mundial de Saúde (OMS, em inglês WHO) reporta que altas taxas de infecções resistentes foram documentadas em todos os continentes (WHO, 2022a). Infecções com microrganismos resistentes podem ter consequências diretas de grande impacto, como períodos mais longos de doença, mortalidade aumentada, prolongamento da internação etc., assim como aumento dos gastos associados ao tratamento (WHO, 2015). Por esses e outros motivos, a resistência bacteriana aos antibióticos está entre as dez principais ameaças à saúde global e é um problema de saúde pública urgente, com amplo impacto socioeconômico (WHO, 2022a).

Considerando que o uso extensivo e inadequado de antibióticos contribui para o desenvolvimento de cepas de bactérias resistentes aos tratamentos disponíveis e que existem poucas novas drogas promissoras em desenvolvimento, é de interesse da sociedade como um todo tomar medidas a fim de garantir a continuidade da eficiência dos tratamentos já existentes (WHO, 2015) e um dos objetivos propostos pela OMS é otimizar o uso de medicamentos antimicrobianos na saúde humana (WHO, 2005; 2015).

Idealmente, a prescrição de um tratamento antibiótico deve ser sempre baseada em evidências (WHO, 2005); no entanto, estima-se que metade de todo o uso de antibióticos seja inadequado de alguma forma, seja pelo uso em situações desnecessárias ou pela escolha de um antibiótico com espectro excessivamente amplo (WHO, 2022b). Além disso, a organização relata que as decisões de prescrição de antibióticos são frequentemente realizadas de modo empírico. Existe, então, uma necessidade de recursos simples para guiar e melhorar a qualidade da prescrição empírica de antibióticos globalmente (WHO, 2022b); ao mesmo tempo são necessárias ferramentas de diagnóstico eficazes, rápidas e de baixo custo para orientar o uso ideal de antibióticos na medicina (WHO, 2005). Hospitais são os locais em que se encontram infecções resistentes com maior frequência (WHO, 2015) e, em particular, as infecções em Unidades de Terapia Intensiva (UTIs) representam uma grande preocupação, devido à sua gravidade e à necessidade de tratamento imediato e preciso (KOLLEF, 2021).

Ao prever corretamente o agente infeccioso, é possível determinar os antibióticos mais adequados para o tratamento, contribuindo para uma terapia mais eficaz e personalizada. Além disso, ao prescrever antibióticos de forma mais restrita e precisa, é possível reduzir o risco de desenvolvimento e de disseminação de resistência bacteriana, uma ameaça crescente à saúde pública global e que muitas vezes pode ser fatal para outros pacientes que dividem o mesmo espaço da UTI (KOLLEF, 2021; WHO, 2015).

Assim, a resistência bacteriana aos antibióticos representa uma grande ameaça à saúde global, intensificada pelo uso inadequado de antibióticos. A falta de ferramentas de diagnóstico eficazes e rápidas agrava o problema, resultando em tratamentos menos eficientes e no aumento da disseminação de resistência bacteriana. Portanto, há uma necessidade urgente de desenvolver modelos que possam auxiliar na prescrição mais precisa e restrita de antibióticos, baseados em dados clínicos e laboratoriais.

Dessa maneira, o objetivo geral deste trabalho é auxiliar na construção de modelos de classificação binária, que tem como objetivo classificar se um paciente está ou não infectado por uma dada bactéria ou classificação multiclasse, que tem como objetivo classificar qual bactéria infecta um paciente, a partir de dados clínicos e laboratoriais. Para isso, partiremos de dados diversos de atendimentos e internações do SUS no Hospital das Clínicas da UFMG. Ao longo do POC I, realizamos um processo de organização e limpeza dos dados para que pudéssemos obter tanto *features* do paciente, quanto *targets* de bactérias identificadas através de resultados de culturas bacterianas. Construímos bases de dados primárias e realizamos um trabalho extensivo de engenharia de *features*. Com o dado tratado, testamos uma primeira versão de modelo baseado em *Gradient Boosting*, a partir das *features* já descritas e *targets* de culturas, utilizando como métrica de otimização a ROC (Receiver Operating Characteristic) AUC (Area Under Curve).

Neste trabalho, exploramos um subset menor do problema. Focamos esforços apenas em pacientes com passagem por UTI, ao invés de todo o contexto do hospital, e segundo recomendações médicas, focamos apenas nas 6 bactérias mais comuns nos hospitais: *Acinetobacter baumannii*, *Enterococcus faecalis*, *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* e *Staphylococcus aureus*. Tratamos o problema com duas abordagens distintas: uma classificação binária e classificação multiclasse. Para a primeira, treinamos, para cada bactéria uma das seis bactérias, um modelo que predizia 1, se um paciente está infectado por uma dada bactéria ou 0, caso o paciente não esteja infectado pela bactéria, totalizando 6 modelos para todas as bactérias. Por outro, para a segunda abordagem, treinamos apenas um modelo que retornava qual bactéria está infectando, tratando como um problema de classificação multiclasse.

Realizamos experimentações e trouxemos o resultado de 3 abordagens: A primeira, seguimos com um modelo baseado em *Gradient Boosting* com balanceamento de classes, utilizando como métrica de otimização a ROC (Receiver Operating Characteristic) AUC (Area Under Curve), tal qual aplicada no último trabalho. Na segunda abordagem, utilizamos um *RandomizedSearchCV* para realizar a busca aleatória por hiperparâmetros do *LightGBM Classifier*, modelo descrito anteriormente, otimizando parâmetros como *num_leaves*, *learning_rate* e *n_estimators* para maximizar a qualidade do modelo. Por fim, a terceira abordagem, utiliza para pré-processamento o método *Multivariate Imputation by Chained Equations* (MICE) para lidar com valores ausentes, e é seguida de *ZScore Normalization* para padronizar os dados, em seguida, treina um modelo *Support Vector Machine* (SVM) com probabilidade de predição habilitada.

Para isso, temos como objetivos específicos reuniões semanais com médicos do Hospital das Clínicas, especialistas da área de infectologia do hospital, visando entender os desafios no diagnóstico e tratamento de infecções. Além disso, passamos por processo de tratamento dos dados para que possamos realizar transformações pertinentes e filtrar os dados desejados focando nas bactérias já citadas e em pacientes que tiveram passagem pela UTI.

Com nosso dado filtrado e tratado, partimos para seleção de modelos, treinamento e experimentação com os modelos. Assim, passaremos pelo treinamento e validação dos modelos, utilizando método como *K-fold Cross Validation* e apresentaremos tabelas de métricas sobre a qualidade dos modelos gerados.

As seguintes seções deste trabalho estão divididas em: referencial teórico, em que abordaremos temas fundamentais para o entendimento das atividades desenvolvidas, desenvolvimento, em que abordamos os tópicos apresentação dos datasets primários, treinamento dos modelos e discussão de resultados. Ademais teremos uma sessão de conclusões tiradas do desenvolvimento deste trabalho.

II. Referencial Teórico

A. Features

Os autores (DONG, 2018) nos explicam que em aprendizado de máquina, mineração de dados e análise de dados, uma *feature* ou característica é um atributo ou variável usada para descrever algum aspecto de objetos de dados individuais. Eles exemplificam que features podem incluir idade e cor dos olhos para uma pessoa ou curso e média ponderada para um estudantes. No caso do nosso trabalho, as features referem-se a aferição de sinais vitais e resultados de exames de um paciente no contexto do hospital das clínicas.

Os autores (DONG, 2018) também afirmam que, *features*, variáveis, ou atributos informativas são a base da análise de dados e que são essenciais para descrever os objetos subjacentes e para distinguir diferentes grupos de objetos, sejam esses grupos explícitos ou não. Eles completam que as *features* são essenciais para a criação de modelos preditivos precisos e de fácil interpretação, resultando em bons desempenhos em várias tarefas de análise de dados.

Features podem possuir vários tipos, como categóricas, ordinais e numéricas. Em nosso contexto, após a limpeza dos nossos dado, todas as nossas features são numéricas.

B. LightGBM

O *LightGBM* é um algoritmo desenvolvido por (KE, 2017), projetado visando a eficiência e escalabilidade em árvores de decisão de reforço gradiente (GBDT). Eles explicam que o modelo utiliza de duas técnicas principais: a Amostragem de Um Lado Baseada em Gradiente (GOSS) e o Agrupamento Exclusivo de *Features* (EFB). GOSS melhora a eficiência computacional ao focar em instâncias de dados com grandes

gradientes, que são mais importantes para o aprendizado, excluindo uma porção significativa de instâncias com pequenos gradientes. Enquanto o EFB é método que reduz o número de *features* combinando aquelas que raramente assumem valores diferentes de zero simultaneamente, diminuindo a dimensionalidade e acelerando o processo de treinamento sem afetar significativamente a precisão.

Segundo os autores (KE, 2017), essas características garantem que o tempo de treinamento do *LightGBM* seja muito mais rápido do que os métodos tradicionais de GBDT, mantendo alta precisão. Isso torna o modelo especialmente útil para lidar com grandes conjuntos de dados e dados de alta dimensionalidade. No nosso caso, estamos lidando com um problema de classificação binária, onde a saída do nosso modelo é 0 ou 1 e com um problema de classificação multiclasse, em que a saída é qual das bactérias o paciente está infectado.

C. SVC

O *Support Vector Classifier* (SVC) é uma variação do *Support Vector Machine* (SVM) voltada para problemas de classificação. Conforme descrito por Cortes e Vapnik (1995), o SVC funciona construindo hiperplanos em um espaço multidimensional que separa as diferentes classes de dados de forma ótima. O objetivo é maximizar a margem entre as classes, ou seja, a distância entre o hiperplano e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte.

