Aprendizagem de máquina aplicada à estratificação de risco de mortalidade de recém-nascidos associados a partos prematuros

Yasmin S. de P. Oliveira*,Gabriel R. de Freitas*, Wysterlânya K. P. Barros*, Luísa C. de Souza*
Karolayne S. Azevedo*, Raquel de M. Barbosa†, e Marcelo A. C. Fernandes*‡§
*InovAI Lab, nPITI/IMD, UFRN, Natal, RN, Brasil.
†Departamento de Farmácia (DEFAR), UFRN, Natal, RN, Brasil.
‡Centro Multiusuário de Bioinformática (BioME) - IMD, UFRN, Natal, RN, Basil.
§Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, RN, Basil.
Email: yasminsangela@hotmail.com, gabrielfreitas2601@gmail.com, kyurybarros@gmail.com, luisa.souza.103@ufrn.edu.br
karolayneazevsantos@gmail.com, raquel.melo@ufrn.br,mfernandes@dca.ufrn.br

***Resumo*—Este artigo apresenta um estudo que utiliza técnicas de aprendizado de máquina não supervisionado, com foco na técnica t-SNE, para estratificar o risco de mortalidade em recém-nascidos prematuros. A metodologia adotada envolve a coleta de dados das bases governamentais do Sistemas de Informações sobre Nascidos Vivos (SINASC) e do Sistemas de Informações sobre Mortalidade (SIM), a seleção de variáveis relevantes e a criação de um conjunto de dados unificado. Após a remoção de outliers e dados faltantes, o conjunto de dados final contém informações de 52.673 recém-nascidos. A técnica t-SNE é aplicada para identificar padrões e estruturas nos dados, permitindo a estratificação do risco de mortalidade neonatal. Os resultados obtidos são avaliados e interpretados, destacando as variáveis mais relevantes para a estratificação do risco. Essa abordagem inovadora tem o potencial de aprimorar os cuidados e os resultados de saúde para recém-nascidos prematuros, fornecendo insights valiosos para a tomada de decisões informadas no âmbito da saúde pública.**

***Index Terms*—Aprendizado não supervisionado, estratificação de risco, prematuridade, t-SNE, mortalidade neonatal**

I. Introdução

Complicações do parto prematuro, definido como um nascimento que ocorre antes da 37ª semana de gestação, são a causa mais comum de mortalidade em crianças de 5 anos de idade ou menos [1], [2]. Além disso, foi demonstrado como um fator crítico para a sobrevivência de recém-nascidos [1]. Bebês nascidos prematuramente apresentam um grande desafio para a assistência médica, que precisa complementar seus órgãos vitais ainda não totalmente desenvolvidos [3].

A utilização de técnicas de aprendizado de máquina (Machine Learning - ML) na predição e estratificação de mortalidade em recém-nascidos prematuros desempenha um papel crucial na área da saúde. A aplicação dessas técnicas permite uma análise abrangente dos dados, classificando padrões e relacionamentos complexos entre as variáveis que podem influenciar o risco de mortalidade neonatal. Existem vários trabalhos na literatura que tem utilizado técnicas de ML supervisionadas e não-supervisionadas na predição e estratificação de risco de mortalidade em recen-nascidos prematuros como descrito nos trabalhos apresentados em [4]–[11].

Ao utilizar algoritmos de aprendizado de máquina não supervisionados, como o t-SNE (*t-Distributed Stochastic Neighbor Embedding*), é possível extrair informações valiosas dos dados disponíveis, permitindo a identificação precoce de bebês prematuros com maior vulnerabilidade [12], [13]. Isso facilita a adoção de medidas preventivas e intervenções adequadas, direcionando recursos e esforços para melhorar os resultados de saúde e reduzir a mortalidade nessa população vulnerável. Essa abordagem inovadora fortalece os esforços para aprimorar a qualidade dos cuidados neonatais e contribui para a tomada de decisões informadas e eficazes no âmbito da saúde pública [14], [15].

Um estudo apresentado em [7] desenvolveu um modelo de previsão para a mortalidade neonatal em prematuros, identificando fatores pré-natais que contribuem para esse risco. Eles descobriram que a diminuição da idade gestacional é um fator preditivo forte para o aumento da taxa de mortalidade. Além disso, fatores como classificação como pequeno para a idade gestacional, oligoidrâmnio, transtornos psiquiátricos maternos, terapia antibiótica pré-natal e gêmeos monocoriônicos também foram identificados como contribuintes significativos para o aumento da mortalidade. Outro estudo relevante apresentado em [8], que realizou uma análise sistemática de métodos de aprendizado de máquina para a previsão da mortalidade neonatal em prematuros. Eles destacaram a importância da sensibilidade dos classificadores na avaliação desses modelos de previsão. Por meio dessa análise, eles puderam identificar os métodos de aprendizado de máquina mais eficazes para a previsão da mortalidade neonatal, fornecendo insights valiosos para a melhoria dos cuidados e dos resultados para esses bebês vulneráveis.

O trabalho apresentado em [9] os autores avaliaram o índice preditivo de mortalidade e morbidade em neonatos prematuros. Eles compararam os escores CRIB e CRIBII na previsão de mortalidade e morbidade em recém-nascidos prematuros,

demonstrando que esses escores podem ser utilizados como ferramentas eficazes para prever complicações e melhorar os resultados em neonatos prematuros. Ja em [10] foi proposto um modelo de floresta aleatória para prever o risco de mortalidade em bebês prematuros. O estudo abordou a limitação dos modelos de previsão clínica existentes que não levam em conta o restante do curso hospitalar. O trabalho mostrou que o modelo de floresta aleatória tem um desempenho superior em comparação com modelos convencionais, com sensibilidade de 88% e área sob a curva (AUC) de 0,93.

