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. For this purpose, we used data from the Hospital das Clínicas at UFMG. Throughout this work, we will carry out a process of organizing and cleaning the data to extract both patient *features* and bacterial *targets* identified through bacterial culture results. To achieve this, we held weekly meetings with doctors from the Hospital das Clínicas, specialists in the hospital's infectious disease area. This work enabled us to build primary databases that served as input and output for a *Gradient Boosting*-based model, trained using ROC (Receiver Operating Characteristic) AUC (Area Under Curve) as the optimization metric. Throughout this work, we underwent an extensive data cleaning and aggregation process to ensure that the data could serve as a foundation for this and other Artificial Intelligence (AI) studies in healthcare. Given the complexity of the problem, our models achieved reasonable results.**

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, 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 deste trabalho realizaremos um processo de organização e limpeza dos dados para que possamos obter tanto *features* do paciente, quanto *targets* de bacterias identificadas através de resultados de culturas bacterianas. Construiremos bases de dados primária que serviram de entrada e saída para um 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). Além disso, antes de treinarmos o modelo passaremos por um trabalho de engenharia de *features*.

Para isso, temos como objetivos específicos a exploração inicial dos dados, 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 um extensivo processo de limpeza e agrupamento dos dados para que possam servir de base para esse e outros estudos para Inteligência Artificial (IA) na área da saúde. Esse processo de limpeza passa tanto pela conversão de valores e correção/eliminação de valores inválidos e *outliers*, quanto pela manipulação e agrupamento de dados.

Com nosso dado limpo e tratado, partiremos para um treinamento inicial de modelos para que possamos ter noções preliminares da aplicabilidade do dado para o problema. 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 construção dos datasets primários, limpeza dos dados, engenharia de *features* e treinamento dos primeiros modelos. Ademais teremos uma sessão de conclusões e discussões 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*, variaveis, 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. Engenharia de Features

Segundo (DONG, 2018), engenharia de *features* pode envolver processos como transformação, geração, extração, seleção, análise e avaliação de features. Segundo eles, a transformação de *features* é como a construção de novas features a partir das existentes, frequentemente utilizando mapeamentos matemáticos, alguns desses mapeamento podem incluir médias, medianas e outras estatísticas. Enquanto isso, a seleção de *features* é descrita como a escolha de um conjunto reduzido de *features* a partir de um grande conjunto inicial. Essa redução torna computacionalmente viável o uso de alguns algoritmos e pode melhorar a qualidade dos resultados obtidos. Neste trabalho, a seleção de features foi realizada manualmente com base no conhecimento especializado dos médicos infectologistas e as transformações partiram de funções diversas, como máximos, mínimos, médias, medianas e etc.

C. 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.

D. 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 garantir que nosso modelo seja bem avaliado e generalize bem para novos dados.

E. 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.

III. Desenvolvimento

A. Construção dos datasets primários

Inicialmente existiam diversas tabelas, que não estavam conectadas entre si, de áreas distintas 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)

Os dados de interesse para o trabalho podem ser agrupados em: Atendimentos, Evoluções, Exames Laboratoriais e Sinais Vitais. Entre as tabelas existentes foram selecionadas bases que continham informações sobre atendimentos, especialidades médicas, leitos, altas, origem de atendimentos, pacientes, prestadores de serviço, procedimentos do SUS, serviços, setores, acomodações e internações. A base de atendimento utilizada foi gerada a partir da junção dessas tabelas, 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.

Além das tabelas descritas, existiam ainda bases com evoluções dos pacientes, que eram compostas por descrições textuais dos atendimentos. Nesse escopo do projeto essa base não foi utilizada, apesar de outras linhas que exploram o tratamento dos dados de evoluções utilizando *Large Language*

TABLE II: Base de Laboratórios

Coluna	Valor Absoluto (%)
prontuarios	83932
pedidos	388591
procedimento	
HMG	5719334 (35.22)
UROT	1750195 (10.78)
GASOA	640439 (3.94)
CREASA	598371 (3.68)
TAP	338993 (2.09)
BILISA	301482 (1.86)
UROC	234282 (1.44)
GASOVS	231979 (1.43)
ATBURO	224140 (1.38)
UREISA	218581 (1.35)
TRIGSA	202354 (1.25)
COLESA	197394 (1.22)
COLHDL	195671 (1.21)
COLLDL	193583 (1.19)
CONHDL	186120 (1.15)
others	5005304 (30.82)

TABLE III: Base de Sinais Vitais

Coluna	Valor Absoluto (%)
id_atendimentos	46472
sinal vital	
TEMPERATURA(C)	1263915 (16.93)
FREQUENCIA CARDIACA	1232306 (16.50)
FREQUENCIA RESPIRATORIA	1151504 (15.42)
PRESSÃO ARTERIAL SISTOLICA	1114870 (14.93)
PRESSÃO ARTERIAL DISTOLICA	1113045 (14.90)
SATURAÇÃO DE OXIGÊNIO	555652 (7.44)
GLICEMIA CAPILAR	286651 (3.84)
PAM	285208 (3.82)
OXIGENIOTERAPIA	91214 (1.22)
FIO2	88314 (1.18)
PEEP	87686 (1.17)
PPI	78664 (1.05)
FR VENTILADOR	44613 (0.60)
PS	29987 (0.40)
CAPNOGRAFIA	19588 (0.26)
others	24479 (0.33)

Models (LLMs) foram desenvolvidas em paralelo.

