UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

COLEGIADO DE ESTATÍSTICA

ELIAS RIBEIRO ROSA JUNIOR

MÉTODOS DE REAMOSTRAGEM PARA DADOS DESBALANCEADOS: UMA APLICAÇÃO PARA CLASSIFICAÇÃO DE COVID-19 NA POPULAÇÃO MATERNA

VITÓRIA

2022

ELIAS RIBEIRO ROSA JUNIOR

MÉTODOS DE REAMOSTRAGEM PARA DADOS DESBALANCEADOS: UMA APLICAÇÃO PARA CLASSIFICAÇÃO DE COVID-19 NA POPULAÇÃO MATERNA

Trabalho de Conclusão de Curso apresentada ao curso de graduação em Estatística do Departamento de Estatística, do Centro de Ciências Exatas da Universidade Federal do Espírito Santo, como requisito para obtenção do grau de Bacharel em Estatística.

Orientadora: Agatha Sacramento Rodrigues

Co-orientadora: Luciana Graziela de Godoi

VITÓRIA

2022

SUMÁRIO

1 INTRODUÇÃO

de grandes volumes de dados.

O uso de algoritmos de Aprendizado de Máquina (do inglês, *Machine Learning*) tem sido cada vez mais ~~utilizados~~ devido à sua alta capacidade de lidar com armazenamento e processamento de dados. Pesquisas nas últimas duas décadas têm focado em diferentes métodos estatísticos e de aprendizado de máquina para resolver problemas de classificação e de regressão (YANG H., 2022). Entretanto, no mundo real, a modelagem de classificação~~, por sua vez,~~ pode se deparar com o problema de dados desequilibrados, isto é, quando as classes não estão aproximadamente igualmente representadas, tornando o modelo incapaz de aprender bem com a classe menor e de gerar interpretações verossímeis. ~~Em um conjunto de dados de saúde, por exemplo, em que é comum lidar com doenças raras, há uma classe menor (geralmente a doença rara) e outra consideravelmente maior, mas que devem ser igualmente consideradas.~~

O desempenho dos algoritmos de aprendizado de máquina é normalmente avaliado usando a predição da acurácia. No entanto, isso não é apropriado quando os dados são desequilibrados (CHAWLA, 2002). Chawla cita como exemplo a classificação de pixels em imagens de mamografia como possível identificador de regiões cancerígenas. ~~pontos cancerígenos~~. Para se ter uma ideia, dados típicos de mamografia podem conter 98% de pixels normais e 2% de pixels anormais. Portanto, uma predição da acurácia simples não seria apropriada para essa situação, visto que o classificador favoreceria a acurácia da classe majoritária (pixels normais), sendo que o interesse maior é a acurácia da classe minoritária (pixels anormais).

itálico

Uma das abordagens que vem se tornando popular para resolver a problemática do desbalanceamentos dos dados é sobreamostrar a classe minoritária ou subamostrar a classe majoritária (JAPKOWICZ, 2000). Existem diferentes métodos de subamostragem e sobreamostragem. As mais simples, por exemplo, apenas removem (*down-sample*) ou replicam (*up-sample*), respectivamente, as informações do conjunto de dados aleatoriamente. Entretanto, essas técnicas podem causar outro problema: o de sobreajuste (*overfitting*, em inglês)(FERNANDEZ, 2018).

Exolicar o que é esse problema na prática do machine learning, quais os impactos.

Uma alternativa mais robusta de reamostragem foi proposta por (CHAWLA, 2002), que é a de gerar dados sintéticos (ou artificiais). Dessa forma, além de superar o desequilíbrio nos conjuntos de dados originais ao gerar amostras com dados artificais (HE HAIBO, 2008), classificar as classes minoritárias e majoritárias é mais eficiente. A primeira a ganhar notoriedade foi a técnica SMOTE (CHAWLA, 2002), que posteriormente foi adaptada nas técnicas ~~Boderline~~ Borderline SMOTE (HAN, 2005), ADASYN (HE HAIBO, 2008) e DBSMOTE (BUNKHUMPORNPAT, 2012). Embora generalizações, todas, incluindo a Smote, geram dados sintéticos na classe minoritária, realizando uma sobreamostragem dos dados. Além desses métodos, existem as técnicas de subamostragem baseadas em algoritmos de aprendizado de máquina, como a Links de Tomek (TOMEK, 1976) e Near Miss (ZHANG, 2003) e a técnica ROSE (MENARDI, 2014), que realiza ao mesmo tempo subamostragem e sobreamostragem.

mostrou-se mais eficiente pois ...