Este trabalho tem como objetivo utilizar técnicas de aprendizado de máquina não supervisionado, especificamente o t-SNE, para estratificar o risco de mortalidade em recém-nascidos prematuros. A análise é baseada em dados do Sistemas de Informações sobre Nascidos Vivos (SINASC) e do Sistemas de Informações sobre Mortalidade (SIM), que são sistemas de informações sobre nascidos vivos e mortalidade, respectivamente, referentes aos anos de 2015, 2016 e 2017. As colunas selecionadas para análise incluem informações sobre idade da mãe, quantidade de filhos, semanas de gestação, tipo de parto, número de consultas pré-natal, entre outras. O objetivo é identificar as variáveis mais relevantes que podem contribuir para a estratificação do risco de mortalidade em recém-nascidos prematuros, o que pode ser fundamental para direcionar intervenções e cuidados adequados para essa população vulnerável.

Este trabalho permite uma abordagem avançada e abrangente na análise e estratificação do risco de mortalidade em recém-nascidos prematuros. Utilizando técnicas de aprendizado de máquina não supervisionado, como o t-SNE, é possível identificar padrões e relações complexas entre as variáveis dos dados, auxiliando na identificação precoce de bebês prematuros com maior vulnerabilidade. Essa estratificação de risco proporciona uma oportunidade de intervenção e cuidados personalizados, direcionando recursos e esforços para garantir o melhor resultado possível para esses bebês e suas famílias. Além disso, contribui para aprimorar as práticas de saúde pública, oferecendo subsídios para o desenvolvimento de políticas e programas de prevenção e tratamento mais eficazes, com o objetivo de reduzir a mortalidade neonatal em recém-nascidos prematuros.

II. t-SNE

O t-SNE é uma técnica de aprendizado de máquina não supervisionado utilizada para visualização de dados de alta dimensionalidade. Foi proposto por Maaten e Hinton em 2008 [**?**] como uma extensão do método SNE (*Stochastic Neighbor Embedding*). O t-SNE mapeia os dados de entrada para um espaço de menor dimensão, preservando as relações de proximidade entre os pontos. É especialmente eficaz na visualização de agrupamentos e estruturas não lineares nos dados.

A principal ideia por trás do t-SNE é modelar a semelhança entre pares de pontos nos espaços de alta e baixa dimensão. Ele faz isso construindo distribuições de probabilidade para medir a proximidade dos pontos nas duas dimensões. No processo, o t-SNE calcula a similaridade entre os pontos de alta dimensão usando uma função de similaridade baseada na distância Euclidiana ou em outras métricas. Em seguida, ele mapeia essas similaridades para um espaço de baixa dimensão, onde a semelhança entre os pontos é medida usando uma distribuição t de Student.

O t-SNE possui dois parâmetros principais: a perplexidade e a taxa de aprendizado (*learning rate*). A perplexidade controla o número de vizinhos considerados ao calcular as similaridades entre os pontos de alta dimensão. Uma perplexidade mais alta significa que mais vizinhos são levados em consideração, resultando em agrupamentos mais difusos. Já a taxa de aprendizado determina a rapidez com que o algoritmo ajusta as posições dos pontos no espaço de baixa dimensão. É importante encontrar um equilíbrio adequado para obter uma visualização coerente.

O t-SNE tem sido amplamente utilizado em diversas áreas, como análise exploratória de dados, reconhecimento de padrões, processamento de imagens e bioinformática. Sua capacidade de revelar estruturas complexas e agrupamentos não lineares torna-o uma ferramenta valiosa para a visualização e interpretação de conjuntos de dados complexos. No entanto, é importante ressaltar que o t-SNE é um algoritmo estocástico e pode produzir resultados diferentes em diferentes execuções, portanto, é recomendado realizar várias execuções e avaliar os resultados de forma consistente [16], [17].

III. Metodologia

O presente estudo utilizou técnicas de aprendizado de máquina não supervisionado, com ênfase na técnica t-SNE, para estratificar o risco de mortalidade em recém-nascidos prematuros. A metodologia adotada compreendeu as seguintes etapas.

Inicialmente, foram coletados os dados das bases governamentais do SINASC e do SIM referentes aos anos de 2015, 2016 e 2017. Essas duas bases de dados foram unificadas em um único conjunto de dados a partir da variável "NUMERODN", referente ao número da Declaração de Nascido Vivo (DN).

Posteriormente, com base nos trabalhos apresentados em [18], [19], [20] e [21] foi realizada uma etapa de seleção de colunas (ou variáveis) relevantes para o estudo, incluindo idade da mãe (IDADEMAE), peso do feto em gramas (PESO), quantidade de gestações anteriores (QTDGESTANT), quantidade de partos normais anteriores (QTDPARTNOR), quantidade de partos cesáreos anteriores (QTDPARTCES), quantidade de filhos vivos (QTDFILVIVO), quantidade de filhos mortos (QTDFILMORT), semana de gestão (SEMAGESTAC), Apgar no primeiro minuto de vida (APGAR1) e no quinto minuto de vida (APGAR5), número de consultas pré-natal (CONSULTAS), Tipo de gravidez (GRAVIDEZ), índice de Kotelchuck (KOTELCHUCK), tipo de parto (PARTO), raça e cor da mãe (RACACORMAE) e sexo do feto (SEXO). Em seguida, foi criada uma nova coluna ou variável denominada "Vivo", que representa uma variável categórica com valores 0 ou 1. Essa coluna indica se o indivíduo estava vivo (1 - Sim) ou não

(0 - Não) após um ano do nascimento. Essa rotulação foi realizada com base nas informações do SIM no período de acompanhamento considerado. A Tabela I descreve de forma detalhada as variáveis associado ao datataset final utilizado.