Uma base de resultados laboratoriais do hospital também foi utilizada e nela estavam contidos resultados de diversos exames. Cada linha dessa tabela representava linhas de exames digitalizadas pelo hospital, a partir de resultados apresentados no formato PDF. Na Tabela II, podemos ver a distribuição dos exames antes dos tratamentos realizados. Assim, fez-se necessário o tratamento dessas linhas para a reconstrução e seleção dos exames de interesse. De maneira similar, também existia uma base com sinais vitais de pacientes, que trazia diversos valores coletados dos pacientes ao longo do tempo — como pressão arterial, frequência cardíaca, saturação de oxigênio, frequência respiratória, etc. Na Tabela III podemos ver a distribuição dos sinais vitais existentes.

A principal contribuição deste trabalho foi o entendimento e tratamento dos dados descritos acima, para que possam ser utilizados por modelos de predição. Assim, o objetivo é, a partir do tratamento desses dados, gerar um conjunto de features sobre os pacientes em datas específicas de atendimento, que serão utilizadas pelo modelo, para tentar prever a bactéria identificada em seu resultado de cultura bacteriana, sendo a presença ou ausência de bactéria o *target* do modelo. Para isso foram realizados diversos tratamentos e manipulações das tabelas existentes, descritos em detalhes nos próximos parágrafos.

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

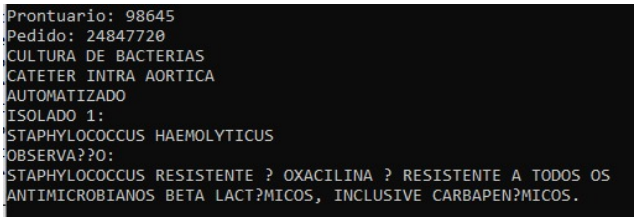

Fig. 2: Resultado de cultura bacteriana (2).

O primeiro passo foi montar o *target* a partir da base de laboratórios. Para isso, seguindo conhecimento especialista dos médicos infectologistas do Hospital das Clínicas, foram filtrados procedimentos do tipos: *UROC, HEMOC e CULBAC*, que representavam resultados de culturas bacterianas. Com a base filtrada, foi realizada a junção de diversas linhas para que fosse possível reconstruir os resultados de cultura, 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.

Assim, com o auxílio de expressões regulares (regex) e uma lista de nomes de bactérias conhecidas, foi possível gerar uma tabela de *target* que possuia um prontuário, uma data e um resultado de cultura, com o nome de bacteria ou um resultado negativo. Para a utilização nos modelos, separamos o dado por bacteria e colocamos 1 se infectado e 0 se não infectado. A lista de bactérias foi construída com base em conhecimento

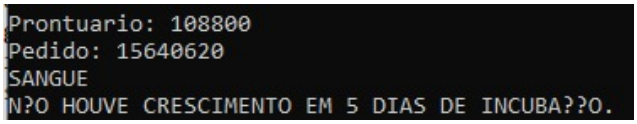

Fig. 3: Resultado de cultura bacteriana (3).

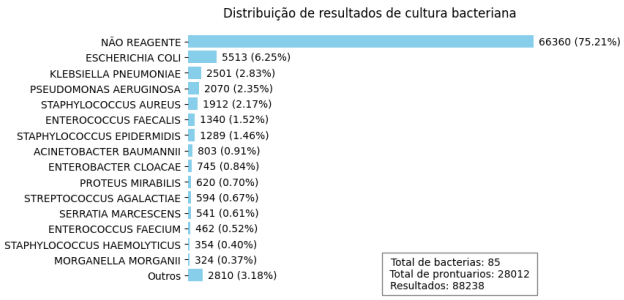

Fig. 4: Distribuição de resultados de cultura do target.

especialista e inclui 850 espécies de bactérias. A lista completa se encontra disponível para download no link.

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%). Modelos para predição dessas bactérias podem funcionar melhor em espécies que aparecem com maior frequência na nossa base de *target*. Dessa maneira, com os resultados de cultura bacteriana tratados, realizou-se o filtro da base de atendimento considerando apenas prontuários que possuíam algum tipo de resultado de cultura bacteriana, fossem eles positivos ou negativos.