SMOTE

que será o evento

Nesta monografia, os métodos de reamostragem propostos, após discutidos e avaliados, serão utilizados em um problema de classificação com dados desbalanceados reais. Nessa aplicação, serão considerados os dados de Síndrome Respiratória Aguda Grave (SRAG), disponibilizados pelo Ministério da Saúde do Brasil no portal DATASUS, que em 2020 passou a ter informações sobre o coronavírus SARS-CoV-2, o COVID-19. Exceto pela categoria “não especificado", que pode ser qualquer um dos tipos de SRAG, a base de dados proposta contém a informação da classificação do tipo de síndrome respiratória aguda grave, o evento de interesse deste estudo. O grupo de indivíduos analisado será a de mulheres gestantes e puérperas, bastante afetadas pela gravidade do novo coronavírus (TAKEMOTO, 2021). Um estudo feito pelo Observatório Obstétrico Brasileiro, aponta que durante o primeiro ano pandêmico no Brasil, um aumento considerável de mortes de mulheres no ciclo gravídico-puerperal elevou a taxa de mortalidade materna em cerca de 20% (FRANCISCO, 2021).

et. al, não?

Posto isso, este trabalho utilizará os métodos de reamostragem para o problema do desequilíbrio dos dados, uma vez que a frequência de casos de infecção pelo COVID-19 é consideravelmente maior do que a soma do casos em que o agente etiológico não era o COVID-19. Ao corrigir esse desbalanceamento, a qualidade dessas técnicas será atestada utilizando o modelo de aprendizado de máquina XGBoost. O modelo irá prever, dentre os casos de gestantes e puérperas classificadas como SRAG não especificada, quantos casos eram de fato casos de COVID-19 e quantos casos eram não COVID-19, ou seja, aqueles causados por influenza, outros vírus e outros agentes.

Fiquei um pouco perdida aqui. A gente identificou os dados desbalanceados com respeito a COVID, não-COVID. Aplicou as técnicas de reamostragem na amostra treino, fez a modelagem via XGBoost.

2 OBJETIVO

2.1 Objetivo Geral

Analisar e avaliar os métodos de reamostragem para dados desbalanceados, com uma aplicação direta nos dados relacionados à pandemia de COVID-19.

2.2 Objetivos Específicos

- ~~Identificar qual método de reamostragem melhor se aplica aos dados utilizados.~~
- ~~Estudar e discutir os métodos propostos.~~
- ~~Avaliar em que tipos de situações, a depender dos dados, o método melhor se encaixa.~~
- ~~Reunir informações de pacotes e funções atualizadas no software R, que podem ser utilizadas para cada método específico, destacando suas diferenças, vantagens e desvantagens.~~

* Apresentar diferentes metodologias de reamostragem.

* Comparar essas metodologias via estudos de simulação.

* Reunir informações de pacotes e funções atualizadas no software R, que podem ser utilizadas para cada método específico, destacando suas diferenças, vantagens e desvantagens.

* Comparar a eficiência dos métodos em termos de medidas de desempenho e de interpretabilidade na fase de teste do modelo XGBoost.

* Indicar a metodologia de reamostragem que forneceu o melhor ajuste do modelo na fase treinamento. Usar esse modelo para classificar os casos não especificados de SRAG durante a pandemia.

ver o artigo em que publicamos e reescrever essa parte, oa dados foram obtidos do SIVEP-GRIPE, não do SRAG.

3 MATERIAIS E MÉTODOS

aplicados

sem vírgula

3.1 Os dados

Os métodos de reamostragem, estudados nessa monografia, serão ~~utilizados~~ em um conjunto de dados de doenças respiratórias agudas graves de 2020 e 2021 obtido no SRAG, disponível no Open Data SUS, site do Ministério da Saúde do Brasil que disponibiliza dados de saúde para a população, https://opendatasus.saude.gov.br/.

O banco de dados de Síndrome Respiratória Aguda Grave, criada pelo Ministério da Saúde, iniciou-se na pandemia de influenza A (H1N1) em 2009, e desde então monitora os casos de influenza e outros virús respiratórios. Em 2020, a rede implantou os dados sobre a COVID-19, infecção do novo Coronavírus, que causou uma nova pandemia ao redor do mundo.

A base de dados contém a classificação final do indivíduo, que indica o diagnóstico final do caso, informando qual é o tipo de SRAG. Os tipos são influenza, outro vírus respiratório, outro agente etiológico, COVID-19 e também existe a classificação de SRAG não especificada, ~~sendo a mesma por diversos motivos, como a morte antes de um diagnóstico, escassez de testes~~ no ~~início da pandemia, entre outros~~.

Na Tabela 1 estão presentes as variáveis utilizadas no banco de dados SRAG e suas respectivas descrições.

Variável	Descrição
DT_NOTIFIC	data de notificação do caso
DT_INTERNA	data de internação do paciente
DT_SIN_PRI	data do primeiro sintoma
CS_SEXO	sexo
NU_IDADE_N	idade
CLASSI_FIN	classificação final da SRAG
CS_GESTANT	gestante ou não gestante
PUERPERA	puérpera ou não puérpera
CS_RACA	raça
CS_ESCOL_N	escolaridade
VACINA	vacina contra a gripe sim ou não
FEBRE	febre sim ou não
TOSSE	tosse sim ou não
GARGANTA	dor de gargante sim ou não
DISPNEIA	dispneia sim ou não
DESC_RESP	desconforto respiratório sim ou não
SATURACAO	saturação oxigénio sim ou não

DIARREIA	diarreia sim ou não
CARDIOPATI	cardiopata sim ou não
PNEUMOPATI	pneumopática sim ou não
RENAL	doença renal sim ou não
OBESIDADE	obesidade sim ou não
UTI	interção na uti sim ou não
SUPORT_VEN	suporte ventilatório não, não invasivo e invasivo
EVOLUCAO	evolução do caso óbito ou cura
SG_UF	estado federal
DOR_ABD	dor abdominal sim ou não
FADIGA	fadiga sim ou não
PERD_OLFT	perda de olfato sim ou não
PERD_PALA	perda do paladar sim ou não
DIABETES	diabetes sim ou não

Tabela 1 – Dicionário das variáveis do conjunto de dados.