O Índice de Kotelchuck é uma medida utilizada para avaliar a adequação do cuidado pré-natal recebido por uma gestante. Ele leva em consideração a quantidade de consultas pré-natais realizadas, o início do cuidado pré-natal, a regularidade das consultas e a duração do cuidado. O índice varia de 1 a 5 (ou de 0 a 4), sendo que valores mais altos indicam uma maior adequação do cuidado pré-natal. Um índice de Kotelchuck mais alto está associado a melhores resultados de saúde para a mãe e o bebê. Portanto, o Índice de Kotelchuck é uma ferramenta importante para avaliar a qualidade e a efetividade do cuidado pré-natal.

Após a realização de um processo de eliminação das informações faltantes e remoção de outliers (associados os dados numéricos) o dataset final conseguiu agregar a informação de 52.673 indivíduos (recen-nascidos). Foi utilizado um percentil de 95% para remoção de outliers. A Tabela II detalha a estatística das variáveis numéricas do dataset, ou seja, as variáveis: IDADEMAE, PESO, QTDGESTANT, QTDPARTNOR, QTDPARTCES, QTDFILVIVO, QTDFILMORT e SEMAGESTAC. Já a Tabela III detalha a estatística das variáveis categóricas, isto é APGAR1, APGAR5, CONSULTAS, GRAVIDEZ, KOTELCHUCK, PARTO, RACACORMAE, SEXO e Vivo.

Com base nas variáveis de IDADEMAE, PESO, QTDGESTANT, QTDPARTNOR, QTDPARTCES, QTDFILVIVO, QTDFILMORT, SEMAGESTAC, APGAR1, APGAR5, CONSULTAS, GRAVIDEZ, KOTELCHUCK, PARTO, RACACORMAE e SEXO aplicou-se a técnica de t-SNE para projetar os dados em um espaço de menor dimensão, preservando as relações e características relevantes. Essa técnica permitiu identificar padrões e estruturas intrínsecas nos dados, auxiliando na estratificação do risco de mortalidade (descrito na variável Vivo) em recém-nascidos prematuros.

Por fim, os resultados obtidos foram avaliados e interpretados com base nas características e relações identificadas. Destacaram-se as variáveis que apresentaram diferenças explícitas entre os grupos estratificados, sendo consideradas as mais importantes para a estratificação do risco de mortalidade. A metodologia adotada neste estudo, que combina técnicas de aprendizado de máquina com o uso de bases de dados governamentais, proporcionou uma análise abrangente e eficaz na estratificação do risco de mortalidade em recém-nascidos prematuros. Essa abordagem pode fornecer insights valiosos para aprimorar os cuidados e os resultados clínicos nessa população vulnerável.

IV. Resultados

A Figura 1 ilustra o resultado da aplicação da técnica t-SNE as 16 variáveis utilizadas, incluindo idade da mãe (IDADEMAE), peso do recém-nascido (PESO), quantidade de gestações anteriores (QTDGESTANT), quantidade de partos normais anteriores (QTDPARTNOR), quantidade de partos cesáreos anteriores (QTDPARTCES), quantidade de filhos vivos (QTDFILVIVO), quantidade de filhos mortos (QTDFILMORT), semanas de gestação (SEMAGESTAC), índice APGAR no primeiro minuto de vida (APGAR1), índice APGAR no quinto minuto de vida (APGAR5), número de consultas pré-natal (CONSULTAS), tipo de gravidez (GRAVIDEZ), índice de Kotelchuck (KOTELCHUCK), tipo de parto (PARTO), raça e cor da mãe (RACACORMAE) e sexo do feto (SEXO). Após a aplicação da técnica t-SNE, a variável "Vivo"foi utilizado para mapear o agrupamento realizado pelo t-SNE.

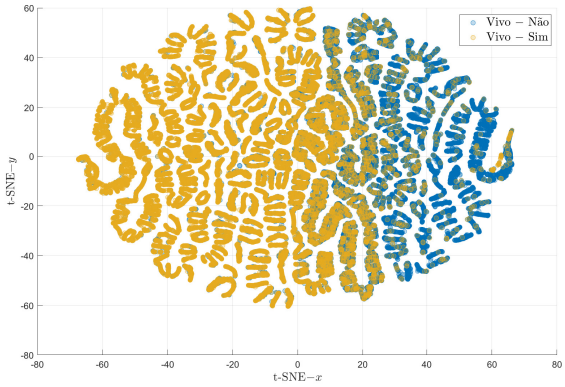

Figura 1: Resultado da aplicação da técnica t-SNE as 16 variáveis utilizadas.

Os resultados do mapeamento bidimensional gerado pelo t-SNE mostram que os recém-nascidos que vieram a óbito (Vivo - Não) se agruparam mais à esquerda do eixo horizontal (t-SNE$-x$), enquanto os que permaneceram vivos (Vivo - Sim) se agruparam mais à direita do eixo horizontal (t-SNE$-x$). Observa-se que a dimensão vertical (t-SNE$-y$) não contribui de forma significativa no agrupamento. Essa visualização permite uma melhor compreensão da relação entre as variáveis utilizadas e a variável "Vivo", auxiliando na identificação de possíveis padrões ou tendências.