Com o *target* montado e os pacientes-alvo identificados, o próximo passo foi seguir com o enriquecimento das *features* do paciente. 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. Assim, foram adicionadas às *features* do paciente, filtradas a partir da tabela de exames laboratoriais, resultados de exames que possuíam os seguintes tipos:

UREISA	**CREASA**	**PTFRSA**	**BILISA**
GASOA	**GASOVS**	**SODISA**	**CASA**
CLORSA	**POTASA**	**GLIC**	**LAC**
PCR	**MAGNSA**	**ALTSA**	**GGT**
PTTA	**TAP**	**HMG**	**FALC**
FIBR	**FERRIT**	**TROPUL**	**PROBNP**
DIMERO			

Uma vez que os exames de cultura bactérias vêm da escaneação de exames no formato PDF, eles podem conter mais de uma linha e diversos valores atribuídos. Assim, cada valor existente foi relacionado a sua linha de referência, por exemplo, exames do tipo GASOA possuem várias linhas, que foram representadas como: *GASOA-5, GASOA-6, GASOA-8*.... Cada um desses valores será considerado uma *features* distinta do paciente por trazerem valores distintos dentro de um mesmo exame, cada qual com sua própria unidade e ordem de grandeza.

TABLE IV: Tratamento Inicial Features

Original	Tratamento	Resultado
—-	Removido	-
>>>>>	Removido	-
—	Removido	-
...	Removido	-
.	Removido	-
NO RESULT	Removido	-
RESULTADOFORMATADO	Removido	-
SEXO INDETERMINADO	Removido	-
SEXO FEMININO	Removido	-
SEXO MASCULINO	Removido	-
,	Substituido	.
<x	Substituido	x
>x	Substituido	x
MAIOR QUE x	Substituido	x
ACIMA DE x	Substituido	x
SUPERIOR A x	Substituido	x
INFERIOR A x	Substituido	x

Além do resultados do exame, de maneira similar, os dados de sinais vitais coletados dos pacientes foram utilizados, filtrando os prontuários de interesse. Para o treinamento dos modelos, além do filtro de prontuário pegamos também datas de coleta que coincidissem com as culturas bacterianas desejadas. Assim, os dados de sinais vitais já descritos anteriormente foram agregados à tabela de *features*. Desse modo, ao final dessa junção e tratamento inicial, foi gerada uma tabela com prontuários, datas e diversos resultados de exames e sinais vitais para pacientes com resultados de culturas bacterianas.

Assim, com o dado agrupado, seguimos para tratamentos específicos, uma vez que ainda foram necessários outros tratamentos, como a remoção de caracteres inválidos e a substituição de vírgulas por pontos para possibilitar a conversão de valores. Além disso, alguns resultados de exames vinham como valores do tipo <10, >300, <20, 'MAIOR QUE 90', 'ACIMA DE 1000' e etc, esses casos também foram tratados, sendo feita a substituição das *strings* pelo valor numérico delimitante. Dessa maneira, valores foram substituídos por seus limites, assim valores >x, <x, MAIOR QUE x, MENOR QUE x, tornaram-se x. O tratamento inicial da base de features está representado na Tabela IV.

Assim, ao final desse tratamento inicial já era possível ter uma ideia da caracterização de nossos dados, na Tabela VIII temos os campos gerados para a tabela de features. Com os campos em mãos e convertidos para valores númericos, calculamos estatísticas de cada coluna, como, contagem de valores nulos e não nulos e suas respectivas porcentagens, quantidade de valores únicos, valor médio, mediana, valor mínimo, valor máximo, primeiro e terceiro quartil. A tabela em questão é bastante extensa e pode ser encontrada para download no link.

Dessa maneira já tínhamos uma boa visão do dado e podíamos passar para a etapa de limpeza e engenharia de features.

B. Limpeza dos dados

Após o tratamento dos dados temos uma boa visão estatística da *features* existentes. Assim, foi possível identificar algumas inconsistências do nosso dado, que podem ter se originado de problemas tanto no preenchimento dos campos por funcionários do Hospital, quanto por problemas do processo de escaneamento dos exames e outros fatores fora do escopo deste trabalho.

As estatísticas foram analisadas pelos médicos infectologistas que sugeriram algumas limpezas iniciais, como a eliminação de colunas que não seriam úteis para nosso problema e colunas que pareciam estar preenchidas de maneira inadequada e seriam, portanto, inválidas. Além disso, foram eliminados também resultados altamente discrepantes do esperado para os exames — por exemplo *feature* temperatura continha como valor máximo um valor de 3708ºC, provavelmente fruto de um erro de digitação do valor 37.08ºC.

De maneira similar, foram removidos *outliers* de colunas como a referente a frequência cardíaca, que apresentava o valor máximo de 8836 batimentos por minuto (bpm), e a de frequência respiratória, com valor máximo de 3293 movimentos respiratórios por minuto (mrm), valores que são humanamente impossíveis. Nesses casos, foram utilizados limites inferiores e superiores fixos considerando o conhecimento médico.

Por outro lado, colunas como *PEEP, volume corrente, PS e FR ventilador*, segundo conhecimento especialista dos médicos, apresentavam dados que não seriam relevantes para discernir o resultado da cultura e, por isso, foram removidas da base. Porém, para algumas colunas não foi necessário nenhum tipo de tratamento. Ao final do tratamento foi realizada a remoção de linhas com resultados de exame que possam ter ficado completamente nulas.