Para lidar com problemas não linearmente separáveis, o SVC utiliza uma função *kernel*, que mapeia os dados para um espaço dimensional mais alto, onde eles podem ser separados linearmente. Em nosso trabalho, utilizamos o SVC com o *kernel* RBF (Radial Basis Function), que captura relações complexas entre as variáveis.

D. Random Search

O *Random Search* é uma técnica para busca de hiperparâmetros proposta por Bergstra e Bengio (2012). Ao contrário da busca exaustiva, onde todas as combinações possíveis são avaliadas, o *Random Search* seleciona aleatoriamente um subconjunto de combinações para avaliação. Isso permite cobrir uma maior variedade de valores em menos tempo, tornando-o uma alternativa eficiente para problemas com espaços de busca extensos.

No nosso trabalho, utilizamos o *Random Search* para otimizar os hiperparâmetros dos nossos modelos de aprendizado de máquina, como a taxa de aprendizado e o número de estimadores e muitos outros, garantindo que alcancemos um melhor desempenho dos modelos.

E. Multivariate Imputation by Chained Equations (MICE)

O *Multivariate Imputation by Chained Equations* (MICE) é uma técnica amplamente utilizada para lidar com dados ausentes em conjuntos de dados multivariados. Segundo Buck (1960) e van Buuren e Groothuis-Oudshoorn (2011), a técnica funciona modelando cada variável com valores ausentes como uma função das demais variáveis, de forma iterativa. O MICE utiliza modelos estatísticos para prever os valores ausentes, considerando as relações entre as variáveis, e repete o processo em cadeia até que a imputação alcance estabilidade.

No trabalho, utilizamos a implementação do *IterativeImputer* da biblioteca *scikit-learn*, que adota o princípio do MICE.

F. Z-Score Normalization

A normalização por *z-score* é uma técnica de pré-processamento amplamente utilizada para padronizar os dados, assegurando que todas as *features* tenham média igual a 0 e desvio padrão igual a 1. Segundo Han, Kamber e Pei (2011), esse método é especialmente útil em algoritmos sensíveis à escala, como *Support Vector Machines* (SVM), regressão logística e redes neurais, garantindo que nenhuma *feature* com maior amplitude domine as demais.

A transformação é realizada calculando o valor padronizado de cada *feature*. Esse processo consiste em subtrair a média dos valores originais da *feature* e, em seguida, dividir o resultado pelo desvio padrão correspondente. Assim, o valor original da *feature* é ajustado para ter uma média igual a zero e um desvio padrão igual a um, garantindo que todas as *features* sejam comparáveis e contribuam de maneira equilibrada para o modelo. A média e o desvio padrão utilizados são calculados com base no conjunto de treino.

No contexto deste trabalho, aplicamos o *Z-Score Normalization* usando a implementação do *StandardScaler* da biblioteca *scikit-learn*. Etapa que visa padronizar as *features* numéricas relacionadas a sinais vitais e resultados de exames, reduzindo vieses induzidos por diferenças de escala com intuito de melhorar o desempenho dos modelos utilizados.

G. K-fold Cross Validation

A validação cruzada K-fold é um método robusto para avaliar a performance de modelos de aprendizado de máquina. (Fushiki, 2011). Neste método, os dados são divididos em K subconjuntos ou *folds*. Em cada iteração, um dos folds é usado como conjunto de teste, enquanto os K-1 folds restantes são usados como conjunto de treinamento. Esse processo é repetido K vezes, de modo que cada fold seja utilizado uma vez como conjunto de teste. A média das métricas de desempenho ao longo das K iterações é calculada para fornecer uma estimativa mais precisa da performance do modelo. O método é vantajoso porque permite que todos os dados sejam usados tanto para treinamento quanto para teste, reduzindo a variação associada ao uso de um único conjunto de teste No nosso trabalho, utilizamos a validação cruzada K-fold para todas as etapas do treinamento dos modelos.

H. Hyperparameters Tuning

O ajuste de hiperparâmetros (*Hyperparameters Tuning*) é um processo essencial para melhorar o desempenho de modelos de aprendizado de máquina. Segundo Bergstra e Bengio (2012), hiperparâmetros são parâmetros definidos antes do treinamento do modelo, como a taxa de aprendizado, o número de árvores em algoritmos baseados em *boosting*, ou o parâmetro de regularização em modelos de regressão.

Existem diversas técnicas para ajuste de hiperparâmetros, incluindo busca exaustiva (*Grid Search*) e busca aleatória (*Random Search*). Mais recentemente, métodos baseados em otimização Bayesiana e redes neurais também foram introduzidos para lidar com espaços de busca mais complexos. No contexto deste trabalho, combinamos *Random Search* com validação cruzada para selecionar os hiperparâmetros que maximizam a métrica AUC (*Area Under the Curve*).

I. Métricas

Para o treinamento do modelo utilizamos a métrica, *Area Under the Curve* (AUC) que se refere à área sob a curva ROC *(Receiver Operating Characteristic)*, uma métrica comum para avaliar modelos de classificação binária. (SOFAER, 2019) Ela avalia a capacidade do modelo em distinguir entre as classes positivas e negativas. Uma AUC de 1 indica um modelo perfeito, enquanto uma AUC de 0,5 indica um modelo sem discriminação melhor que o acaso. Além disso reportamos métricas como:

- **Sensibilidade (ou recall)**: mede a proporção de verdadeiros positivos (VP) corretamente identificados pelo modelo em relação ao total de casos positivos reais (VP + falsos negativos, FN), indicando a capacidade do modelo de detectar casos positivos.
- **Especificidade**: mensura a proporção de verdadeiros negativos (VN) corretamente identificados em relação ao total de casos negativos reais (VN + falsos positivos, FP), mostrando a capacidade do modelo de identificar corretamente os casos negativos.
- **F1 Score**: É a média harmônica da precisão e da sensibilidade, fornecendo uma única métrica que equilibra ambas. Esta métrica é especialmente útil em cenários com classes desbalanceadas, pois considera tanto a capacidade de identificar casos positivos quanto a exatidão dessas predições.
- **Acurácia (ou accuracy)**: representa a proporção de previsões corretas, verdadeiros positivos (VP) e verdadeiros negativos (VN), em relação ao total de amostras avaliadas, verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). A acurácia é uma métrica global que reflete o quão frequentemente o modelo faz previsões corretas, sendo mais apropriada quando as classes estão balanceadas.

III. Desenvolvimento

A. Apresentação do dataset

Realizamos, ao longo do POC I, um processo de organização e limpeza dos dados para que pudéssemos obter tanto *features* do paciente, quanto *targets* de bactérias identificadas através de resultados de culturas bacterianas. Construímos bases de dados primárias e realizaremos um trabalho extensivo de engenharia de *features*.

Inicialmente existiam diversas tabelas do Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG). O prontuário foi utilizado como a principal chave de conexão entre as bases, uma vez que ele representa uma pessoa no contexto de dados do hospital e, independente de quantas vezes um paciente é atendido, ele sempre manterá o mesmo número de prontuário. Assim, é possível acompanhar a trajetória de um indivíduo por meio deste identificador.

TABLE I: Base de atendimentos

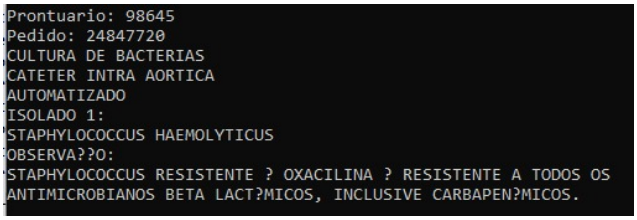
Coluna	Valor Absoluto (%)
prontuarios	53790
data atendimento	
min	2016-01-01 01:49:00
max	2020-12-31 20:51:00
convenio	
sus - internacao	86447 (100.00)
sexo[paciente]	
F	49318 (57.05)
M	37127 (42.95)
I	2 (0.00)
origem atendimento	
hosp.das clinicas da ufmg	64726 (74.87)
dermatologia - laudo	8702 (10.07)
urgencia - hsg inat 09/09/19	4749 (5.49)
pa urgência/emergencia hc	3455 (4.00)
neonatologia - 4. andar hc	1985 (2.30)
hemodinamica	1688 (1.95)
maternidade 4º andar	501 (0.58)
outros	641 (0.75)
servico	
clinica geral	21261 (24.59)
ginecologia	15328 (17.73)
pediatria	7544 (8.73)
cardiologia clinica	5587 (6.46)
oftalmologia	5584 (6.46)
cirurgia geral	4307 (4.98)
pediatria neonatologia	4269 (4.94)
gastroenterologia	4267 (4.94)
urologia	2982 (3.45)
ortopedia e traumatologia	1641 (1.90)
outros	13677 (15.82)
tipo acomodacao[leito]	
enfermaria	54035 (62.51)
enfermaria 3 leitos-hospital d	13718 (15.87)
pronto socorro internado	6478 (7.49)
observacao	6237 (7.21)
apartamento simples	3608 (4.17)
uti - neonatal	861 (1.00)
uti - adulto	674 (0.78)
uti - cardiologica	628 (0.73)
uti - infantil	207 (0.24)

A base de atendimento utilizada foi gerada a partir da junção tabelas que continham dados de Atendimentos, Evoluções, Exames Laboratoriais e Sinais Vitais, que continham entre seus dados o prontuário de um paciente e sua data de atendimento. A Tabela I trás uma visão de algumas colunas da base de atendimento. Elas tratam de atendimentos de internações feitas via SUS no Hospital das Clínicas da UFMG, no período de Janeiro de 2016 até Dezembro de 2020, recorte que foi feito a fim de evitar o ruído gerado pela COVID no sistema público de saúde.