Baseada nas informações obtidas através da técnica t-SNE, foi possível gerar um histograma do eixo horizontal (t-SNE$-x$) com o objetivo de identificar a quantidade de ocorrências em cada grupo, ou seja, entre os recém-nascidos classificados como "Vivo-Não"e "Vivo-Sim". A Figura 2 apresenta o histograma do eixo horizontal (t-SNE$-x$), revelando que a maior quantidade de ocorrências do grupo "Vivo-Não"(recém-nascidos que chegaram a óbito) está localizada à esquerda do eixo horizontal (t-SNE$-x$), enquanto a maior quantidade de ocorrências do grupo "Vivo-Sim"(recém-nascidos que não chegaram a óbito) está localizada à direita do eixo horizontal (t-SNE$-x$).

Esse histograma proporciona uma visualização mais clara e quantitativa da distribuição dos grupos em relação ao eixo horizontal (t-SNE$-x$), permitindo verificar a separação entre os recém-nascidos que chegaram a óbito e os que permaneceram vivos. Após a geração do histograma, foi calculado o valor

Tabela I: Descrição detalhada das variáveis associadas ao dataset utilizado.

Variáveis	Tipo	Descrição	Categorias
IDADEMAE	Numérico	Idade da mãe.	Nenhuma.
PESO	Numérico	Peso do feto em gramas.	Nenhuma.
QTDGESTANT	Numérico	Quantidade de gestações anteriores.	Nenhuma.
QTDPARTNOR	Numérico	Quantidade de partos normais anteriores.	Nenhuma.
QTDPARTCES	Numérico	Quantidade de partos cesáreos anteriores.	Nenhuma.
QTDFILVIVO	Numérico	Quantidade de filhos vivos.	Nenhuma.
QTDFILMORT	Numérico	Quantidade de filhos mortos.	Nenhuma.
SEMAGESTAC	Numérico	Semanas de gestação.	Nenhuma.
APGAR1	Categórico	Índice APGAR no 1º minuto de vida.	Valor inteiro de 0 a 10.
APGAR5	Categórico	Índice APGAR no 5º minuto de vida.	Valor inteiro de 0 a 10.
CONSULTAS	Categórico	Número de consultas pré-natal.	1 = Nenhuma. 2 = de 1 a 3. 3 = de 4 a 6. 4 = 7 e mais.
GRAVIDEZ	Categórico	Tipo de gravidez.	1 = Única. 2 = Dupla. 3 = Tripla e mais.
KOTELCHUCK	Categórico	Índice de Kotelchuck.	1 = Não tem. 2 = Inadequado. 3 = Intermediário. 4 = Adequado. 5 = Mais que adequado.
PARTO	Categórico	Tipo de parto.	1 = Vaginal 2 = Cesáreo
RACACORMAE	Categórico	Raça e cor da mãe.	1 = Branco 2 = Preto. 3 = Amarelo. 4 = Pardo. 5 = Indígena
SEXO	Categórico	Sexo do feto.	1 = Masculino. 2 = Feminino.
Vivo	Categórico	Indicador de recém-nascido vivo.	0 = Não 1 = Sim

Tabela II: Estatística das variáveis numéricas do dataset gerado e utilizado no trabalho.

Variáveis	Média	Mediana	Desvio padrão	Min	Max
IDADEMAE	25,52	25	6,66	10	39
PESO (gramas)	1823,61	2000	857,9	100	3395
QTDGESTANT	0,87	1	1,02	0	4
QTDPARTNOR	0,45	0	0,78	0	3
QTDPARTCES	0,24	0	0,52	0	2
QTDFILVIVO	0,66	0	0,87	0	4
QTDFILMORT	0,22	0	0,48	0	2
SEMAGESTAC	32,04	34	4,99	19	39

da divergência de Kullback-Leibler para identificar o melhor ponto de separação no eixo horizontal (t-SNE$-x$) entre as duas distribuições. O valor de Kullback-Leibler encontrado foi de t-SNE$-x \approx 2{,}7536$.

Essa análise estatística fornece uma métrica objetiva para determinar um ponto de corte que melhor discrimina os recém-nascidos que chegaram a óbito daqueles que permaneceram vivos no eixo horizontal (t-SNE$-x$). Os resultados obtidos indicam que a maioria das amostras rotuladas como "Vivo-Não"está concentrada em valores menores de t-SNE$-x$, enquanto a maioria das amostras rotuladas como "Vivo-Sim"está concentrada em valores maiores de t-SNE$-x$. Essa separação evidencia a capacidade do t-SNE em identificar padrões distintos entre os dois grupos, contribuindo para uma melhor

Tabela III: Estatística das variáveis categóricas do dataset gerado e utilizado no trabalho.

Variáveis	Categorias										
	Cat-0	Cat-1	Cat-2	Cat-3	Cat-4	Cat-5	Cat-6	Cat-7	Cat-8	Cat-9	Cat-10
APGAR1	1,77%	8,02%	6,23%	5,59%	5,68%	6,24%	7,14%	10,23%	22,8%	24,99%	1,3%
APGAR5	2,14%	4,16%	2,48%	2,29%	2,6%	3,83%	4,86%	8,31%	15,16%	33,46%	20,7%
CONSULTAS	–	2,47%	18,22%	39,07%	40,24%	–	–	–	–	–	–
GRAVIDEZ	–	87,34%	11,98%	0,68%	–	–	–	–	–	–	–
KOTELCHUCK	–	2,4%	23,16%	25,53%	11,6%	37,32%	–	–	–	–	–
PARTO	–	48,6%	51,4%	–	–	–	–	–	–	–	–
RACACORMAE	–	37,3%	6,29%	0,36%	55,41%	0,64%	–	–	–	–	–
SEXO	–	52,63%	47,37%	–	–	–	–	–	–	–	–
Vivo	49,73%	50,27%	–	–	–	–	–	–	–	–	–