Todos os critérios para tratamentos e limpeza dos dados estão disponíveis na Tabela IX. Em resumo, partindo das colunas existentes na Tabela VIII, foram realizadas as limpezas sugeridas pelo conhecimento especialista dos médicos infectologistas da Tabela IX. Com os dados limpos, é possível ver na Tabela X estatísticas, como contagem de valores nulos e não nulos e suas respectivas porcentagens, quantidade de valores únicos, valor médio, mediana, valor mínimo, valor máximo, primeiro e terceiro quartil das colunas finais da nossa base de *features*.

C. Engenharia de Features

Assim, com a tabela primária de *features* limpa foi possível prosseguir para o treinamento de modelos a partir do dado. Uma particularidade desse dado é que alguns exames ou medições do paciente podem ser aferidas múltiplas vezes por dia — como por exemplo dados de temperatura, frequência cardíaca e respiratória. Dessa maneira, não é incomum que existam na base de *features* diversos registros distintos de um mesmo parâmetro realizados em no mesmo dia para um só

	data_evento	prontuario	FREQUENCIA CARDIACA	FREQUENCIA RESPIRATORIA	...	TEMPERATURA(C)
71230	2017-05-16 06:00:00	1095130.0	NaN	NaN	...	NaN
71231	2017-05-16 08:00:00	1095130.0	100.0	20.0	...	36.7
71232	2017-05-16 11:00:00	1095130.0	NaN	NaN	...	NaN
71233	2017-05-16 12:00:00	1095130.0	NaN	NaN	...	36.0
71234	2017-05-16 16:04:00	1095130.0	92.0	19.0	...	36.2
71235	2017-05-16 18:47:00	1095130.0	NaN	NaN	...	36.8
71236	2017-05-16 20:00:00	1095130.0	115.0	20.0	...	37.2
71237	2017-05-16 21:53:00	1095130.0	NaN	NaN	...	NaN

Fig. 5: Exemplo de exames com varios valores ao longo do dia.

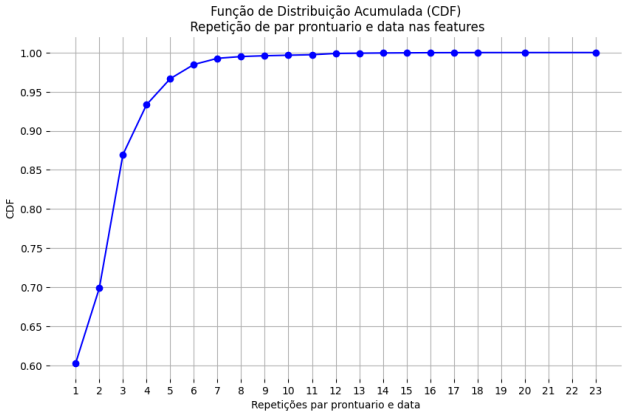

Fig. 6: Função de Distribuição Acumulada (CDF) da frequência de repetição de par prontuario e data na base limpa de features

paciente, como podemos ver representado na Figura 5 para o prontuário 1095130, em que o paciente possui diversas aferições desses valores ao longo do dia.

Na Figura 6, podemos ver a função de distribuição acumulada (CDF) da frequência de repetições do par prontuário e data na base limpa de *features*. A partir dela, podemos ver que 60% dos prontuários tem apenas um registro de resultado de exame por dia, porém os outros 40% possuem mais de um resultado de exame por dia. Apesar de existirem alguns *outliers* com pacientes com até 23 resultados distintos de exame em um único dia, cerca de 97% dos paciente tem no máximo 5 resultados de exames num único dia e desses cerca de 87% aparecem possuem até 3 resultados de exame em um dia.

Assim, porque temos como *target* um resultado de cultura de um paciente em um certo dia, é necessário selecionar e tratar quais valores do dia devemos levar em consideração em nosso modelo. Uma pergunta que surge é qual a melhor maneira de selecionar ou sumarizar os dados de um paciente ao longo do dia? Assim, entramos então no assunto de engenharia de *features*.

Trataremos as *features* em dois conjuntos distintos: o primeiro *features* que aparecem no máximo uma vez ao dia para todos pacientes da base, como VI, e o segundo *features* que aparecem mais de uma vez ao dia para algum paciente da base V.

Para o segundo grupo aplicamos diversos métodos de agrupamento para o par prontuário e dia, seguindo a sugestão dos médicos selecionamoso primeiro valor aferido no dia e o valor mais discrepante da média do paciente. Além disso,

TABLE V: *Features* aparecem mais de uma vez ao dia para algum paciente da base

Feature
ALTURA
FIO2
FREQUENCIA CARDIACA
FREQUENCIA RESPIRATORIA
FREQUÊNCIA DE PULSO (M2BR)
FREQUÊNCIA RESPIRATÓRIA (M2BR)
GLICEMIA CAPILAR
PAM
PAP
PCP
PPI
PRESSÃO ARTERIAL DISTOLICA
PRESSÃO ARTERIAL SISTOLICA
PRESSÃO INTRA-ARTERIAL
PVC
SATURAÇÃO DE O2 (M2BR)
SATURAÇÃO DE OXIGÊNIO
TEMPERATURA(C)
INDICE CARDÍACO