As *features* utilizadas como X dos nossos modelos representam resultados de exames laboratoriais e coletas de sinais vitais de pacientes — como pressão arterial, frequência cardíaca, saturação de oxigênio, frequência respiratória, etc em datas específicas de atendimento. Nosso objetivo é tentar prever a bactéria identificada no resultado de cultura bacteriana do pa-

```
Prontuario: 1426
Pedido: 12826707
URINA
ISOLADO 1: ESCHERICHIA COLI
ISOLADO 2: KLEBSIELLA PNEUMONIAE
100.000 UFC/ML
ANTIBIOTICO                          ISOLADO 1  ISOLADO 2
AMOXACILINA+ AC.CLAVULANICO              S          S
AMICACINA                                S          S
AMPICILINA                               R          R
CEFTAZIDIMA                              S          S
CEFALOTINA                               R          S
CIPROFLOXACINA                           R          R
CEFEPIME                                 S          S
CEFTRIAXONA                              S          S
ERTAPENEM                                S          S
GENTAMICINA                              R          R
MEROPENEM                                S          S
NITROFURANTO?NA                          R          I
NORFLOXACINA                             R          R
PIPERACILINA+TAZOBACTAM                  S          S
SULFAMETOXAZOL/TRIMETOPRIM               R          R
```

Fig. 1: Resultado de cultura bacteriana (1).



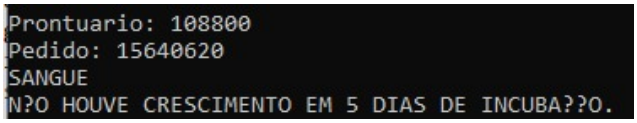
```
Prontuario: 98645
Pedido: 24847720
CULTURA DE BACTERIAS
CATETER INTRA AORTICA
AUTOMATIZADO
ISOLADO 1:
STAPHYLOCOCCUS HAEMOLYTICUS
OBSERVA??O:
STAPHYLOCOCCUS RESISTENTE ? OXACILINA ? RESISTENTE A TODOS OS
ANTIMICROBIANOS BETA LACT?MICOS, INCLUSIVE CARBAPEN?MICOS.
```

Fig. 2: Resultado de cultura bacteriana (2).

ciente em uma determinada data, sendo a presença ou ausência de bactéria o *target* do modelo. Para nosso Y selecionamos resultados de exames do tipo *UROC, HEMOC e CULBAC*, que representavam resultados de culturas bacterianas, exemplos desses resultados podem ser vistos na figuras 1, 2 e 3. Algumas culturas, além da espécie isolada, trazem o perfil de sensibilidade a antibióticos da bactéria identificada, podendo elas serem sensíveis ao antibiótico (S), resistentes ao antibiótico (R) e Intermediário (I), que em que a bactéria é suscetível ao antibiótico, porém não necessariamente é possível realizar o tratamento com este medicamento.

O Gráfico 4 traz informações sobre a distribuição das bactérias encontradas na base de *target*, nela temos 88.238 resultados de culturas bacterianas para 28.012 pacientes. Além disso, da lista de bactérias existentes, 85 espécies foram identificadas nos resultados. É possível perceber que as mais frequentes são: *Escherichia Coli* com 5.513 resultados positivos, representando 6.25% da base, seguida pela, *Klebsiella Pneumoniae* com 2.501 (2.83%) resultados e a *Pseudomonas Aeruginosa* 2.070 (2.35%).

Realizamos o filtro da base de atendimento considerando



```
Prontuario: 108800
Pedido: 15640620
SANGUE
N?O HOUVE CRESCIMENTO EM 5 DIAS DE INCUBA??O.
```

Fig. 3: Resultado de cultura bacteriana (3).

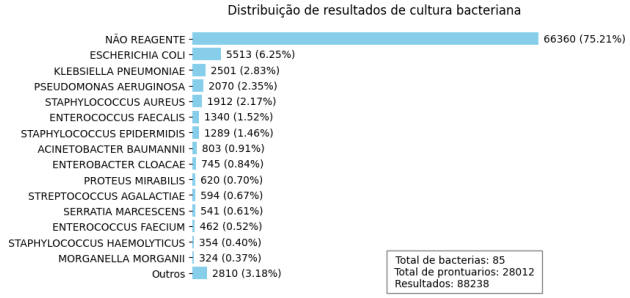


Fig. 4: Distribuição de resultados de cultura do target.

apenas prontuários que possuíam algum tipo de resultado de cultura bacteriana, fossem eles positivos ou negativos. Realizamos também um extenso processo de limpeza de *outliers* e tratamento dos dados, estatísticas sobre a tabela de features pode ser encontrado na Tabela X. Além disso, realizamos o enriquecimento das *features* do paciente por um extenso trabalho de engenharia de features, cujas estatísticas podem ser encontradas nas tabelas XI, XII, XIII . Para a composição da tabela de *features* dos pacientes, foi utilizado o par prontuário e data de atendimento para mapear resultados de exames que fossem da mesma data da coleta da cultura bacteriana.

Além disso, nessa etapa do trabalho focamos em um *subset* específico de bactérias as 6 mais comuns no ambiente hospitalar, *Acinetobacter baumannii*, *Enterococcus faecalis*, *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* e *Staphylococcus aureus*, de pacientes com alguma passagem pela UTI. Dessa maneira, passamos para a etapa de treinamento de modelos e construção de pipelines.

Outras linhas de pesquisa trabalharam com um modelo de *embedding* multilíngue apara gerar representações vetoriais dos textos médicos de evolução de pacientes. O modelo foi configurado para processar sequências de tokens, e um prompt detalhado, que simula um painel de especialistas médicos analisando registros de pacientes em UTI para determinar a provável bactéria causadora da infecção, foi utilizado para orientar a geração dos *embeddings*. Os textos fornecidos foram então codificados em *embeddings* normalizados e agregamos às *features* já descritas anteriormente.

B. Treinamento de modelos

1) LightGBM Classifier: No primeiro pipeline, utilizamos a implementação do *LightGBM Classifier*, um método de *Gradient Boosting* para a classificação, projetada para lidar com grandes conjuntos de dados e alta dimensionalidade. Em nosso pipeline, o modelo é configurado com o parâmetro *class_weight="balanced"*, que ajusta automaticamente o peso das classes inversamente proporcional à sua frequência no conjunto de dados, para lidarmos com o desbalanceadas das classes em nosso dataset. Foi-se utilizado os parâmetros *default* do *LightGBM Classifier*, com *num_leaves=31*, que define o número máximo de folhas em cada árvore , *max_depth=-1* que significa que árvore não tem limite de profundidade, esses parâmetros estão diretamente ligados com a complexidade do

model. Além disso, o modelo utiliza um *learning_rate=0.1*, que controla a distância o tamanho do passo do algoritmo por iteração e o número de iterações do *boosting* é definido como *n_estimators=100*. Durante o treinamento o método, com base nos conjuntos de treino *X_train* e *Y_train*, otimiza os hiperparâmetros internos visando minimizar a função de perda baseada no *ROC AUC*, padrão para classificação em LightGBM.

Além de parâmetros já descritos, outros parâmetros *default* da implementação incluem: *min_child_samples=20*, que é o número mínimo de amostras necessárias em uma folha, uma vez que poucas folhas podem gerar *overfitting*; *subsample=1.0*, que define a proporção de amostras de treinamento usadas para construir cada árvore; *colsample_bytree=1.0*, que define a proporção de *features* usadas para construir cada árvore, *reg_alpha=0.0* e *reg_lambda=0.0*, que penalizam grandes pesos nas árvores de decisão. Cada uma dessas variáveis possui objetivos distintos, entre eles estão: reduzir o *overfitting*, melhorar a generalização e interpretabilidade do modelo. O compilado dos parâmetros pode ser visto na tabela II.

TABLE II: Hiperparâmetros do modelo LightGBM

Parâmetro	Valor
num_leaves	31
max_depth	-1
learning_rate	0.1
n_estimators	100
min_child_samples	20
subsample	1.0
colsample_bytree	1.0
reg_alpha	0.0
reg_lambda	0.0

O modelo foi treinado para a versão binária do problema e para a versão multiclasse. Assim, durante o treinamento utilizamos do método de *k-fold Cross Validation*, com *k = 5*, e reportamos a média das 5 *folds* para métricas que incluem: número de instâncias, *Training ROCAUC*, *Validation ROCAUC*, *Validation Sensitivity/Recall*, *Validation Specificity*, *Validation F1-Score* e *Accuracy*. Lembrando que na versão multiclasse temos um modelo por bactéria onde a saída é 0, se o paciente não está infectado pela bactéria, ou 1, caso esteja. Enquanto no modelo de classificação multiclasse tem como saída qual bactéria está infectando o paciente. Os resultados do modelo binário podem ser vistos na tabela IV e os do modelo multiclasse na tabela VII.