Os dados foram baixados em 17/08/2022 no Open Data SUS citado anteriormente.

Explicar porque você precisa falar das distâncias. Foi muito direto.

3.2 Distância Euclidiana

itálico

Baseada no teorema de Pitágoras, a distância euclidiana é a distância entre dois pontos. Num espaço vetorial de n dimensões, a distância euclidiana entre um ponto $P = (p_1, p_2, \cdots, p_n)$ e $Q = (q_1, q_2, \cdots, q_n)$ é definida por:

ponto

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

3.3 Distância de Gower

algoritmo

A distância de Gower é baseada no ~~algorítimo~~ de Gower (GOWER, 1971), que pode ser usado para calcular a distância entre as variáveis de um banco de dados. O algoritmo analisa conjuntamente variáveis qualitativas e quantitativas de modo que os valores da matriz de distância estejam entre 0 e 1. A distância é definida por:

vírgula

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}$$

onde k é o numero de variáveis, p é o número total de características , i e j são duas observações quaisquer, W_{ijk} é um peso dada a comparação ijk, atribuindo valor 1 para comparações validas e valor 0 para comparações inválidas, S_{ijk} é a contribuição da variável k na similaridade entre as observações i e j, possuindo valores entre 0 e 1. Para uma variável nominal, caso o valor

válidas

itálico

Não está claro quando uma compração é válida e quando é inválida.

de k seja o mesmo para ambas as observações então S_{ijk} é 1, caso contrário, é igual a 0. Para uma variável contínua temos:

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$$

onde x_{ik} e x_{jk} são os valores da variavel k para as observações i e j, respectivamente, e R_k é a amplitude de variação da variável k na amostra.

3.4 Distância de Hamming ou Overlap

Criado em 1950 pelo engenherio R. W. Hamming, o método calcula a distância entre dois padrões considerando 0 se forem iguais, e 1 se forem diferentes da forma que:

$$D_O(p,q) = \sum_{k=1}^{n} \delta(p_k, q_k)$$

$$\delta(p_k, q_k)) = \begin{cases} 0 \text{ se } p_k = q_k \\ 1 \text{ se } p_k \neq q_k \end{cases}$$

3.5 HEOM

A distância HEOM (Heterogeneous Euclidian-overlap Metric) mescla a Distância Euclidiana com a distância ~~overlap~~, calculando as diferentes funções para seus tipos de atribuidos da forma que:

$$D_H(p,q) = \sqrt{\sum_{k=1}^{n} d_k(p_k, q_k)^2}$$

onde se o $k$ésimo for nominal $d_k(p_k, q_k)^2$ é a distância Overlap (3.4) e se for numérico é a distância Euclidiana (3.2).

3.6 KNN

O Algoritmo KNN também conhecido como método K-vizinhos mais próximos, do inglês *k-nearest ~~neighbours~~*, é um método de classificação não paramétrico, ~~que é~~ simples, mas eficaz em muitos casos (HAND, 2001). O algoritmo consiste em estimar a probabilidade condicional dos k-vizinhos mais proximos de uma determinada observação de teste x_0 (IZBICKI, 2020), de forma que,

$$P(Y_i = c | X = x_0) \quad c = \{0, 1\}$$

O KNN funciona encontrando os pontos ou vizinhos de dados mais próximos de um conjunto de dados. Os pontos de dados mais próximos são encontrados de acordo com as distâncias mais próximas do ponto de consulta, geralmente é calculada usando a distância euclidiana. Após localizar os k pontos de dados mais próximos, ele executa uma regra de votação majoritária para

descobrir qual classe apareceu mais. A classe que mais apareceu é considerada a classificação final da consulta (UDDIN, 2022).

~~O Algoritimo, de acordo com Morettin (MORETTIN, 2020) é feito da seguinte forma:~~

De acordo com Morettin (2020), o algoritmo se organiza da seguinte forma:

1. fixe um valor para k e uma observação de teste x_0.
2. identifique os k-vizinhos mais próximos de x_0 segundo algum critério de distância, e crie um conjunto W com esses pontos.
3. estime a probabilidade condicional da observação x_0 pertencer à classe C como sendo a fração dos pontos identificados em (2) que têm seus valores de Y iguais à C, isto é, $P(Y_i = c|X = x_0) = \frac{1}{k}\sum_{i\in W} I(y_i = c)$.
4. classifique x_0 com a classe que resultar em maior probabilidade.