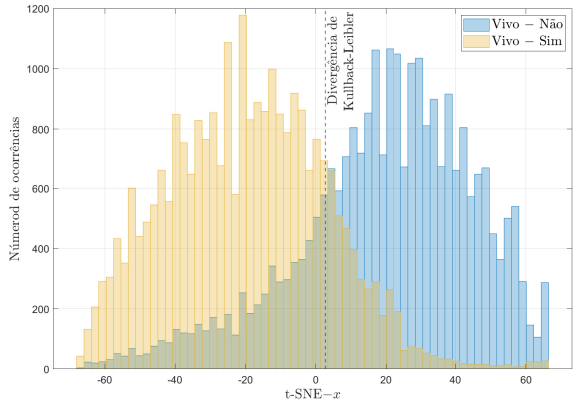

Figura 2: Histograma do eixo horizontal (t-SNE$-x$) representando as ocorrências dos grupos "Vivo-Não"e "Vivo-Sim".

compreensão das características relacionadas à mortalidade neonatal em recém-nascidos prematuros. Assim foram então criados dois grupos chamados de "Grupo-Vivo-Não"e "Grupo-Vivo-Sim". O "Grupo-Vivo-Não"possui um total de 20.848 amostras ($\approx$ 78.20% do total de amostras rotuladas como "Vivo-Não") e o "Grupo-Vivo-Sim"é formado um total de 22.166 amostras ($\approx$ 83.72% do total de amostras rotuladas como "Vivo-Sim")

As Tabelas IV e V apresentam as estatísticas das amostras numéricas com t-SNE$-x > 2{,}7536\%$ (amostras pertencentes ao "Grupo-Vivo-Não") e com t-SNE$-x < 2{,}7536\%$ (amostras pertencentes ao "Grupo-Vivo-Sim"), respectivamente. Pode-se observar que, no caso das variáveis numéricas, as variáveis peso (variável PESO) e semanas de gestação (variável SEMAGESTAC) tiveram resultados com diferenças estatísticas bem perceptíveis. A média do peso para "Grupo-Vivo-Não"ficou com $\approx$ 942 gramas e média para o "Grupo-Vivo-Sim"ficou com $\approx$ 2570 gramas. Ja para caso da variável SEMAGESTAC a média para "Grupo-Vivo-Não"ficou com $\approx$ 27 semanas e média para o "Grupo-Vivo-Sim"ficou com $\approx$ 35 semanas.

As Figuras 3 e 4 apresentam as estatísticas das variáveis categóricas associadas ao Grupo-Vivo-Não e ao Grupo-Vivo-Sim, respectivamente. Ao analisar os dados categóricos, observa-se diferenças significativas nas variáveis do índice APGAR no primeiro minuto de vida (APGAR1), índice

Tabela IV: Estatísticas das variáveis numéricas do subconjunto de dados Grupo-Vivo-Não.

ColumnsNames	Mean	Median	Std	Min	Max
IDADEMAE	25,32	25	6,72	11	39
PESO (gramas)	942,13	840	416	100	2.150
QTDGESTANT	0,83	1	1,02	0	4
QTDPARTNOR	0,41	0	0,75	0	3
QTDPARTCES	0,21	0	0,49	0	2
QTDFILVIVO	0,58	0	0,84	0	4
QTDFILMORT	0,26	0	0,52	0	2
SEMAGESTAC	27,39	27	4,2	19	39

Tabela V: Estatísticas das variáveis numéricas do subconjunto de dados Grupo-Vivo-Sim.

ColumnsNames	Mean	Median	Std	Min	Max
IDADEMAE	25,66	25	6,58	12	39
PESO (gramas)	2.569,73	2.485	381,33	1.565	3.395
QTDGESTANT	0,88	1	1,01	0	4
QTDPARTNOR	0,46	0	0,79	0	3
QTDPARTCES	0,26	0	0,54	0	2
QTDFILVIVO	0,7	0	0,88	0	4
QTDFILMORT	0,19	0	0,45	0	2
SEMAGESTAC	35,58	36	1,99	19	39

APGAR no quinto minuto de vida (APGAR5), número de consultas pré-natal (CONSULTAS) e índice de Kotelchuck (KOTELCHUCK). No Grupo-Vivo-Não, as ocorrências do APGAR1 e APGAR5 estão distribuídas em várias categorias, enquanto no Grupo-Vivo-Sim mais de 80% das ocorrências estão nas categorias 8, 9 e 10. Quanto ao número de consultas pré-natal, mais de 90% das ocorrências no Grupo-Vivo-Sim estão nas categorias 3 e 4, enquanto no Grupo-Vivo-Não ocorre uma redução significativa na categoria 4 (de $\approx$ 57% para $\approx$ 19%) e um aumento significativo na categoria 2 (de $\approx$ 9% para $\approx$ 30%). Em relação ao índice de Kotelchuck (KOTELCHUCK), aproximadamente 64% das amostras do Grupo-Vivo-Sim estão nas categorias 4 e 5, enquanto no Grupo-Vivo-Não esse valor é de apenas cerca de 29%.