2) Random Search para Hyperparameters Tuning do LightGBM Classifier: No segundo pipeline, realizamos *tuning* dos hiperparâmetros do *LightGBM Classifier*, modelo descrito na seção anterior. Para isso, utilizamos um algoritmo de busca aleatória a partir da implementação do *RandomizedSearchCV* do *sklearn*, com ele foi realizada a busca aleatória por distintos conjuntos de hiperparâmetros citados na sessão anterior, considerando os seguintes intervalos:: *num_leaves* entre 20 e 150, *max_depth* entre 3 e 12, *learning_rate* variando de 0.01 a 0.31, *n_estimators* entre 50 e 500, *min_child_samples* entre 5 e 100, *subsample* e *colsample_bytree* variando de 0.6 a 1.0, e os parâmetros de regularização *reg_alpha* e *reg_lambda* entre 0 e 1. O compilado dos intervalos de parâmetros considerado pode ser visto na tabela III.

TABLE III: Espaço de busca da busca aleatória de hiperparâmetros do tuning modelo LightGBM

Parâmetro	Intervalo/Valor
num_leaves	20 a 150
max_depth	3 a 12
learning_rate	0.01 a 0.31
n_estimators	50 a 500
min_child_samples	5 a 100
subsample	0.6 a 1.0
colsample_bytree	0.6 a 1.0
reg_alpha	0 a 1
reg_lambda	0 a 1
Random Search - n_iter	50

Além disso, o modelo também foi configurado com o parâmetro *class_weight="balanced"*, para o balanceamento de cargas, assim como descrito anteriormente. A busca por hiperparâmetros foi realizada com *n_iter=50*, o que significa que 50 combinações aleatórias de parâmetros são testadas, e também utiliza validação cruzada com *cv=5*. Dessa maneira por execução eram realizadas 5 iterações da busca aleatória para cada iteração do das 5 *folds* do *LightGBM Classifier*, tanto para a versão binária do problema e para a versão multiclasse. Os resultados do modelo binário podem ser visto na tabela V e os do modelo multiclasse na tabela VIII.

3) SVC: No terceiro pipeline, temos uma etapa inicial utiliza o método *Multivariate Imputation by Chained Equations* (MICE), implementado pelo *IterativeImputer* do *sklearn*. Esse método preenche valores ausentes em cada *feature* por meio de uma regressão multivariada, usando como preditoras as demais variáveis do conjunto de dados. No caso deste pipeline, o parâmetro *n_nearest_features=5* limita as preditoras a apenas as cinco *features* mais correlacionadas, focando nas variáveis mais relevantes e reduzindo o custo computacional. Utilizamos o parâmetro *keep_empty_features=True* que mantém as *features* completamente ausentes no *dataset*, facilitando a análise mesmo em cenários com dados severamente incompletos. O imputador utiliza como padrão o estimador *BayesianRidge* para a regressão.

Após a imputação, os dados passam por *Z-Score Normalization*, realizada pelo *StandardScaler*. Esse método ajusta as *features* para uma distribuição com média 0 e desvio padrão 1, garantindo que todas as variáveis estejam na mesma escala. Essa normalização evita que features com escalas maiores dominem o modelo, que são bastante sensíveis à escala das features, garantindo que todas contribuam igualmente. Além de melhorar a eficiência do treinamento, e facilitar a convergência do algoritmo de otimização. No caso específico do SVM, a normalização é ainda mais importante porque o modelo constrói um hiperplano de separação baseado na maximização da margem entre as classes. Se as features estiverem em escalas muito diferentes, o cálculo da margem pode ser distorcido, resultando em uma separação subótima dos dados. Foram utilizados os valores padrão do *scikit-learn*, que garantem a padronização dos dados com base na média e

no desvio padrão calculados no conjunto de treino.

Por fim, o pipeline utiliza um classificador SVM implementado pelo *SVC* do *scikit-learn*. Outros parâmetros seguem os valores padrão: o *kernel* é linear, a penalização *C=1.0* controla o equilíbrio entre margem ampla e erro de classificação, e o parâmetro *gamma='scale'* ajusta a influência de cada amostra nos dados. Assim como os outros pipelines, o modelo foi treinado para a versão binária do problema e para a versão multiclasse, utilizando do método de *k-fold Cross Validation*, com $k = 5$, e reportamos a média das 5 *folds* e os resultados podem ser vistos na tabela VI e os do modelo multiclasse na tabela IX.

C. Discussão de resultados

A primeira coisa que se percebe é que com o foco no *subset* de pacientes com passagem pela UTI, o número de amostras positivas para todas as bactérias caiu bastante. A coluna número de instâncias representa quantas instâncias positivas cada problema considerou, lembrando que as tabelas IV, V e VI nós trazem os resultados de seis modelos distintos um para cada bactéria, representando a abordagem binária do problema. Enquanto as tabelas VII, VIII e IX nos trazem o resultado de um modelo cada, representando a abordagem multiclasse do problema.

Em relação à abordagem binária a busca aleatória para o *tuning* de hiperparâmetros do *LightGBM Classifier* se destacou das demais em todos os modelos de bactérias, como podemos ver comparando a ROC AUC do conjunto de validação das tabelas IV, V e VI. Para o pipeline a métrica ficou entre 0.636 para a bactéria *Klebsiella pneumoniae* e 0.701 para a bactéria *Escherichia coli*. O pipeline também se destacou na métrica de sensibilidade quando comparado aos outros o que significa que ele é melhor em identificar casos positivos, com valores entre 0.12 para a bactéria *Enterococcus faecalis* e 0.266 para a bactéria *Pseudomonas aeruginosa*. Em geral, ele se destacou também no F1-Score, que é uma métrica que balanceia a precisão e o recall, com valores entre 0.174 para a bactéria *Enterococcus faecalis* e 0.292 para a bactéria *Acinetobacter baumannii*. Vale ressaltar que a bactéria *Acinetobacter baumannii* foi a que possuiu menor número de instâncias para o problema. O modelo apenas perdeu em especificidade para o pipeline que utilizava *SVC*, métrica que indica a qualidade do modelo de identificar casos negativos, porém o modelo de SVC teve uma sensibilidade e F1-score bastante baixo, o que provavelmente indica que teve uma tendência a considerar a maioria dos pacientes como não infectado. Ainda sim, o pipeline com *tuning* de hiperparâmetros apresentou uma elevada especificidade com valores que variam entre 0.903 para a bactéria *Klebsiella pneumoniae* e 0.975 para a bactéria *Enterococcus faecalis*. Assim, dada a complexidade do problema alcançamos resultados satisfatórios para o problema de classificação binário utilizando o *tuning* de hiperparâmetros do *LightGBM Classifier*. Ademais, o modelo do *LightGBM Classifier* sem *tuning* também teve, em geral, métricas melhores que o pipeline que utilizou *SVC*.

Por outro lado, para a abordagem muiticlasse não tivemos um melhor modelo claro, pelas tabelas VII, VIII e IX. Podemos ver que o modelo *SVC* teve melhores ROC AUC para as bactérias *Acinetobacter baumannii*, *Enterococcus faecalis* e *Pseudomonas aeruginosa*, com valores 0.716, 0.645 e 0.639, respectivamente. O padrão classico de parâmetros do *LightGBM Classifier* se destacou nas demais bacterias *Escherichia coli*, *Klebsiella pneumoniae* e *Staphylococcus aureus*, com valores 0.608, 0.626 e 0.667, respectivamente, empatando com o modelo de *tuning* de hiperparâmetros do *LightGBM Classifier* para a bactéria *Escherichia coli*. Apesar disso, em geral, o *tuning* de hiperparâmetros do *LightGBM Classifier* ganha dos demais modelos outras métricas como sensibilidade e F1-score, o que indica que ele é melhor que os demais para identificar casos positivos de bacteria e trás um bom equilibrio entre sensibilidade e especificidade. Novamente vemos a tendência do *SVC* de inflar sua métrica de ROC AUC baseado em uma elevada especificidade, entre em geral se aproximando do 0.98, mas mantendo uma baixissima sensibilidade que se acerca do 0.05 e um baixo F1-score, como representado na tabela IX. Enquanto o *tuning* de hiperparâmetros do *LightGBM Classifier* nos tras sensibilidades mais elevadas que giram em torno dos 0.2 e F1-scores também mais elevados em torno dos 0.27, como mostrado na tabela VIII. O modelo ainda trás ROC AUC competitivas que variam de 0.608, para a bactéria *Escherichia coli*, até 0.689 para a bactéria *Acinetobacter baumannii*. Dessa maneira, argumentamos que a melhor escolha segue sendo o *tuning* de hiperparâmetros do *LightGBM Classifier* e também trás um resultado satisfatório dada a complexidade do problema.

IV. Conclusões

Com o trabalho desenvolvido foi possível comparar 3 pipelines distintas para o problema de predição de infecções bacterianas no contexto de paciente da UTI do Hospital das Clínicas e foi possível aplicar conhecimento da área de Inteligência Artificial (IA) na área da saúde. Concluímos que o pipeline de que utiliza do busca aleatória para realizar o *tuning* de hiperparâmetros do *LightGBM Classifier* se aplicou melhor ao problema do que os demais pipelines: *LightGBM Classifier* sem *tuning* de hiperparâmetros e *Multivariate Imputation by Chained Equations* (MICE) e *ZScore Normalization* junto ao modelo *Support Vector Machine* (SVM). Além disso, o *tuning* de hiperparâmetros do *LightGBM Classifier* nos trouxe resultados satisfatórios dada a complexidade do problema. Por fim, os modelos podem estar limitados devido a qualidade dos dados e *embeddings* utilizados, futuras experimentações podem se beneficiar de novos tratamentos prévios pré treinamento de modelos.