Ora usa C, ora usa c, e não está claro o que significa, se são coisas diferentes.

É importante ressaltar que na abordagem KNN, é especialmente importante escolher o número de k corretamente, pois essa escolha pode afetar fortemente o poder preditivo do método. Pequenos valores de k levam a overfitting (alta variância), enquanto grandes valores de k resultam em modelos muito tendenciosos (AL-DOSARY, 2021).

3.7 Bootstrap suavizado

itálico

O bootstrap introduzido por Efron em 1979 é uma técnica computacionalmente intensiva que tem se mostrado útil em muitos problemas e aplicações estatísticas. Sua versão suavizada possui melhorias potenciais em relação ao bootstrap padrão (WANG, 1995). Ele consiste em gerar m novas amostras a partir de sorteios com reposição da amostra original.

Suponha que $X_1, X_2, \cdots, X_n$ seja uma amostra de distribuição contínua F desconhecida com densidade f. O interesse é estimar uma nova população de interesse com um estimador bootstrap $\alpha(\hat{F})$, onde $\hat{F}$ é o valor impírico da distribuição F_n ou sua versão suavizada.

Seja K uma função kernel simétrica tal que ela própria é uma função de densidade com variância unitária. O estimador de kernel padrão $\hat{f}_h(x)$ de $f(x)$ é dado por

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

vírgula

onde h é o parâmetro de suavização. A função de distribuição correspondente do estimador é $\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)dt$.

sem vírgula

3.8 Métodos de Reamostragem

minúsculo

Nesta seção, serão abordados os métodos de Reamostragem para lidar com a problemática do desbalanceamento dos dados. Existem três tipos de reamostragem dos dados, a sobreamostra-

gem (do inglês, *upsampling*), a subamostragem (do inglês, *downsampling*) e a junção dos dois tipos ao mesmo tempo.

Os métodos de sobreamostragem, como o próprio nome já diz, têm a finalidade de sobreamostrar o conjunto de dados, aumentando nossa amostra da classe minoritária em relação à majoritária. ~~O mais simples é mencionado na~~ seção 3.8.1, seguido dos mais robustos nas ~~seções~~ 3.8.2, 3.8.3, 3.8.4, 3.8.5 e 3.8.6.

Já os métodos de subamostragem tem a função de subamostrar o conjunto de dados, diminuindo nossa amostra da classe majoritária em relação à minoritária. O método mais simples é mencionado na ~~seção~~ 3.8.7, seguido dos métodos 3.8.8 e 3.8.9.

Por fim, mas não menos importante, a ~~seção~~ 3.8.10 aborda o método ROSE, que realiza ao mesmo tempo uma subamostragem e sobreamostragem.

3.8.1 Up-Sample

O *Up-Sample* é uma técnica de sobreamostragem aleatória. De maneira simples, ele realiza uma sobreamostragem da classe minoritária aleatória com ~~substituição,~~ replicando linhas de um conjunto de dados para igualar a ocorrência das classes.

3.8.2 SMOTE

Proposta por Chawla, em 2002, a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) realiza uma sobreamostragem gerando dados sintéticos por meio de interpolação linear. Ele utiliza o algoritmo KNN (3.6) para identificar os k-vizinhos mais próximos dos exemplos minoritários e introduz exemplos sintéticos ao longo dos segmentos que os unem.

O método seleciona aleatoriamente ~~um exemplo~~ da classe minoritária, calcula a distância euclidiana (3.2) entre ~~o exemplo e um vizinho aleatório dos seus k vizinhos mais próximos, multiplica~~ esse valor por um número aleatório entre 0 e 1. A seguir, ele aplica o resultado obtido ao longo do ponto no espaço entre o exemplo e seu vizinho, gerando assim um novo dado sintético da classe minoritária. Os passos mencionados, podem ser observados na ~~figura~~ 1 .

Como a técnica Smote utiliza a distância euclidiana em seus passos, seria errôneo aplicá-la a dados categóricos. Posto isso, CHAWLA propõe em seu estudo utilizar outras formas de se obter a associação entre as observações para aplicação em dados categorizados, criando assim a abordagem Smote-NC (3.8.3).

3.8.3 SMOTE-NC

O SMOTE-NC (Synthetic Minority Over-sampling Technique-Nominal Continuous) é uma técnica de sobreamostragem generalizada do Smote para lidar com dados ~~nominais e contínuos ao mesmo tempo.~~ categóricos e não categóricos ao mesmo tempo.

A versão mais simples do método é mencionada na Seção 3.8.1.

Seção

Seções

de forma aleatória e com reposição

Faltou citar o ano do trabalho

??? uma unidade amostral

Figura 1

esta unidade amostral e um vizinho aleatoriamente escolhido dentre os seus k-vizinhos na mesma classe minoritária, multiplicando por um número aleatório entre 0 e 1. O novo dado sintético será aquele localizado justamente no ponto do espaço entre a unidade amostral e seu vizinho.