	data_evento	prontuario	TEMPERATURA(C)	FREQUENCIA RESPIRATORIA
629	2017-07-22 00:00:00	44107.0	36.5	NaN
630	2017-07-22 02:30:00	44107.0	38.3	NaN
631	2017-07-22 03:00:00	44107.0	38.0	NaN
632	2017-07-22 06:00:00	44107.0	37.0	NaN
633	2017-07-22 08:00:00	44107.0	36.7	17.0
634	2017-07-22 12:00:00	44107.0	36.5	NaN
635	2017-07-22 14:00:00	44107.0	36.5	16.0
636	2017-07-22 17:00:00	44107.0	NaN	NaN
637	2017-07-22 18:00:00	44107.0	NaN	NaN
638	2017-07-22 20:00:00	44107.0	37.7	12.0

prontuario	44107.0
dia_evento	2017-07-22
FREQUENCIA RESPIRATORIA_first_value	17.0
FREQUENCIA RESPIRATORIA_last_value	12.0
FREQUENCIA RESPIRATORIA_mean	15.0
FREQUENCIA RESPIRATORIA_std	2.645751
FREQUENCIA RESPIRATORIA_median	16.0
FREQUENCIA RESPIRATORIA_max	17.0
FREQUENCIA RESPIRATORIA_min	12.0
FREQUENCIA RESPIRATORIA_most_discrepant_value	12.0
TEMPERATURA(C)_first_value	36.5
TEMPERATURA(C)_last_value	37.7
TEMPERATURA(C)_mean	37.15
TEMPERATURA(C)_std	0.740656
TEMPERATURA(C)_median	36.85
TEMPERATURA(C)_max	38.3
TEMPERATURA(C)_min	36.5
TEMPERATURA(C)_most_discrepant_value	38.3

Fig. 7: Exemplo de engenharia de *feature* para colunas que aparecem mais de uma vez ao dia para um mesmo paciente.

adicionamos valores como, o valor médio, mediano e o desvio padrão ao longo do dia, o valor máximo do dia, o valor mínimo do dia e o último valor aferido. Dessa maneira uma *feature* com múltiplos valores por dia foi transformada em diversas outras *features* e ao final do tratamento cada paciente possui no máximo uma medição por dia de cada uma dessas novas *features*. Esse processo está ilustrado na Figura 7, nela podemos percerber que o prontuario 44107 possuí diversos resultados ao longo do dia, tendo sua frequência respiratória medida as 8:00 da manhã, 14:00 da tarde e 20:00 da noite, e outras diversas aferições de temperatura à meia noite, às 2:30 da manhã, as 3:00 da manhã, às 6:00 da manhã e assim por diante. Dessa maneira, os 10 registros distintos foram agrupados em apenas um registro para o dia, assim, se equivalendo à primeira categoria de *features*.

Dessa maneira, agora se um paciente possui mais de um registro para uma determinada data, eles representam resultados de exames diferentes, nos resta apenas agregar esses valores diferentes em um mesmo registro. Com isso, possuímos as features tratadas e limpas e prontas para serem usadas para o treinamento do nosso modelo, estatisticas das *features* após esse processo de engenharia estão disponíves nas Tabelas XI e XII.

TABLE VI: *Features* que aparecem no máximo uma vez ao dia para todos pacientes da base

Feature
IMC
PADRÃO ATIVIDADE (M2BR)
PERFUSÃO CAPILAR (M2BR)
PFE (M2BR)
PRESSÃO INTRA-ABDOMINAL
ALTSA-1
BILISA-1
BILISA-2
BILISA-3
CLORSA-1
CREASA-1
FERRIT-1
FIBR-1
GASOA-1
GASOA-2
GASOA-3
GASOA-4
GASOA-6
GASOA-7
GGT-1
GLIC-1
HMG-2
LAC-1
MAGNSA-1
PCR-1
POTASA-1
PROBNP-1
PTTA-1
PTTA-2
SODISA-1
TAP-4
TROPUL-1
UREISA-1

D. Treinamento dos primerios modelos

Com o dado tratado foi possível começar o treinamento de modelos e para isso utilizamos a implementação do *LightGBM* implementado pela biblioteca python, disponível no link. Realizamos o treinamento de um modelo para algumas das bactérias existentes na base. Pela na Figura 4 é possível perceber que existem muito mais resultados negativos de cultura do que positivos, por isso para melhorar a qualidade dos modelo realizamos o *oversampling* de resultados positivos, de maneira que durante nosso treino houvessem mais quantidade exemplos de culturas positivas para cada bactéria.

Além disso, alguns tipos de bactérias não possuem *samples* o suficiente para o treinamento de seu modelo, por isso optamos por treinar as versões iniciais dos modelos apenas para as 38 bactérias mais frequentes na base. O nome das bactérias pode ser visto na tabela VII, juntamento com os resultados dos modelos.