Com base nos resultados apresentados nas Tabelas IV e V e Figuras 3 e 4 observa-se que as variáveis PESO, SEMAGESTAC, APGAR1, APGAR5, CONSULTAS, KOTELCHUCK possuem um papel bastante significativo na distinção entre os subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim. As Figuras 5, 6. 7, 8, 9 e 10 ilustram a distribuição e alguns

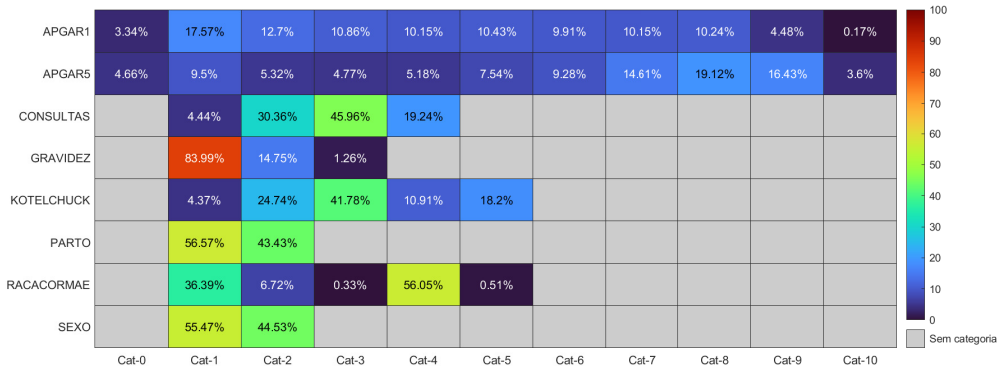

Figura 3: Estatísticas das variáveis categóricas do subconjunto de dados Grupo-Vivo-Não.

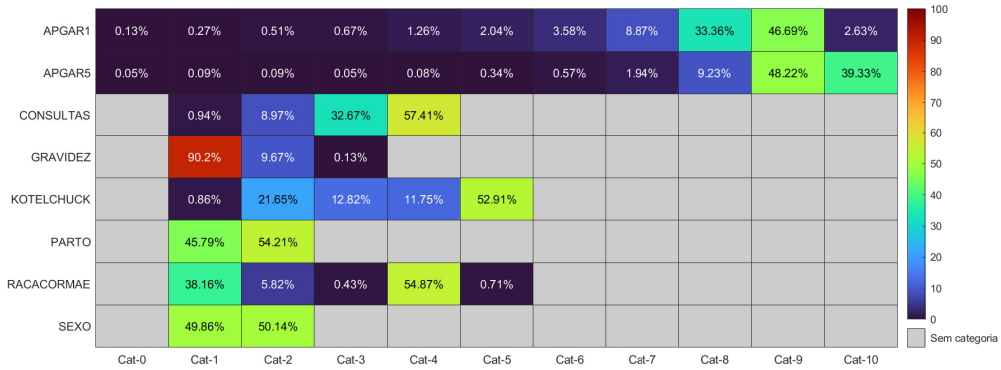

Figura 4: Estatísticas das variáveis categóricas do subconjunto de dados Grupo-Vivo-Sim.

parâmetros estatísticos (junção de um *boxplot* com um *Violin plot* [22]) das variáveis PESO, SEMAGESTAC, APGAR1, APGAR5, CONSULTAS e KOTELCHUCK em cada grupo, respectivamente.

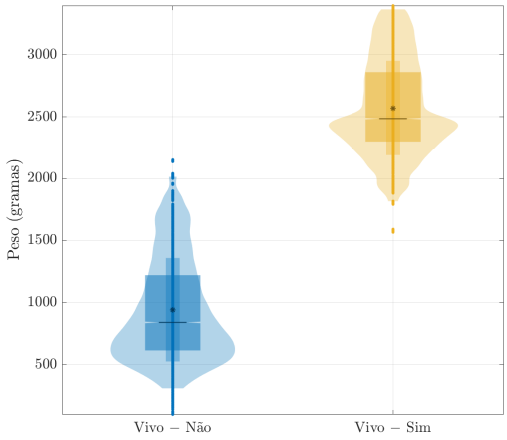

Figura 5: Distribuição da variável PESO nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim (*:média, —:mediana).

Ao comparar o Grupo-Vivo-Sim e o Grupo-Vivo-Não, observa-se uma diferença significativa em relação ao peso médio e mediano dos recém-nascidos. No Grupo-Vivo-Sim, a média de peso foi de 2.569,73 gramas, com uma mediana de 2.485 gramas, indicando um peso mais elevado em comparação ao Grupo-Vivo-Não, que teve uma média de peso de 942,13 gramas e uma mediana de 840 gramas. Essa diferença evidencia uma associação entre o peso do recém-nascido e sua sobrevivência, sugerindo que recém-nascidos com peso mais baixo possuem maior probabilidade de pertencerem ao Grupo-Vivo-Não, ou seja, não chegarem a sobreviver.

Em relação à idade gestacional média e mediana dos recém-nascidos. No Grupo-Vivo-Sim, a média de semanas de gestação foi de 35,58 semanas, com uma mediana de 36 semanas. Por outro lado, no Grupo-Vivo-Não, a média de semanas de gestação foi de 27,39 semanas, com uma mediana de 27 semanas. Essa diferença sugere que recém-nascidos com uma idade gestacional mais avançada têm uma maior probabilidade de pertencerem ao Grupo-Vivo-Sim, ou seja, de sobreviverem. A idade gestacional é um fator importante na determinação da saúde e desenvolvimento do recém-nascido, e compreender essa relação com a sobrevivência pode contribuir para a identificação de possíveis fatores de risco e intervenções

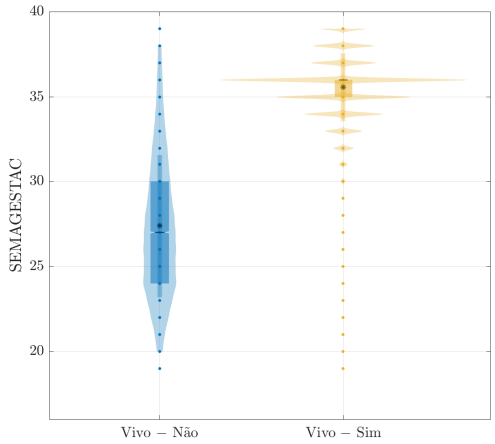

Figura 6: Distribuição da variável SEMAGESTAC nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim (*:média, —:mediana).