Agradecimentos

Gostariamos de agradecer o apoio e suporte contínuo dos médicos e professores da Faculdade de Medicina da UFMG, Saulo Fernandes Saturnino e Helena Duani, cujas contribuições foram fundamentais para a realização desse trabalho.

Além disso, gostariamos de agradecer nosso orientador, Adriano Veloso, e demais membros do grupo de pesquisa.

REFERENCES

[1] L. DONG, G. GUOZHU, and L. HUAN Liu, eds., *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, 2018.

[2] G. KE *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] H. SHANG *et al.*, "Artificial Intelligence and Machine Learning Assisted Drug Delivery for Effective Treatment of Infectious Diseases," *Advanced Drug Delivery Reviews*, vol. 178, art. 113922, 2021. Available: https://doi.org/10.1016/j.addr.2021.113922. Accessed: Mar. 31, 2024.

[4] X. WAN *et al.*, "Risk Factors Analysis of COVID-19 Patients with ARDS and Prediction Based on Machine Learning," *Scientific Reports*, vol. 11, art. 2933, 2011. Available: https://www.nature.com/articles/s41598-021-82492-x. Accessed: Mar. 31, 2024.

[5] S. F. Buck, "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 22, no. 2, pp. 302–306, 1960. Available: https://doi.org/10.1111/j.2517-6161.1960.tb00375.x. Accessed: Jan. 13, 2025.

[6] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. Available: https://doi.org/10.18637/jss.v045.i03. Accessed: Jan. 13, 2025.

[7] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. Available: https://doi.org/10.1007/BF00994018. Accessed: Jan. 13, 2025.

[8] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, no. 2, pp. 281–305, 2012. Available: http://jmlr.org/papers/v13/bergstra12a.html. Accessed: Jan. 13, 2025.

[9] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, 4th ed. Morgan Kaufmann, 2022.

[10] M. H. KOLEFF *et al.*, "Timing of Antibiotic Therapy in the ICU," *Critical Care*, vol. 25, art. 360, 2021. Available: https://ccforum.biomedcentral.com/articles/10.1186/s13054-021-03787-z. Accessed: Mar. 31, 2024.

[11] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, "The Area Under the Precision-Recall Curve as a Performance Metric for Rare Binary Events," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565-577, 2019.

[12] T. Fushiki, "Estimation of Prediction Error by Using K-Fold Cross-Validation," *Statistical Computation and Simulation*, vol. 21, pp. 137–146, 2011. Available: https://doi.org/10.1007/s11222-009-9153-8.

[13] WORLD HEALTH ORGANIZATION (WHO) and FIFTY-EIGHTH WORLD HEALTH ASSEMBLY, *Antimicrobial Resistance: A Threat to Global Health Security - Rational Use of Medicines by Prescribers and Patients*, Geneva: WHO, 2005. Available: https://apps.who.int/gb/archive/pdf_files/WHA58/A58_14-en.pdf. Accessed: Mar. 31, 2024.

[14] WORLD HEALTH ORGANIZATION (WHO), *Global Action Plan on Antimicrobial Resistance*, Geneva: WHO, 2015. Available: https://www.amcra.be/swfiles/files/WHO%20actieplan_90.pdf. Accessed: Mar. 31, 2024.

[15] WORLD HEALTH ORGANIZATION (WHO), *Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022*, Geneva: WHO, 2022a. Available: https://www.who.int/publications/i/item/9789240062702. Accessed: Mar. 31, 2024.

[16] WORLD HEALTH ORGANIZATION (WHO), *The WHO AWaRe (Access, Watch, Reserve) Antibiotic Book*, Geneva: WHO, 2022b. Available: https://www.who.int/publications/i/item/9789240062382. Accessed: Mar. 31, 2024.

[17] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

TABLE IV: Métricas LightGBM Classifier - modelos binários

Bacteria	Metrics						
	Número de instâncias	Training ROCAUC	Validation ROCAUC	Validation Sensitivity/Recall	Validation Specificity	Validation F1-Score	Accuracy
ACINETOBACTER BAUMANNII	115	0.999	0.690	0.270	0.950	0.306	0.879
ENTEROCOCCUS FAECALIS	100	1.000	0.654	0.100	0.978	0.148	0.899
ESCHERICHIA COLI	203	1.000	0.702	0.202	0.935	0.268	0.800
KLEBSIELLA PNEUMONIAE	226	1.000	0.613	0.208	0.906	0.264	0.763
PSEUDOMONAS AERUGINOSA	169	1.000	0.658	0.242	0.931	0.296	0.826
STAPHYLOCOCCUS AUREUS	126	0.999	0.657	0.231	0.953	0.279	0.871

TABLE V: Métricas LightGBM Classifier com tuning de hiperparâmetros - modelos binário

Bacteria	Metrics						
	Número de instâncias	Training ROCAUC	Validation ROCAUC	Validation Sensitivity/Recall	Validation Specificity	Validation F1-Score	Accuracy
ACINETOBACTER BAUMANNII	115	0.999	0.693	0.261	0.945	0.292	0.874
ENTEROCOCCUS FAECALIS	100	1.000	0.682	0.120	0.975	0.174	0.898
ESCHERICHIA COLI	203	1.000	0.701	0.212	0.929	0.276	0.798
KLEBSIELLA PNEUMONIAE	226	1.000	0.636	0.217	0.903	0.273	0.763
PSEUDOMONAS AERUGINOSA	169	1.000	0.670	0.266	0.923	0.313	0.823
STAPHYLOCOCCUS AUREUS	126	0.999	0.667	0.247	0.943	0.283	0.864

TABLE VI: Métricas SVC - modelos binário

Bacteria	Metrics						
	Número de instâncias	Training ROCAUC	Validation ROCAUC	Validation Sensitivity/Recall	Validation Specificity	Validation F1-Score	Accuracy
ACINETOBACTER BAUMANNII	115	0.981	0.640	0.017	0.997	0.032	0.895
ENTEROCOCCUS FAECALIS	100	0.992	0.629	0.010	0.999	0.018	0.910
ESCHERICHIA COLI	203	0.991	0.661	0.000	0.994	0.000	0.812
KLEBSIELLA PNEUMONIAE	226	0.975	0.623	0.040	0.992	0.072	0.798
PSEUDOMONAS AERUGINOSA	169	0.979	0.627	0.053	0.988	0.093	0.846
STAPHYLOCOCCUS AUREUS	126	0.983	0.598	0.000	0.996	0.000	0.883

TABLE VII: Métricas LightGBM Classifier - modelo multi-classes

Bacteria	Metrics						
	Número de instâncias	Training ROCAUC	Validation ROCAUC	Validation Sensitivity/Recall	Validation Specificity	Validation F1-Score	Accuracy
ACINETOBACTER BAUMANNII	3280	0.999	0.703	0.231	0.948	0.271	0.874
ENTEROCOCCUS FAECALIS	3280	1.000	0.582	0.020	0.988	0.033	0.901
ESCHERICHIA COLI	3280	0.999	0.608	0.849	0.350	0.358	0.442
KLEBSIELLA PNEUMONIAE	3280	1.000	0.626	0.203	0.940	0.279	0.790
PSEUDOMONAS AERUGINOSA	3280	1.000	0.597	0.191	0.938	0.243	0.824
STAPHYLOCOCCUS AUREUS	3280	0.999	0.667	0.200	0.960	0.263	0.874

TABLE VIII: Métricas LightGBM Classifier com tuning de hiperparâmetros - modelo multiclasses

Bacteria	Metrics						
	Número de instâncias	Training ROCAUC	Validation ROCAUC	Validation Sensitivity/Recall	Validation Specificity	Validation F1-Score	Accuracy
ACINETOBACTER BAUMANNII	3280	0.999	0.689	0.276	0.943	0.311	0.874
ENTEROCOCCUS FAECALIS	3280	1.000	0.615	0.035	0.986	0.056	0.900
ESCHERICHIA COLI	3280	0.998	0.608	0.792	0.391	0.351	0.464
KLEBSIELLA PNEUMONIAE	3280	0.999	0.623	0.216	0.924	0.279	0.780
PSEUDOMONAS AERUGINOSA	3280	1.000	0.613	0.208	0.930	0.255	0.819
STAPHYLOCOCCUS AUREUS	3280	0.999	0.652	0.200	0.952	0.252	0.866

TABLE IX: Métricas SVC - modelo multiclasses

Bacteria	Metrics						
	Número de instâncias	Training ROCAUC	Validation ROCAUC	Validation Sensitivity/Recall	Validation Specificity	Validation F1-Score	Accuracy
ACINETOBACTER BAUMANNII	3280	0.978	0.716	0.061	0.984	0.098	0.888
ENTEROCOCCUS FAECALIS	3280	0.995	0.645	0.000	1.000	0.000	0.910
ESCHERICHIA COLI	3280	0.922	0.583	0.983	0.097	0.327	0.260
KLEBSIELLA PNEUMONIAE	3280	0.956	0.619	0.076	0.985	0.130	0.799
PSEUDOMONAS AERUGINOSA	3280	0.971	0.639	0.082	0.982	0.136	0.845
STAPHYLOCOCCUS AUREUS	3280	0.983	0.618	0.008	0.994	0.014	0.882