TABLE VII: Resultados modelos por bacteria

Bacteria	#Samples	#Train	#Test	Roc Auc	Sensitivity	Specificity	F1 score
STAPHYLOCOCCUS COAGULASE NEGATIVE	20	16138	16138	0.885 ± 0.134	0.999 ± 0.000	0.350 ± 0.255	0.395 ± 0.282
STAPHYLOCOCCUS CAPRAE	11	16152	16152	0.747 ± 0.278	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
STAPHYLOCOCCUS SAPROPHYTICUS	92	16024	16024	0.728 ± 0.044	0.993 ± 0.002	0.011 ± 0.022	0.012 ± 0.024
HAEMOPHILUS INFLUENZAE	45	16098	16098	0.715 ± 0.080	0.998 ± 0.001	0.022 ± 0.044	0.029 ± 0.057
SALMONELLA SPP	154	15924	15924	0.706 ± 0.052	0.990 ± 0.004	0.059 ± 0.025	0.066 ± 0.020
ESCHERICHIA COLI	3812	10072	10072	0.681 ± 0.010	0.544 ± 0.021	0.722 ± 0.026	0.583 ± 0.009
ACINETOBACTER BAUMANNII	378	15566	15566	0.682 ± 0.026	0.942 ± 0.003	0.169 ± 0.058	0.126 ± 0.041
STREPTOCOCCUS AGALACTIAE	393	15542	15542	0.683 ± 0.029	0.891 ± 0.006	0.287 ± 0.058	0.144 ± 0.026
STREPTOCOCCUS ORALIS	127	15968	15968	0.664 ± 0.052	0.982 ± 0.005	0.087 ± 0.029	0.073 ± 0.034
STAPHYLOCOCCUS AUREUS	653	15126	15126	0.657 ± 0.015	0.862 ± 0.012	0.286 ± 0.022	0.175 ± 0.018
STAPHYLOCOCCUS EPIDERMIDIS	636	15152	15152	0.659 ± 0.017	0.855 ± 0.018	0.289 ± 0.023	0.169 ± 0.016
ENTEROCOCCUS FAECIUM	235	15794	15794	0.660 ± 0.017	0.972 ± 0.003	0.038 ± 0.009	0.035 ± 0.009
PROVIDENCIA STUARTII	16	16144	16144	0.646 ± 0.134	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
STREPTOCOCCUS SANGUINIS	12	16152	16152	0.647 ± 0.173	0.999 ± 0.000	0.067 ± 0.133	0.067 ± 0.133
PROVIDENCIA RETTGERI	25	16130	16130	0.643 ± 0.159	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
STAPHYLOCOCCUS CAPITIS	46	16096	16096	0.643 ± 0.084	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
STENOTROPHOMONAS MALTOPHILIA	91	16024	16024	0.637 ± 0.027	0.997 ± 0.001	0.022 ± 0.027	0.032 ± 0.039
STREPTOCOCCUS PNEUMONIAE	56	16080	16080	0.644 ± 0.031	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
SERRATIA MARCESCENS	259	15756	15756	0.613 ± 0.023	0.969 ± 0.007	0.093 ± 0.048	0.079 ± 0.031
PSEUDOMONAS AERUGINOSA	836	14832	14832	0.583 ± 0.016	0.825 ± 0.007	0.249 ± 0.029	0.156 ± 0.016
STAPHYLOCOCCUS LUGDUNENSIS	27	16128	16128	0.575 ± 0.094	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
CITROBACTER FREUNDII	71	16056	16056	0.574 ± 0.060	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
ENTEROCOCCUS FAECALIS	852	14808	14808	0.557 ± 0.022	0.812 ± 0.014	0.263 ± 0.037	0.160 ± 0.026
KLEBSIELLA PNEUMONIAE	1552	13686	13686	0.555 ± 0.013	0.734 ± 0.013	0.352 ± 0.022	0.250 ± 0.013
STAPHYLOCOCCUS HAEMOLYTICUS	208	15838	15838	0.556 ± 0.049	0.959 ± 0.007	0.063 ± 0.040	0.041 ± 0.028
MORGANELLA MORGANII	193	15862	15862	0.556 ± 0.031	0.966 ± 0.004	0.052 ± 0.028	0.038 ± 0.021
ENTEROBACTER CLOACAE	405	15522	15522	0.551 ± 0.019	0.905 ± 0.013	0.114 ± 0.026	0.066 ± 0.009
KLEBSIELLA OXYTOCA	105	16002	16002	0.555 ± 0.054	0.998 ± 0.001	0.038 ± 0.019	0.061 ± 0.031
STAPHYLOCOCCUS HOMINIS	107	16000	16000	0.551 ± 0.077	0.996 ± 0.002	0.019 ± 0.023	0.023 ± 0.028
PROTEUS MIRABILIS	368	15582	15582	0.544 ± 0.019	0.910 ± 0.014	0.101 ± 0.037	0.057 ± 0.018
STREPTOCOCCUS SALIVARIUS	10	16154	16154	0.537 ± 0.193	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
STREPTOCOCCUS CONSTELLATUS	12	16152	16152	0.544 ± 0.153	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
PROTEUS VULGARIS	16	16144	16144	0.544 ± 0.087	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
STREPTOCOCCUS GALLOLYTICUS	20	16138	16138	0.533 ± 0.166	0.999 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
CITROBACTER KOSERI	86	16032	16032	0.514 ± 0.060	0.998 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
STREPTOCOCCUS ANGINOSUS	58	16078	16078	0.510 ± 0.066	0.999 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
STAPHYLOCOCCUS WARNERI	34	16116	16116	0.449 ± 0.130	1.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
PROTEUS PENNERI	18	16142	16142	0.378 ± 0.137	0.998 ± 0.001	0.000 ± 0.000	0.000 ± 0.000