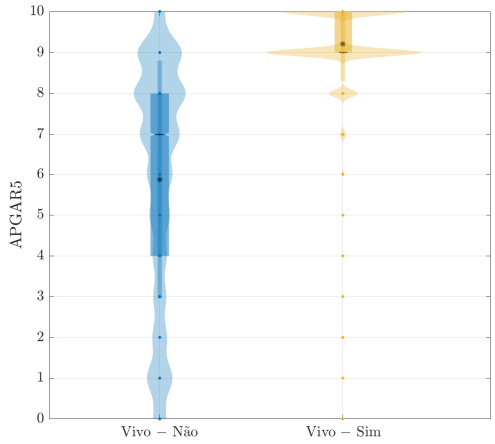

Figura 8: Distribuição da variável APGAR5 nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim (*:média, —:mediana).

adequadas.

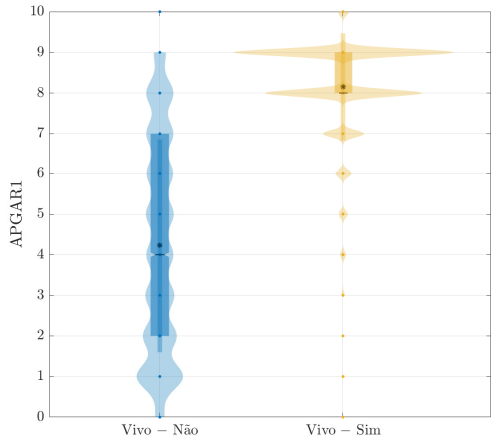

Figura 7: Distribuição da variável APGAR1 nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim (*:média, —:mediana).

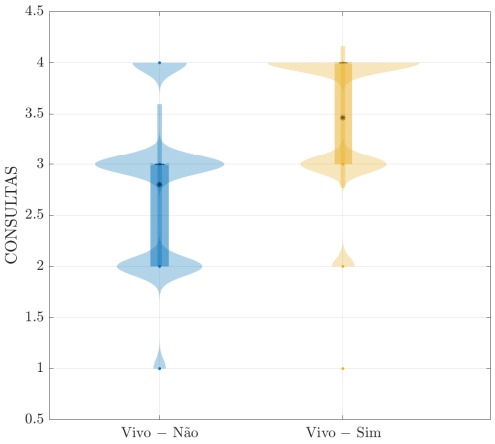

Figura 9: Distribuição da variável CONSULTAS nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim (*:média, —:mediana).

Ao comparar as variáveis APGAR1 e APGAR5 nos subconjunto Grupo-Vivo-Não e Grupo-Vivo-Sim, observa-se diferenças significativas nas distribuições. No Grupo-Vivo-Não, a média e a mediana das pontuações no APGAR1 são tem valores bastante próximos em torno de 4, enquanto no APGAR5 são de 6 e 7, respectivamente. Já no subconjunto Grupo-Vivo-Sim, a média e a mediana das pontuações no APGAR1 são semelhantes com um valor 8, enquanto no APGAR5 são também semelhantes com um valor 9, respectivamente. Esses resultados indicam que os recém-nascidos que permaneceram vivos apresentaram pontuações mais altas nos índices APGAR em comparação com aqueles que vieram a óbito.

No contexto da análise do conjunto de dados, foram investigadas as variáveis CONSULTAS e KOTELCHUCK nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim. Os resultados revelaram diferenças significativas entre os grupos em relação a essas variáveis. No subconjunto Grupo-Vivo-Não, observou-se uma média de $\approx$ 3 consultas pré-natal e uma mediana de $\approx$ 3 consultas, indicando um menor número médio de consultas em comparação ao subconjunto Grupo-Vivo-Sim, que apresentou uma média de 3,5 consultas e uma mediana de $\approx$ 4 consultas. Quanto ao índice de Kotelchuck, no subconjunto Grupo-Vivo-Não, a média foi de $\approx$ 3 e a mediana também de $\approx$ 3, enquanto no subconjunto Grupo-Vivo-Sim a média foi de $\approx$ 4 e a mediana de $\approx$ 5. Esses resultados sugerem uma diferença significativa na frequência de consultas pré-natal e no índice de Kotelchuck entre os subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim.

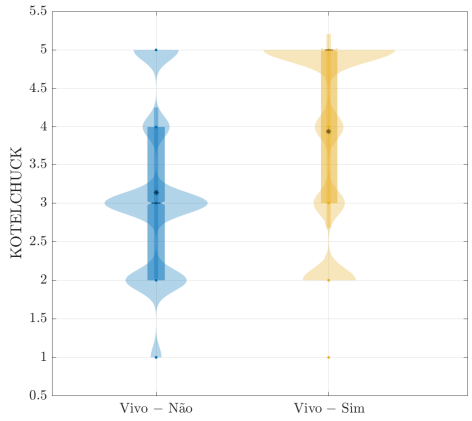

Figura 10: Distribuição da variável KOTELCHUCK nos subconjuntos Grupo-Vivo-Não e Grupo-Vivo-Sim (*:média, —:mediana).