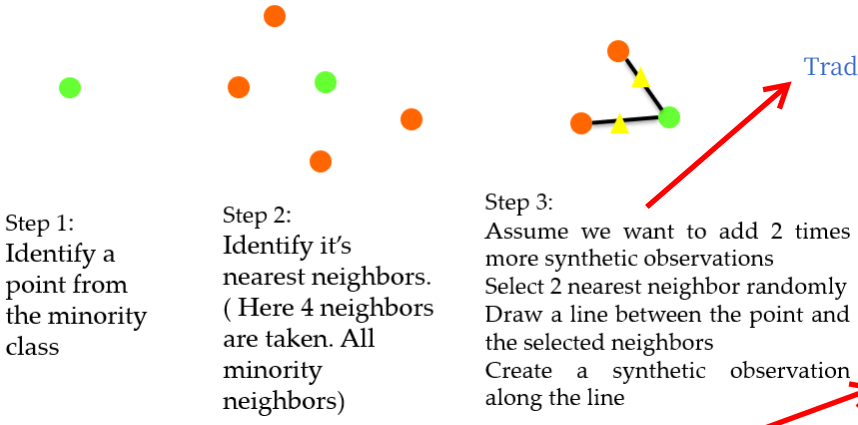

Figura 1 – Synthetic Minority Observation Generation Process
FONTE: GOSWAMI

No caso em que se tem dados categóricos e contínuos, CHAWLA propõe utilizar distância euclidiana da mediana dos desvios padrões das características contínuas para penalizar a diferença de características nominais. Portanto, ele usa apenas as informações fornecidas pela parte numérica das amostras, levando a suposições de distâncias inverossímeis. Além disso, ele ignora qualquer informação dos atributos categóricos, como a quantidade de categorias e sua distribuição.

Com a finalidade de fornecer uma sobreamostragem adequada para todos os tipos de dados, ou seja, dados categóricos puros ou numéricos e do tipo misto, destinados a problemas binários, foi criada a abordagem GSMOTE (*Generalized Synthetic Minority Oversampling TEchnique*)(CERNUDA, 2021). No que difere do Smote, o GSMOTE calcula as distâncias dos k-vizinhos mais próximos por meio de uma transformação do Coeficiente Geral de Similaridade de Gower(GOWER, 1971). Em sequência, é calculada uma medida de similaridade originalmente desenvolvida para agrupamento hierárquico de dados categóricos, Entropia variável (do inglês, *Variavel Entropy*) (SULC, 2019).

Por ser um método mais eficiente que o proposto inicialmente por CHAWLA, sua implementação foi feita aos ~~algoritimos~~ existentes do SMOTE-NC, assim, algumas funções do SMOTE-NC foram adaptadas à utilização da distância de Gower (3.3).

3.8.4 ~~Boderline~~ SMOTE

Em 2005, HAN desenvolveu a técnica ~~Boderline~~-SMOTE, subdivida em dois métodos, 1 e 2. Generalizada do método SMOTE, a técnica ~~Boderlie~~-SMOTE considera apenas as observações nas fronteiras entre as classes, reamostrando as observações da classe minoritária, que tem como k-vizinhos observações da classe minoritária e majoritária.

O método foi proposto para lidar com a má classificação nas ~~froneteiras~~ entre as classes dos outros métodos de reamostragem. Por exemplo, se houver observações na classe minoritária que ficam próximas às fronteiras com a classe majoritária, a técnica ~~Smote~~ pode criar uma ponte

entre essas classes, e gerar dados artificais entre as observações de forma errônea.

O primeiro método, Boderline-SMOTE1, classifica qualquer observação minoritária como um ponto de ruído se todos os k-vizinhos mais próximos forem da classe majoritária, e assim desconsiderado pelo algorítmo. Ele considera apenas os casos, onde o número de k-vizinhos mais próximos da classe majoritária são maiores do que os da minoritária. Desconsiderando também, os casos onde os k-vizinhos mais próximos da classe minoritaria sejam maiores do que o da classe majoritária.

Já o segundo método, Boderline-SMOTE2, não apenas considerar os casos onde o número de k-vizinhos mais próximos da classe majoritária são maiores do que os da minoritária, mas também considera os casos onde os k-vizinhos mais próximos da classe minoritária sejam maiores. A diferença, é que no segunco caso, seguido pelo calculo da distância euclidiana, a diferença dos pontos é multiplicado por um numero aleatório entre 0 e 0.5 e não 0 e 1.

Uma desvatangem desse método, é que ele acaba dando mais atenção às observações entre as fronteiras, fazendo com que o resto dos dados continue da mesma forma. Posteriormente, outras abordagens foram desenvolvidas, que por sua vez conseguiram lidar com a importância das informações nas fronteiras como também no restante dos dados, as técnicas ADASYN (3.8.5) e DBSMOTE (3.8.6).

3.8.5 ADASYN

Reescrever e excluir a palavra exemplo.

A técnica ADASYN (*Adaptive Synthetic Sampling*), desenvolvida em 2008 por HE HAIBO, tem como idéia principal utilizar uma distribuição ponderada no tipo de exemplos minoritários de acordo com sua complexidade para o aprendizado. A quantidade de dados sintéticos para cada um está associada ao nível de dificuldade de cada exemplo minoritário.