TABLE X: Estatísticas Tabela de Features Pós Limpeza dos Dados

	#Não nulos	#Nulos	(%)	#Unicos	Média	Min	1º Quartil	Mediana	3º Quartil	Max
ALTSA-1	74607	477462	13.51	3506	58.93	4.0	20.0	29.4	45.3	18249.0
ALTURA	519	551550	0.09	116	42.85	1.0	1.58	1.66	36.825	1515.0
BILISA-1	54696	497373	9.91	2516	2.13	0.0	0.44	0.69	1.41	71.29
BILISA-2	54054	498015	9.79	2315	1.5	0.0	0.21	0.36	0.63	63.86
BILISA-3	54124	497945	9.80	734	0.65	-0.3	0.15	0.34	0.76	16.74
CLORSA-1	28310	523759	5.13	530	104.23	56.4	101.3	104.3	107.1	161.9
CREASA-1	136713	415356	24.76	1366	1.23	0.05	0.6	0.83	1.29	19.81
FERRIT-1	15497	536572	2.81	2395	291.28	2.44	33.8	91.2	257.0	80000.0
FIBR-1	3990	548079	0.72	815	290.03	16.0	162.0	255.0	375.0	1497.0
FIO2	12611	539458	2.28	42	40.99	21.0	30.0	40.0	40.0	4023.0
FREQUENCIA CARDIACA	321882	230187	58.30	217	87.11	20.0	73.0	84.0	98.0	250.0
FREQUENCIA RESPIRATORIA	297854	254215	53.95	76	19.86	6.0	18.0	19.0	20.0	60.0
FREQUÊNCIA DE PULSO (M2BR)	31	552038	0.01	19	77.29	36.0	68.5	76.0	88.5	120.0
FREQUÊNCIA RESPIRATÓRIA (M2BR)	54	552015	0.01	9	19.67	15.0	17.25	18.0	20.0	86.0
GASOA-1	32151	519918	5.82	893	7.4	5.0	7.35	7.416	7.467	7.782
GASOA-2	32081	519988	5.81	2353	91.31	11.0	66.0	85.1	108.8	523.9
GASOA-3	32150	519919	5.82	951	38.79	8.2	31.5	36.3	42.6	191.0
GASOA-4	32149	519920	5.82	503	23.1	1.7	19.6	22.8	26.1	83.4
GASOA-6	32139	519930	5.82	557	-1.57	-34.4	-4.9	-1.3	2.0	48.8
GASOA-7	31701	520368	5.74	599	92.19	40.0	92.4	95.9	97.6	100.0
GGT-1	48637	503432	8.81	6490	156.08	10.0	27.0	52.4	141.0	10318.0
GLIC-1	54023	498046	9.79	2592	103.94	20.0	84.0	92.3	106.0	600.0
GLICEMIA CAPILAR	70318	481751	12.74	622	151.09	20.0	106.0	134.0	177.0	600.0
HMG-2	148463	403606	26.89	4815	9.07	0.0	5.02	7.11	10.03	846.93
IMC	2	552067	0.00	2	45.15	19.3	32.225	45.15	58.075	71.0
INDICE CARDÍACO	145	551924	0.03	49	4.43	1.0	3.6	4.4	5.1	8.2
LAC-1	4509	547560	0.82	565	2.05	0.5	1.16	1.64	2.4	20.22
MAGNSA-1	79480	472589	14.40	336	1.96	1.0	1.75	1.95	2.16	5.93
PADRÃO ATIVIDADE (M2BR)	3	552066	0.00	3	102.67	78.0	84.0	90.0	115.0	140.0
PAM	53930	498139	9.77	161	87.19	20.0	74.0	85.0	98.0	193.0
PAP	291	551778	0.05	47	27.99	7.0	21.0	26.0	33.5	62.0
PCP	112	551957	0.02	41	21.06	1.0	12.0	15.5	24.25	131.0
PCR-1	65355	486714	11.84	15433	68.74	5.0	7.59	27.62	75.005	706.4
PERFUSÃO CAPILAR (M2BR)	4	552065	0.00	3	45.0	2.0	2.0	41.0	84.0	96.0
PFE (M2BR)	1	552068	0.00	1	18.0	18.0	18.0	18.0	18.0	18.0
POTASA-1	88688	463381	16.06	605	4.45	1.16	4.07	4.42	4.8	13.7
PPI	10914	541155	1.98	55	20.44	1.0	16.0	20.0	23.0	2006.0
PRESSÃO ARTERIAL DISTOLICA	287157	264912	52.01	173	71.67	5.0	60.0	70.0	80.0	160.0
PRESSÃO ARTERIAL SISTOLICA	287481	264588	52.07	229	118.31	40.0	103.0	118.0	130.0	270.0
PRESSÃO INTRA-ABDOMINAL	10	552059	0.00	7	34.5	1.0	2.0	9.0	79.5	97.0
PRESSÃO INTRA-ARTERIAL	122	551947	0.02	47	83.06	22.0	74.0	83.0	90.0	130.0
PROBNP-1	1185	550884	0.21	874	4380.6	11.1	232.0	1030.0	4310.0	87300.0
PTTA-1	40666	511403	7.37	7	28.71	26.5	27.0	28.8	30.9	33.8
PTTA-2	40654	511415	7.36	934	33.73	0.2	26.8	30.2	34.9	1200.0
PVC	295	551774	0.05	29	12.85	0.0	9.0	13.0	16.0	60.0
SATURAÇÃO DE O2 (M2BR)	23	552046	0.00	9	95.17	90.0	93.5	96.0	97.0	98.0
SATURAÇÃO DE OXIGÊNIO	129048	423021	23.38	147	96.72	40.0	95.0	97.0	98.0	9997.0
SODISA-1	84469	467600	15.30	552	139.17	104.6	137.0	139.5	142.0	202.2
TAP-4	50406	501663	9.13	698	1.4	0.72	1.01	1.12	1.37	120.0
TEMPERATURA(C)	331919	220150	60.12	125	36.23	33.0	35.8	36.2	36.6	41.0
TROPUL-1	961	551108	0.17	654	4776.88	1.5	6.4	30.2	264.9	1205000.0
UREISA-1	129531	422538	23.46	2356	50.38	4.0	26.3	37.2	60.0	299.0
data_evento	552069	0	100.00	126999	-	-	-	-	-	-
prontuario	552069	0	100.00	24959	-	-	-	-	-	-