Assim, mapeamos os *targets* para cada modelo a partir da tabela *features* pós engenharia, selecionando *matches* de prontuário e data. Treinamos *LightGBM* com *targets* binários, ou seja, o paciente pode ou não ter sido infectado pela bactéria, e utilizamos a métrica ROC AUC, descrita em nosso referencial teórico, para a otimização do modelo. Durante o treinamento utilizamos do método de *k-fold Cross Validation*, com $k = 5$, e reportamos a média das 5 *folds* para métricas como ROC AUC, sensibilidade, especificidade e F1-Score. Os resultados dos modelos, juntamente com a quantidade de valores positivos e tamanho de conjuntos de treino e teste, podem ser encontrados na Tabela VII.

Em geral, temos a métrica ROC AUC acima de 0.5 para quase todos os modelos, o que indica os modelos funcionam melhor que a aleatoriedade, além disso, os melhores 5 modelos tiveram a métrica acima do 0.7, com o melhor deles se aproximando de uma ROC AUC de 0.9. Em geral os modelos possuem a métrica entre 0.5 e 0.6, o que indica que se aproximam da aleatoriedade, mais ainda sim são um pouco melhores do que ela.

Além disso, praticamente todos os modelos apresentam uma sensibilidade alta, quando não igual, proxima à 1, que indica que o modelo tem uma excelente capacidade em identificar corretamente as instâncias positivas de cultura. Por outro lado, a maioria dos modelos tem uma especificidade muito baixa, se aproximando do zero, o que significa que são péssimos em identificar corretamente as instâncias com resultados negativos para as culturas, talvez um problema fruto do balanceamento de carga. Tal comportamento, afeta bastante a nossa métrica de F1-Score, que apesar da alta sensibilidade é bastante baixa para os modelos, sinalizando que ainda é necessário melhorar aspectos dos modelos.

Assim, os resultados dos modelos são razoáveis, dado que, diferenciar bactérias utilizando apenas medidas de exames e sinais vitais, sem informações específicas dos prontuários, é uma tarefa extremamente difícil, que nem mesmo médicos conseguem tirar conclusões apenas com base em dados brutos de exames.

Além disso, dada a quantidade de dados que temos por bactérias, talvez a seleção de intervalos de apenas um dia tenha limitado demais a quantidade de *features* existentes para cada bactéria. Talvez se fossem realizados outro métodos de *match* entre a base de *features* e a base de *targets*, haveriam mais dados e consequentemente melhores modelos, ou quem

sabe reavaliar o método de *oversampling* também traria uma melhora de performance.

IV. Conclusões

Com o trabalho desenvolvido foi possível entregar uma base de dados limpa e estruturada que pode servir de base para esse e outros estudos para Inteligência Artificial (IA) na área da saúde. Devido ao enfoque no tratamento dos dados, os modelos treinados possuem performances razoáveis dada a complexidade do problema, porém provavelmente podem ser melhorados e apresentam um potencial elevado. Assim, trabalhos futuros podem envolver experimentações com outros hiperparâmetros para o treinamento dos modelos, além da experimentação com outras implementações da engenharia de *features* e agrupamentos e tratamento dos dados. Além disso, outras linhas de pesquisa que explorem o tratamento dos dados de evolução de paciente utilizando Large Language Models (LLMs), podem melhorar e muito a qualidade dos 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, Luan Borges, Luiz Henrique Melo e Vítor Corrêa Silva, pela colaboração e apoio durante a realização desse estudo.