Seus objetivos incluem reduzir o viés introduzido pelo desequilíbrio de classe e aprender de forma adaptativa o limite de decisão de classificação dos exemplos difíceis nos dados. Dessa forma, mais dados sintéticos são gerados para exemplos de classes minoritárias que são mais difíceis de aprender em comparação com os exemplos minoritários que são mais fáceis.

O método aplica maior peso às observações minoritárias que tem mais k-vizinhos próximos da classe majoritária, visto que esses são mais difíceis de classificar, e menos peso as outras observações. Desta maneira, o método estabelece um critério para decidir automaticamente o número de amostras sintéticas que precisam ser geradas para cada observação minoritária.

3.8.6 DBSMOTE

Criada por BUNKHUMPORNPAT, em 2012, a técnica DBSMOTE (*Density-Based Minority Over-sampling Technique*) se baseia em uma abordagem de agrupamento baseada em densidade chamada DBSCAN (ESTER, 1996). O método realiza uma sobreamostragem gerando amostras sintéticas ao longo de um caminho mais curto de cada instância minoritária para um

~~pseudocentroide~~ de um cluster de classe minoritária.

O DBSMOTE foi inspirado pelo Borderline SMOTE no sentido de operar em uma região de fronteiras, mas ao contrário do Borderline SMOTE, o método opera em regiões seguras, aquelas regiões onde as observações minoritária tem todos os k-vizinhos das classes minoritárias, melhorando assim as taxas de aprendizado de classes minoritárias.

O método também se assemelha ao ADASYN, porém ao invés de atribuir pesos as observações da classe minoritária, a técnica as agrupam em clusters.

3.8.7 Down-Sample

O *Down-Sample* é uma técnica de subamostragem aleatória. De maneira simples, ele realiza uma subamostragem da classe minoritária aleatória ~~com substituição~~, removendo linhas de um conjunto de dados para igualar a ocorrência das classes.

3.8.8 Near-miss

A método Near-miss é uma técnica de subamostragem baseada no ~~algorítimo~~ Near-miss que observa a distribuição de classes e elimina as observações da classe majoritária que possuem a menor distância média aos k-vizinhos mais ~~proximos~~ na outra classe, minoritária.

Por exemplo, se o algoritmo encontra um caso em que duas observações próximas pertencem a classes diferentes, uma majoritária e outra minoritária, ele remove automaticamente a observação da classe superior a fim de garantir um equilíbrio das classes.

3.8.9 Links de Tomek

Em 1968, foi desenvolvida o CNN (*Condensed Nearest Neighbors*), uma ~~ténica~~ de subamostragem. A abordagem CNN é semelhante ao KNN (3.6), porém foi proposta para reduzir os requisitos computacionais de memória que o KNN exigia na época (HART, 1968).

Ivan Tomek, em 1976, propôs uma modificação do método CNN, que ao invés de escolher amostras ~~aleatóriamente~~, o método encontre pares de exemplos, uma majoritária e outra minoritária, em que exista a menor distância euclidiana entre ambas. Assim, esses pares de exemplos receberam os nomes de *Tomek Links* (TOMEK, 1976).

Aplicado às técnicas de subamostragem, o algoritmo pode ser usado para localizar todos os vizinhos mais próximos entre classes, removendo as observações da classe majoritária que tem as menores distâncias euclidianas com os vizinhos da classe minoritária, os ~~links de tomek~~.

3.8.10 ROSE

O método ROSE (*Random OverSampling Examples*), proposto por MENARDI, é uma técnica de sobreamostragem dentro de um conjunto de dados para obter regras de classificação em dados desbalanceados. Ele é estabelecido a partir da geração de novos dados artificiais

pseudocentróide
itálico
acho que não tem isso aqui
Está esquisito essa definição. Não são escolhidos aleatoriamente unidades amostrais na classe majoritária para serem excluídas? Verifique a teoria.
algoritmo
próximos
técnica
aleatoriamente
???
exemplos
???
Links de Tomek.
referenciar

itálico

das classes, de acordo com uma forma bootstrap suavizada (3.7). O algoritmo amostra uma nova instância usando a distribuição de probabilidade centrada em um exemplo selecionado aleatoriamente e dependendo de uma matriz de suavização de parâmetros de escala.

Diferente das outras técnicas de sobreamostragem e subamostragem apresentadas, ROSE combina técnicas de sobreamostragem e subamostragem gerando novos dados artificiais tanto para as classes majoritárias, quanto minoritárias.

3.9 XGBoost

Dentre os métodos de aprendizado supervisionado estatístico, existem os modelos baseados em arvores de decisão (JAMES G., 2013). Eles servem tanto para regressão tanto para classificação, ~~extratificando~~ ou segmentando o espaço de predição em varias regiões simples, comumente chamado de nós.

Vários algoritmos tem como base o modelo de árvore de decisão, dentre ...