V. Conclusão

Este estudo demonstrou que a utilização de técnicas de aprendizado de máquina não supervisionado, como o algoritmo t-SNE, é uma abordagem eficaz para estratificar o risco de mortalidade em recém-nascidos prematuros. A análise dos dados do SINASC e SIM permitiu identificar variáveis relevantes que podem influenciar o risco de mortalidade neonatal, proporcionando uma visão abrangente e detalhada da população estudada. A identificação precoce de recém-nascidos prematuros com maior vulnerabilidade ao risco de mortalidade possibilita a adoção de medidas preventivas e intervenções adequadas, direcionando recursos e esforços para melhorar os cuidados e os resultados de saúde. Essa abordagem inovadora fortalece os esforços para aprimorar a qualidade dos cuidados neonatais e contribui para a tomada de decisões informadas e eficazes no âmbito da saúde pública. Os resultados deste estudo podem ser utilizados para orientar políticas e programas de prevenção e tratamento, com o objetivo de reduzir a mortalidade neonatal e melhorar os resultados para os recém-nascidos prematuros.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

Referências

[1] B. Modell, R. Berry, C. A. Boyle, A. Christianson, M. Darlison, H. Dolk, C. P. Howson, P. Mastroiacovo, P. Mossey, and J. Rankin, "Global regional and national causes of child mortality," *The Lancet*, vol. 380, no. 9853, p. 1556, 2012.

[2] World Health Organization, *Born Too Soon: The Global Action Report on Preterm Birth*. Geneva, Switzerland: World Health Organization, 2012. [Online]. Available: https://apps.who.int/iris/handle/10665/44864

[3] A. S. Butler, R. E. Behrman *et al.*, "Preterm birth: causes, consequences, and prevention," 2007.

[4] M. L. B. Lopes, R. d. M. Barbosa, and M. A. C. Fernandes, "Unsupervised learning applied to the stratification of preterm birth risk in brazil with socioeconomic data," *International Journal of Environmental Research and Public Health*, vol. 19, no. 9, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/9/5596

[5] M. Podda, D. Bacciu, A. Micheli, R. Bellù, G. Placidi, and L. Gagliardi, "A machine learning approach to estimating preterm infants survival: development of the preterm infants survival assessment (pisa) predictor," *Scientific reports*, vol. 8, no. 1, p. 13743, 2018.

[6] R. Del Río, M. Thió, M. Bosio, J. Figueras, and M. Iriondo, "Prediction of mortality in premature neonates. an updated systematic review," *Anales de Pediatría (English Edition)*, vol. 93, no. 1, pp. 24–33, 2020.

[7] M. J. Vincer, B. A. Armson, V. M. Allen, A. C. Allen, D. A. Stinson, R. Whyte, and L. Dodds, "An algorithm for predicting neonatal mortality in threatened very preterm birth," *Journal of Obstetrics and Gynaecology Canada*, vol. 37, no. 11, pp. 958–965, 2015.

[8] J. Jaskari, J. Myllärinen, M. Leskinen, A. B. Rad, J. Hollmén, S. Andersson, and S. Särkkä, "Machine learning methods for neonatal mortality and morbidity classification," *Ieee Access*, vol. 8, pp. 123 347–123 358, 2020.

[9] A. J. Motlagh, R. Asgary, and K. Kabir, "Evaluation of clinical risk index for babies to predict mortality and morbidity in neonates admitted to neonatal intensive care unit." *Electronic Journal of General Medicine*, vol. 17, no. 5, 2020.

[10] J. Lee, J. Cai, F. Li, and Z. A. Vesoulis, "Predicting mortality risk for preterm infants using random forest," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[11] M. Son and E. S. Miller, "Predicting preterm birth: cervical length and fetal fibronectin," in *Seminars in perinatology*, vol. 41, no. 8. Elsevier, 2017, pp. 445–451.

[12] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[13] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature communications*, vol. 10, no. 1, p. 5415, 2019.

[14] E. Bajal, V. Katara, M. Bhatia, and M. Hooda, "A review of clustering algorithms: comparison of dbscan and k-mean with oversampling and t-sne," *Recent Patents on Engineering*, vol. 16, no. 2, pp. 17–31, 2022.

[15] T. Włodarczyk, S. Płotka, T. Szczepański, P. Rokita, N. Sochacki-Wójcicka, J. Wójcicki, M. Lipa, and T. Trzciński, "Machine learning methods for preterm birth prediction: A review," *Electronics*, vol. 10, no. 5, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/5/586

[16] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[17] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.

[18] M. d. C. Leal, A. P. Esteves-Pereira, M. Nakamura-Pereira, J. A. Torres, M. Theme-Filha, R. M. S. M. Domingues, M. A. B. Dias, M. E. Moreira, and S. G. Gama, "Prevalence and risk factors related to preterm birth in brazil," *Reproductive health*, vol. 13, pp. 163–174, 2016.

[19] C. Gao, S. Osmundson, D. R. V. Edwards, G. P. Jackson, B. A. Malin, and Y. Chen, "Deep learning predicts extreme preterm birth from electronic health records," *Journal of biomedical informatics*, vol. 100, p. 103334, 2019.

[20] L. C. Alves, C. E. Beluzo, N. M. Arruda, R. C. Bresan, and T. Carvalho, "Assessing the performance of machine learning models to predict neonatal mortality risk in brazil, 2000-2016," *medRxiv*, pp. 2020–05, 2020.

[21] T. A. H. Rocha, E. B. A. F. de Thomaz, D. G. de Almeida, N. C. da Silva, R. C. de Sousa Queiroz, L. Andrade, L. A. Facchini, M. L. L. Sartori, D. B. Costa, M. A. G. Campos *et al.*, "Data-driven risk stratification for preterm birth in brazil: a population-based study to develop of a machine learning risk assessment approach," *The Lancet Regional Health-Americas*, vol. 3, p. 100053, 2021.

[22] A. Legouhy, "al_goodplot - boxblot & violin plot," https://www.mathworks.com/matlabcentral/fileexchange/91790-al_goodplot-boxblot-violin-plot, June 2023, mATLAB Central File Exchange.