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

TABLE VIII: Campos tabela primária de features pré tratamento

ALTSA-1	**ALTURA**	**BILISA-1**
BILISA-2	**BILISA-3**	**CAPNOGRAFIA**
CLORSA-1	**CREASA-1**	**CREASA-2**
CREASA-3	**DC**	**DIMERO-1**
FALC-1	**FERRIT-1**	**FIBR-1**
FIO2	**FR VENTILADOR**	**FREQUENCIA CARDIACA**
FREQUENCIA RESPIRATORIA	**FREQUÊNCIA DE PULSO (M2BR)**	**FREQUÊNCIA RESPIRATÓRIA (M2BR)**
GASOA-1	**GASOA-10**	**GASOA-11**
GASOA-12	**GASOA-13**	**GASOA-14**
GASOA-2	**GASOA-3**	**GASOA-4**
GASOA-5	**GASOA-6**	**GASOA-7**
GASOA-8	**GASOA-9**	**GASOVS-1**
GASOVS-2	**GASOVS-3**	**GASOVS-4**
GASOVS-5	**GASOVS-6**	**GASOVS-7**
GGT-1	**GLIC-1**	**GLICEMIA (M2BR)**
GLICEMIA CAPILAR	**HMG-10**	**HMG-11**
HMG-12	**HMG-13**	**HMG-14**
HMG-15	**HMG-16**	**HMG-17**
HMG-18	**HMG-19**	**HMG-2**
HMG-20	**HMG-21**	**HMG-22**
HMG-23	**HMG-24**	**HMG-31**
HMG-32	**HMG-33**	**HMG-34**
HMG-35	**HMG-36**	**HMG-37**
HMG-38	**HMG-39**	**HMG-5**
HMG-52	**HMG-6**	**HMG-7**
HMG-8	**HMG-9**	**IMC**
INDICE CARDÍACO	**LAC-1**	**MAGNSA-1**
OXIGENIOTERAPIA	**PADRÃO ATIVIDADE (M2BR)**	**PAM**
PAP	**PCP**	**PCR-1**
PEEP	**PERFUSÃO CAPILAR (M2BR)**	**PESO**
PFE (M2BR)	**PH URINÁRIO**	**PIC**
POTASA-1	**PPI**	**PRESSÃO ART DIASTÓLICA (M2BR)**
PRESSÃO ART SISTÓLICA (M2BR)	**PRESSÃO ARTERIAL DISTOLICA**	**PRESSÃO ARTERIAL SISTOLICA**
PRESSÃO INTRA-ABDOMINAL	**PRESSÃO INTRA-ARTERIAL**	**PROBNP-1**
PS	**PTFRSA-1**	**PTFRSA-2**
PTFRSA-3	**PTFRSA-4**	**PTTA-1**
PTTA-2	**PVC**	**SATURAÇÃO DE O2 (M2BR)**
SATURAÇÃO DE OXIGÊNIO	**SODISA-1**	**TAP-1**
TAP-2	**TAP-3**	**TAP-4**
TEMPERATURA (M2BR)	**TEMPERATURA(C)**	**TROPUL-1**
UREISA-1	**VOLUME CORRENTE**	**data_evento**
prontuario		

TABLE IX: Critérios de Limpeza das Features

Parâmetro	Critério de Exclusão de Outliers	Comentários
Temperatura	< 33 ou > 42	Excluir outliers
Frequência cardíaca	< 20 ou > 250	Excluir outliers
Frequência respiratória	< 5 ou > 60	Excluir outliers
Pressão arterial sistólica	< 40 ou > 280	Excluir outliers
Pressão arterial diastólica	< 5 ou > 160	Excluir outliers
Creasa-1	> 20	Excluir outliers
Ureisa-1	> 300	Excluir outliers
Saturação de oxigênio	< 40	Excluir outliers
Potasa-1	-	Sem modificações
Sodisa-1	-	Sem modificações
Creasa-2	-	Retirar do dataset
Creasa-3	-	Retirar do dataset
Magnsa-1	< 1 ou > 6	Excluir outliers.
Glicemia capilar	< 15 ou > 600	Excluir outliers
Altsa-1	-	Sem modificações
PCR-1	-	Sem modificações
Bilisa-1	-	Sem modificações
Bilisa-2	< 0	Excluir valores negativos
Bilisa-3	-	Sem modificações
TAP-1, 2, 3, 4	-	TAP4 RNI, manter e excluir 1,2 e 3
FALC-1	-	Retirar do dataset
GLIC-1	< 15 ou > 600	Excluir outliers
GGT-1	-	Sem modificações
PTTA-1	-	Sem modificações
PTTA-2	-	Sem modificações
GASOA-1 (pH)	-	Sem modificações
GASOA-2 (PaO2)	-	Sem modificações
GASOA-3 (PaCO2)	-	Sem modificações
GASOA-4 (HCO3)	-	Sem modificações
GASOA-6 (BEecf)	-	Sem modificações
GASOA-7 (Saturação de oxigênio)	< 40	Excluir outliers
GASOA-5, 8, 9, 10, 11, 12, 13, 14	-	Retirar do dataset
Clorsa-1	-	Sem modificações
Gasovs-1, 2, 3, 4, 5, 6, 7	-	Retirar do dataset
PTFRSA-1, 2, 3, 4	-	Retirar do dataset
PAM	< 20 ou > 200	Excluir outliers
Ferrit-1	-	Sem modificações
Oxigênioterapia	-	Retirar do dataset
Lac-1	-	Sem modificações
Peso	-	Retirar do dataset
Fibr-1	-	Sem modificações
FiO2	< 21	Excluir outliers
PEEP, volume corrente, PS, FR ventilador, capno...	-	Ocultar no dataset e excluir da análise
Dimero-1	-	Retirar do dataset
Probnp-1	-	Sem modificações
Pressão intra-arterial	-	Sem modificações
Tropul-1	-	Sem modificações
PH urinário	-	Retirar do dataset
PIC	-	Retirar do dataset
DC	-	Retirar do dataset
Glicemia (M2BR)	-	Retirar do dataset
Pressão art diastólica (M2BR)	-	Retirar do dataset
Pressão arterial sistólica (M2BR)	-	Retirar do dataset
Temperatura (M2BR)	-	Retirar do dataset
HMG	-	Retirar, manter HMG2 (leucócitos).

TABLE X: Estatísticas Tabela de Features Limpa

	#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