TABLE XI: Estatisticas após Engenharia de Features - Tabela A

	#Não nulos	#Nulos	%	Média	Min	Quartil 25%	Mediana	Quartil 75%	Max	#Únicos
ALTSA-1	74607	200844	27.09	58.93	4.00	20.00	29.40	45.30	18249.00	3506
ALTURA_first_value	515	274936	0.19	42.40	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_last_value	515	274936	0.19	42.16	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_max	515	274936	0.19	42.40	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_mean	515	274936	0.19	42.28	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_median	515	274936	0.19	42.28	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_min	515	274936	0.19	42.16	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_most_discrepant_value	515	274936	0.19	42.40	1.00	1.58	1.65	5.50	1515.00	115
ALTURA_std	3	275448	0.00	29.23	0.00	0.00	0.00	43.84	87.68	2
BILISA-1	54696	220755	19.86	2.13	0.00	0.44	0.69	1.41	71.29	2516
BILISA-2	54054	221397	19.62	1.50	0.00	0.21	0.36	0.63	63.86	2315
BILISA-3	54124	221327	19.65	0.65	-0.30	0.15	0.34	0.76	16.74	734
CLORSA-1	28310	247141	10.28	104.23	56.40	101.30	104.30	107.10	161.90	530
CREASA-1	136713	138738	49.63	1.23	0.05	0.60	0.83	1.29	19.81	1366
FERRIT-1	15497	259954	5.63	291.28	2.44	33.80	91.20	257.00	80000.00	2395
FIBR-1	3990	271461	1.45	290.03	16.00	162.00	255.00	375.00	1497.00	815
FIO2_first_value	1706	273745	0.62	44.74	21.00	40.00	40.00	40.00	100.00	28
FIO2_last_value	1706	273745	0.62	41.33	21.00	30.00	40.00	40.00	100.00	30
FIO2_max	1706	273745	0.62	52.26	21.00	40.00	40.00	50.00	4023.00	32
FIO2_mean	1706	273745	0.62	42.83	21.00	35.00	40.00	42.47	371.92	266
FIO2_median	1706	273745	0.62	41.65	21.00	35.00	40.00	40.00	100.00	39
FIO2_min	1706	273745	0.62	39.29	21.00	30.00	40.00	40.00	100.00	32
FIO2_most_discrepant_value	1706	273745	0.62	50.33	21.00	40.00	40.00	50.00	4023.00	35
FIO2_std	1482	273969	0.54	5.81	0.00	0.00	0.00	5.27	1149.79	318
FC_first_value	114254	161197	41.48	87.68	20.00	73.00	84.00	99.00	250.00	199
FC_last_value	114254	161197	41.48	88.13	20.00	74.00	85.00	100.00	235.00	195
FC_max	114254	161197	41.48	94.06	20.00	80.00	91.00	105.00	250.00	186
FC_mean	114254	161197	41.48	87.99	20.00	75.00	84.50	98.00	210.00	2730
FC_median	114254	161197	41.48	87.93	20.00	74.50	84.00	98.00	210.00	328
FC_min	114254	161197	41.48	82.04	20.00	68.00	79.00	92.00	210.00	198
FC_most_discrepant_value	114254	161197	41.48	88.20	20.00	73.00	85.00	100.00	250.00	211
FC_std	93714	181737	34.02	7.90	0.00	3.51	6.24	10.54	100.46	7103
FR_first_value	106897	168554	38.81	20.02	6.00	18.00	19.00	20.00	60.00	70
FR_last_value	106897	168554	38.81	20.10	6.00	18.00	20.00	20.00	60.00	71
FR_max	106897	168554	38.81	21.33	7.00	19.00	20.00	22.00	60.00	72
FR_mean	106897	168554	38.81	20.04	6.67	18.00	19.00	20.00	60.00	1026
FR_median	106897	168554	38.81	20.01	7.00	18.00	19.00	20.00	60.00	119
FR_min	106897	168554	38.81	18.83	6.00	17.00	18.00	20.00	60.00	61
FR_most_discrepant_value	106897	168554	38.81	20.21	6.00	18.00	20.00	20.00	60.00	74
FR_std	84946	190505	30.84	1.63	0.00	0.71	1.15	2.08	31.11	3542
FdP (M2BR)_first_value	28	275423	0.01	75.75	36.00	68.00	76.00	81.25	120.00	18
FdP (M2BR)_last_value	28	275423	0.01	76.75	36.00	68.00	76.00	87.75	120.00	18
FdP (M2BR)_max	28	275423	0.01	76.82	36.00	68.00	76.00	87.75	120.00	18
FdP (M2BR)_mean	28	275423	0.01	76.25	36.00	68.00	76.00	84.25	120.00	20
FdP (M2BR)_median	28	275423	0.01	76.25	36.00	68.00	76.00	84.25	120.00	20
FdP (M2BR)_min	28	275423	0.01	75.68	36.00	68.00	76.00	81.25	120.00	18
FdP (M2BR)_most_discrepant_value	28	275423	0.01	75.75	36.00	68.00	76.00	81.25	120.00	18
FdP (M2BR)_std	3	275448	0.00	7.54	1.41	5.30	9.19	10.61	12.02	3
FR (M2BR)_first_value	49	275402	0.02	19.71	15.00	18.00	18.00	20.00	86.00	7
FR (M2BR)_last_value	49	275402	0.02	19.92	15.00	18.00	18.00	20.00	86.00	9
FR (M2BR)_max	49	275402	0.02	19.92	15.00	18.00	18.00	20.00	86.00	9
FR (M2BR)_mean	49	275402	0.02	19.82	15.00	18.00	18.00	20.00	86.00	9
FR (M2BR)_median	49	275402	0.02	19.82	15.00	18.00	18.00	20.00	86.00	9
FR (M2BR)_min	49	275402	0.02	19.71	15.00	18.00	18.00	20.00	86.00	7
FR (M2BR)_most_discrepant_value	49	275402	0.02	19.71	15.00	18.00	18.00	20.00	86.00	7
FR (M2BR)_std	4	275447	0.00	1.77	0.00	0.00	1.41	3.18	4.24	3
GASOA-1	32151	243300	11.67	7.40	5.00	7.35	7.42	7.47	7.78	893
GASOA-2	32081	243370	11.65	91.31	11.00	66.00	85.10	108.80	523.90	2353
GASOA-3	32150	243301	11.67	38.79	8.20	31.50	36.30	42.60	191.00	951
GASOA-4	32149	243302	11.67	23.10	1.70	19.60	22.80	26.10	83.40	503
GASOA-6	32139	243312	11.67	-1.57	-34.40	-4.90	-1.30	2.00	48.80	557
GASOA-7	31701	243750	11.51	92.19	40.00	92.40	95.90	97.60	100.00	599
GGT-1	48637	226814	17.66	156.08	10.00	27.00	52.40	141.00	10318.00	6490
GLIC-1	54023	221428	19.61	103.94	20.00	84.00	92.30	106.00	600.00	2592

Features abreviadas para possibilitar exibição: (FC) FREQUENCIA CARDIACA, (FR) FREQUENCIA RESPIRATORIA, (GC) GLICEMIA CAPILAR, (PAD) PRESSÃO ARTERIAL DISTOLICA, (PAS) PRESSÃO ARTERIAL SISTOLICA, (PI-ART) PRESSÃO INTRA-ARTERIAL, (SdO) SATURAÇÃO DE OXIGÊNIO, (Sd) SATURAÇÃO DE, (T) TEMPERATURA, (FdP) FREQUÊNCIA DE PULSO

TABLE XII: Estatisticas após Engenharia de Features - Tabela B

	#Não nulos	#Nulos	%	Média	Min	Quartil 25%	Mediana	Quartil 75%	Max	#Unicos
GC_first_value	30701	244750	11.15	139.39	20.00	100.00	123.00	162.00	600.00	548
GC_last_value	30701	244750	11.15	149.61	20.00	106.00	132.00	174.00	600.00	548
GC_max	30701	244750	11.15	165.55	20.00	114.00	145.00	198.00	600.00	578
GC_mean	30701	244750	11.15	145.45	20.00	107.33	131.00	169.50	599.00	2558
GC_median	30701	244750	11.15	144.70	20.00	106.00	130.00	168.00	599.00	839
GC_min	30701	244750	11.15	126.55	20.00	95.00	114.00	145.00	599.00	514
GC_most_discrepant_value	30701	244750	11.15	146.49	20.00	101.00	127.00	171.00	600.00	577
GC_std	19555	255896	7.10	33.01	0.00	12.02	24.04	45.25	287.09	7320
HMG-2	148463	126988	53.90	9.07	0.00	5.02	7.11	10.03	846.93	4815
IMC	2	275449	0.00	45.15	19.30	32.22	45.15	58.08	71.00	2
IC_first_value	36	275415	0.01	4.08	1.00	2.98	3.95	4.95	7.50	25
IC_last_value	36	275415	0.01	4.19	1.00	3.10	3.95	5.28	8.20	24
IC_max	36	275415	0.01	4.68	1.00	3.68	4.50	5.70	8.20	29
IC_mean	36	275415	0.01	4.20	1.00	3.43	4.11	5.15	6.85	33
IC_median	36	275415	0.01	4.17	1.00	3.30	4.20	5.06	6.90	30
IC_min	36	275415	0.01	3.72	1.00	2.90	3.55	4.35	6.40	25
IC_most_discrepant_value	36	275415	0.01	4.21	1.00	2.98	4.05	5.28	8.20	24
IC_std	28	275423	0.01	0.50	0.00	0.28	0.39	0.60	1.58	28
LAC-1	4509	270942	1.64	2.05	0.50	1.16	1.64	2.40	20.22	565
MAGNSA-1	79480	195971	28.85	1.96	1.00	1.75	1.95	2.16	5.93	336
PADRÃO ATIVIDADE (M2BR)	3	275448	0.00	102.67	78.00	84.00	90.00	115.00	140.00	3
PAM_first_value	6201	269250	2.25	87.17	32.00	74.00	85.00	98.00	188.00	128
PAM_last_value	6201	269250	2.25	88.02	20.00	75.00	86.00	99.00	193.00	138
PAM_max	6201	269250	2.25	100.87	32.00	88.00	99.00	112.00	193.00	139
PAM_mean	6201	269250	2.25	87.60	30.17	76.71	85.92	97.00	180.80	1981
PAM_median	6201	269250	2.25	87.41	29.00	76.00	85.50	97.50	180.80	197
PAM_min	6201	269250	2.25	75.14	20.00	64.00	73.00	84.00	180.80	124
PAM_most_discrepant_value	6201	269250	2.25	89.27	20.00	72.00	88.00	103.00	193.00	156
PAM_std	5807	269644	2.11	9.26	0.00	6.22	8.51	11.40	54.45	4785
PAP_first_value	69	275382	0.03	28.45	8.00	22.00	28.00	33.00	60.00	31
PAP_last_value	69	275382	0.03	27.35	9.00	21.00	25.00	34.00	56.00	29
PAP_max	69	275382	0.03	32.45	13.00	24.00	32.00	38.00	62.00	33
PAP_mean	69	275382	0.03	27.66	11.20	21.57	25.44	32.67	50.25	59
PAP_median	69	275382	0.03	27.15	8.00	22.00	25.00	32.00	53.50	38
PAP_min	69	275382	0.03	23.42	7.00	18.00	22.00	29.00	43.00	27
PAP_most_discrepant_value	69	275382	0.03	28.65	9.00	22.00	26.00	37.00	60.00	34
PAP_std	59	275392	0.02	4.51	0.00	1.93	3.83	6.23	17.74	49
PCP_first_value	35	275416	0.01	27.94	1.00	11.00	19.00	28.50	131.00	26
PCP_last_value	35	275416	0.01	27.74	1.00	11.00	21.00	28.00	131.00	29
PCP_max	35	275416	0.01	30.08	1.00	14.50	22.00	30.50	131.00	26
PCP_mean	35	275416	0.01	27.71	1.00	11.88	19.12	28.00	131.00	33
PCP_median	35	275416	0.01	27.67	1.00	12.50	19.00	27.50	131.00	30
PCP_min	35	275416	0.01	25.60	1.00	9.00	16.00	27.00	131.00	27
PCP_most_discrepant_value	35	275416	0.01	28.68	1.00	12.00	20.00	30.50	131.00	29
PCP_std	21	275430	0.01	3.58	0.58	1.53	3.00	4.04	17.68	18
PCR-1	65355	210096	23.73	68.74	5.00	7.59	27.62	75.00	706.40	15433
PERFUSÃO CAPILAR (M2BR)	4	275447	0.00	45.00	2.00	2.00	41.00	84.00	96.00	3
PFE (M2BR)	1	275450	0.00	18.00	18.00	18.00	18.00	18.00	18.00	1
POTASA-1	88688	186763	32.20	4.45	1.16	4.07	4.42	4.80	13.70	605
PPI_first_value	1549	273902	0.56	20.09	1.00	16.00	19.00	23.00	400.00	42
PPI_last_value	1549	273902	0.56	19.40	1.00	15.00	19.00	22.00	219.00	45
PPI_max	1549	273902	0.56	23.94	1.00	18.00	21.00	25.00	2006.00	51
PPI_mean	1549	273902	0.56	19.83	1.00	16.08	19.00	22.29	217.80	535
PPI_median	1549	273902	0.56	19.61	1.00	16.00	19.00	22.00	108.00	64
PPI_min	1549	273902	0.56	17.28	1.00	14.00	17.00	20.00	108.00	41
PPI_most_discrepant_value	1549	273902	0.56	21.66	1.00	15.00	19.00	23.00	2006.00	54
PPI_std	1398	274053	0.51	2.77	0.00	0.73	1.63	2.80	628.31	753
PAD_first_value	93997	181454	34.12	73.03	5.00	62.00	70.00	80.00	160.00	151
PAD_last_value	93997	181454	34.12	72.27	5.00	60.00	70.00	80.00	160.00	154
PAD_max	93997	181454	34.12	79.20	5.50	70.00	80.00	88.00	160.00	141
PAD_mean	93997	181454	34.12	72.43	5.50	65.00	71.67	80.00	160.00	2077

Features abreviadas para possibilitar exibição: (FC) FREQUENCIA CARDIACA, (FR) FREQUENCIA RESPIRATORIA, (GC) GLICEMIA CAPILAR, (PAD) PRESSÃO ARTERIAL DISTOLICA, (PAS) PRESSÃO ARTERIAL SISTOLICA, (PI-ART) PRESSÃO INTRA-ARTERIAL, (SdO) SATURAÇÃO DE OXIGÊNIO, (Sd) SATURAÇÃO DE, (T) TEMPERATURA, (FdP) FREQUÊNCIA DE PULSO

TABLE XIII: Estatisticas após Engenharia de Features - Tabela C

	#Não nulos	#Nulos	%	Média	Min	Quartil 25%	Mediana	Quartil 75%	Max	#Unicos
PAD_median	93997	181454	34.12	72.29	5.50	64.00	70.00	80.00	160.00	223
PAD_min	93997	181454	34.12	65.91	5.00	60.00	65.00	72.00	160.00	146
PAD_most_discrepant_value	93997	181454	34.12	73.17	5.00	60.00	70.00	80.00	160.00	170
PAD_std	81944	193507	29.75	7.96	0.00	4.83	7.07	10.60	71.42	6808
PAS_first_value	94090	181361	34.16	118.46	40.00	104.00	120.00	130.00	260.00	208
PAS_last_value	94090	181361	34.16	118.31	40.00	104.00	119.00	130.00	270.00	209
PAS_max	94090	181361	34.16	126.99	41.00	111.00	125.00	140.00	270.00	214
PAS_mean	94090	181361	34.16	117.98	41.00	106.00	116.67	129.00	233.00	2761
PAS_median	94090	181361	34.16	117.77	41.00	105.00	117.50	130.00	233.00	311
PAS_min	94090	181361	34.16	109.24	40.00	100.00	110.00	120.00	233.00	188
PAS_most_discrepant_value	94090	181361	34.16	118.83	40.00	102.00	120.00	130.00	270.00	224
PAS_std	82004	193447	29.77	10.55	0.00	5.77	9.50	14.15	91.92	8136
PRESSÃO INTRA-ABDOMINAL	10	275441	0.00	34.50	1.00	2.00	9.00	79.50	97.00	7
PI-ART_first_value	51	275400	0.02	83.42	22.00	74.50	83.00	93.00	130.00	34
PI-ART_last_value	51	275400	0.02	82.89	22.00	73.00	82.00	92.80	130.00	35
PI-ART_max	51	275400	0.02	88.31	22.00	79.50	86.00	99.00	130.00	31
PI-ART_mean	51	275400	0.02	82.75	22.00	75.25	81.33	91.80	130.00	40
PI-ART_median	51	275400	0.02	82.32	22.00	73.75	82.50	92.80	130.00	36
PI-ART_min	51	275400	0.02	77.76	22.00	70.50	76.00	84.00	130.00	32
PI-ART_most_discrepant_value	51	275400	0.02	83.05	22.00	73.50	82.00	94.00	130.00	35
PI-ART_std	35	275416	0.01	8.47	0.00	3.21	7.53	12.04	28.69	32
PROBNP-1	1185	274266	0.43	4380.60	11.10	232.00	1030.00	4310.00	87300.00	874
PTTA-1	40666	234785	14.76	28.71	26.50	27.00	28.80	30.90	33.80	7
PTTA-2	40654	234797	14.76	33.73	0.20	26.80	30.20	34.90	1200.00	934
PVC_first_value	75	275376	0.03	13.41	0.00	9.50	13.00	16.50	60.00	24
PVC_last_value	75	275376	0.03	13.21	0.00	10.00	13.00	16.50	60.00	22
PVC_max	75	275376	0.03	15.87	0.00	12.00	15.00	18.50	60.00	25
PVC_mean	75	275376	0.03	13.20	0.00	9.58	13.00	16.00	60.00	58
PVC_median	75	275376	0.03	13.09	0.00	10.00	12.50	16.00	60.00	31
PVC_min	75	275376	0.03	10.65	0.00	7.00	10.00	13.00	60.00	23
PVC_most_discrepant_value	75	275376	0.03	13.71	0.00	9.00	13.00	17.50	60.00	27
PVC_std	63	275388	0.02	2.70	0.00	1.34	2.02	3.67	8.50	54
SdO2(M2BR)_first_value	20	275431	0.01	95.30	90.00	93.75	96.00	97.00	98.00	8
SdO2(M2BR)_last_value	20	275431	0.01	95.25	90.00	93.75	96.00	97.00	98.00	9
SdO2(M2BR)_max	20	275431	0.01	95.35	90.00	93.75	96.00	97.00	98.00	8
SdO2(M2BR)_mean	20	275431	0.01	95.29	90.00	93.75	96.00	97.00	98.00	10
SdO2(M2BR)_median	20	275431	0.01	95.28	90.00	93.75	96.00	97.00	98.00	9
SdO2(M2BR)_min	20	275431	0.01	95.25	90.00	93.75	96.00	97.00	98.00	9
SdO2(M2BR)_most_discrepant_value	20	275431	0.01	95.35	90.00	93.75	96.00	97.00	98.00	8
SdO2(M2BR)_std	2	275449	0.00	0.64	0.58	0.61	0.64	0.67	0.71	2
SdO_first_value	55859	219592	20.28	96.83	45.00	95.00	97.00	98.00	9997.00	113
SdO_last_value	55859	219592	20.28	96.75	44.00	95.00	97.00	98.00	9997.00	107
SdO_max	55859	219592	20.28	97.66	46.00	96.00	98.00	98.00	9997.00	136
SdO_mean	55859	219592	20.28	96.70	46.00	95.00	97.00	98.00	9997.00	728
SdO_median	55859	219592	20.28	96.69	46.00	95.00	97.00	98.00	9997.00	149
SdO_min	55859	219592	20.28	95.88	40.00	94.00	96.00	98.00	9997.00	101
SdO_most_discrepant_value	55859	219592	20.28	96.76	40.00	95.00	97.00	98.00	9997.00	139
SdO_std	26345	249106	9.56	1.90	0.00	0.71	1.15	2.08	3105.71	1830
SODISA-1	84469	190982	30.67	139.17	104.60	137.00	139.50	142.00	202.20	552
TAP-4	50406	225045	18.30	1.40	0.72	1.01	1.12	1.37	120.00	698
T(C)_first_value	115150	160301	41.80	36.14	33.00	35.80	36.10	36.50	41.00	95
T(C)_last_value	115150	160301	41.80	36.30	33.00	36.00	36.30	36.70	41.00	98
T(C)_max	115150	160301	41.80	36.55	33.00	36.20	36.50	36.80	41.00	95
T(C)_mean	115150	160301	41.80	36.21	33.00	35.90	36.20	36.50	40.50	2255
T(C)_median	115150	160301	41.80	36.22	33.00	35.90	36.20	36.50	40.50	175
T(C)_min	115150	160301	41.80	35.87	33.00	35.50	35.90	36.20	40.00	95
T(C)_most_discrepant_value	115150	160301	41.80	36.19	33.00	35.70	36.20	36.60	41.00	104
T(C)_std	95479	179972	34.66	0.43	0.00	0.21	0.36	0.57	3.61	9977
TROPUL-1	961	274490	0.35	4776.88	1.50	6.40	30.20	264.90	1205000.00	654
UREISA-1	129531	145920	47.03	50.38	4.00	26.30	37.20	60.00	299.00	2356
dia_evento	275451	0	100.00	-	-	-	-	-	-	2196
prontuario	275451	0	100.00	-	-	-	-	-	-	24959

Features abreviadas para possibilitar exibição: (FC) FREQUENCIA CARDIACA, (FR) FREQUENCIA RESPIRATORIA, (GC) GLICEMIA CAPILAR, (PAD) PRESSÃO ARTERIAL DISTOLICA, (PAS) PRESSÃO ARTERIAL SISTOLICA, (PI-ART) PRESSÃO INTRA-ARTERIAL, (SdO) SATURAÇÃO DE OXIGÊNIO, (Sd) SATURAÇÃO DE, (T) TEMPERATURA, (FdP) FREQUÊNCIA DE PULSO