~~Existem vários algortimos na qual são utilizados os modelos de árvores de decisão como base~~, dentre eles existem o algoritmo de aprendizagem em conjunto Boosting. O método de Boosting é um processo sequencial, onde cada árvore subsequente tenta corrigir os erros da anterior, ou seja, cada árvore $\hat{f}^i(x)$ é construída a partir dos resíduos de uma árvore $\hat{f}^{i-1}(x)$ previamente ajustada.

Seja $\hat{f}(x)$ a função de predição estimada para o modelo. Em cada iteração i, uma nova árvore $\hat{f}^i(x)$ é somada à $\hat{f}(x)$ ponderada por um parâmetro de encolhimento $\lambda < 0$ responsável por controlar a velocidade com que o algoritmo aprende. Então temos que

$$\begin{aligned}
\hat{f}^{(0)}(x) &= 0 \quad i = 0, \\
\hat{f}^{(1)}(x) &= \hat{f}^{(0)}(x) + \lambda \hat{f}^{(1)}(x) \quad i = 1, \\
\hat{f}^{(2)}(x) &= \hat{f}^{(1)}(x) + \lambda \hat{f}^{(2)}(x) \quad i = 2, \\
&\vdots \\
\hat{f}^{(n)}(x) &= \sum_{k=1}^{n} f_k(x) = \hat{f}^{(n-1)}(x) + \lambda \hat{f}^{(n)}(x) \quad i = n.
\end{aligned}$$

XGBoost

algoritmos

algoritmo XGBoost

número

O parâmetro de encolhimento λ, o ~~numero~~ e o tamanho das árvores podem ser obtidos através do método de validação cruzada (AMORIM, 2019).

Os ~~algoritimos~~ de Boosting possuem várias implementações e variações, uma das mais principais é o ~~algorítimo XGboost~~ ou Extreme Gradient Boosting. Formalmente definido como um aprendizado de máquinas por agrupamento baseado em árvores de decisão, o ~~XGboost~~ é uma aplicação do método de Gradient Boosting, que tem como objetivo minimizar uma função de custo $L(y, f(x))$ através do ~~algorítmo~~ de Gradient Descent (FRIEDMAN, 2001). Uma nova

algoritmo

árvore é ajustada utilizando o gradiente da função de custo anterior a cada iteração de forma que

$$g_i = -\left[\frac{\partial L(y_i, f(x_i)}{\partial f(x_i)}\right]_{f(x)=f^{i-1}(x)} \qquad i = 1, \cdots, n.$$

falta parêntesis aqui

algoritmos

Minimizando assim, a função perda a cada iteração.

O XGBoost funciona de maneira mais rápida e otimizada comparada aos outros ~~algorítimos~~ de Boosting pelo uso de software e combinações de hardware, utilizando os mesmos ao seu favor para ~~torna-lo~~ poderoso computacionalmente para algoritmos de árvores aprimadoradas. Em especial, ele reduz a complexidade computacional para encontrar a melhor divisão das árvores de decisão de maneira mais rápida.

torná-lo

italico

nã entendi, acho que quis dizer aprimoradas.

4 CRONOGRAMA

Os itens descritos a seguir detalham os próximos passos para a Monografia 2.

Apresentar de forma detalhada os algoritmos utilizados no processo de reamostragem.

1. ~~Descrever a idéia dos algoritmos de reamostragem, e seus respectivos passos na Metodologia do Trabalho.~~

2. ~~Acrescentar metodologias mais complexas utilizadas a alguns métodos de reamostragem, como DBSCAN no DSBMOTE (3.8.6) e Entropia variável no SMOTE-NC (3.8.3).~~

3. ~~Aplicar o método SMOTE-NC ao conjunto de dados completo (categórico e numérico).~~ de covariáveis

4. ~~Aplicar os métodos ADASYN, SMOTE e Links de Tomek com dados categóricos através do pacote "UBL", o mesmo disponibiliza trocar a forma de cálculo entre as distâncias dos dados de forma categórica, usando a distância Overlap (3.4) ou HEOM (3.5).~~

5. ~~Aplicar o modelo XGBoost nos dados completos com os métodos SMOTE-NC e com os métodos do item 4 e comparar a eficiência dos métodos em termos de medidas de desempenho e medidas de interpretabilidade. Concluindo assim qual método melhor se adequou aos dados propostos.~~

concluir

6. ~~Realizar um estudo de simulações para analisar os métodos e poder concluir quais dados específicos o melhor método se encaixa.~~

7. ~~Testar os métodos que tem mais de um pacote disponível, informando qual pacote o método tem um melhor desempenho computacional, rapidez e destacar suas diferenças.~~

1. Acrescentar metodologias mais complexas utilizadas a alguns métodos de reamostragem, como DBSCAN no DSBMOTE (3.8.6) e Entropia variável no SMOTE-NC (3.8.3). Meses:

2. Como diferentes metodologias de reamostragem são aplicadas a variáveis de mesma natureza, realizar um estudo de simulação de forma a indicar qual seria a metodologia ótima. Adicionalmente, como muitas das metodologias estão implementadas em diferentes pacotes no R, comparar o desempenho dos pacotes. Meses:

3. Aplicar os métodos ADASYN, SMOTE e Links de Tomek com dados categóricos através do pacote "UBL", o mesmo disponibiliza trocar a forma de cálculo entre as distâncias dos dados de forma categórica, usando a distância Overlap (3.4) ou HEOM (3.5). Meses:

4. Considerando todas as covariáveis do banco de dados, aplicar as diferentes metodologias de reamostragem citadas no Item 3, incluso análise com o SMOTE-NC, na fase de treinamento do modelo XGBoost. Comparar a eficiência dos métodos em termos de medidas de desempenho e de interpretabilidade na fase de teste do modelo. Meses:

5. Indicar a metodologia de reamostragem que forneceu o melhor ajuste do modelo na fase treinamento. Usar esse modelo para classificar os casos não especificados de SRAG durante a pandemia. Meses:

5 RESULTADOS PARCIAIS

Os resultados e análises descritas nesta seção foram obtidos utilizando o *software* R, versão 4.1.0, sob a IDE RStudio.

5.1 Análise exploratória dos dados

Seção

Os dados utilizados da base SRAG, descritos na ~~seção~~ 3.1, foram tratados e filtrados de forma que contivessem apenas informações adequadas de gestantes e puérperas, isto é, sexo feminino e idade entre 10 e 55 anos. A base contém 38.877 registros de sindrome respiratória aguda grave em mulheres no ~~clico gravpidico~~ puerperal, datados de janeiro de 2016 a novembro de 2021, contendo assim informações de SRAG antes da pandemia, originada em 2020 no Brasil, como depois.

ciclo

gravídico

Dos 38.877 registros, 1.521 (3,9%) foram classificados como SRAG por Influenza, 601 (1,5%) como SRAG outros vírus respiratórios, 108 (0,3%) como SRAG por outro agente etiológico, 16.408 (41,3%) como SRAG não especificada, 19.136 (49.2%) como SRAG por COVID-19 e 1.463 (3,7%) como observações em branco e ignorada. Como podemos observar, a frequência de casos de COVID-19 é a maior se comparada as outras categorias de SRAG. O tipo de SRAG não especificada é a segunda maior frequência dos dados, ou seja, os casos de COVID-19 podem ter sido subestimados.

Com a finalidade de reamostrar as informações de COVID-19 e outras síndromes respiratórias agudas graves, primeiro fez-se a junção de SRAG por influenza, outros vírus respiratórios e outros agentes etiológicos em uma categoria, com o nome de não COVID-19. Selecionando apenas os casos de COVID-19 e não COVID-19, a base de dados, por fim, registra 21.366 observações, sendo 19.136 (89.6%) SRAG por COVID-19 e 2.230 (10.4%) por não COVID-19.

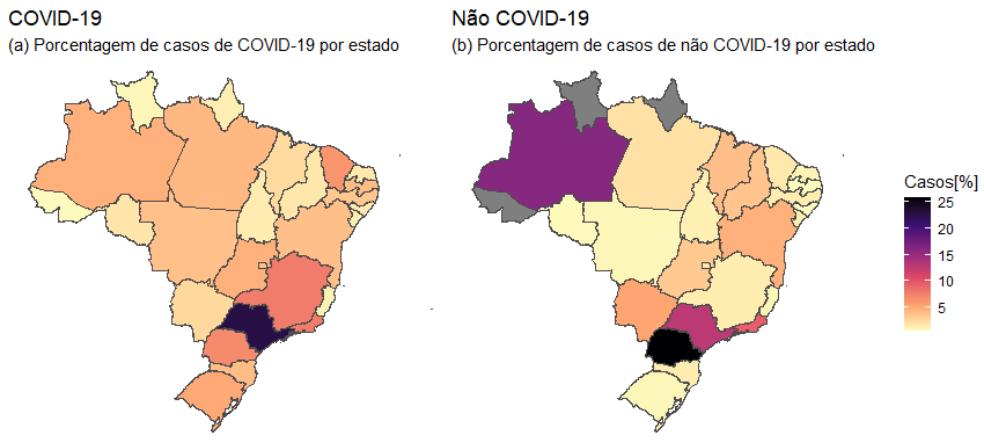

Figura 2 – Mapas dos estados para os casos de SRAG por COVID-19 e não COVID-19

Na Figura 2 podemos observar a distribuição dos casos de COVID-19 e não COVID-19 por estado no Brasil. No mapa 2(b), os estados em cinza mostram que não existiu nenhum registro de outras síndromes respiratórias agudas grave em gestantes e puérperas. Podemos notar que o estado com maior porcentagem de número de casos é o estado Paraná com 25,76% dos casos, seguido pelo Amazonas com 16,03% e São Paulo com 12,79%. Para os casos de COVID-19, é observado que o estado de São Paulo detém a maior porcentagem de casos, com aproximadamente 22,81% dos casos no Brasil, o que é esperado devido a sua grande população. Observamos também que a região sudeste representa a região com estados com maior porcentagem de casos de COVID-19.

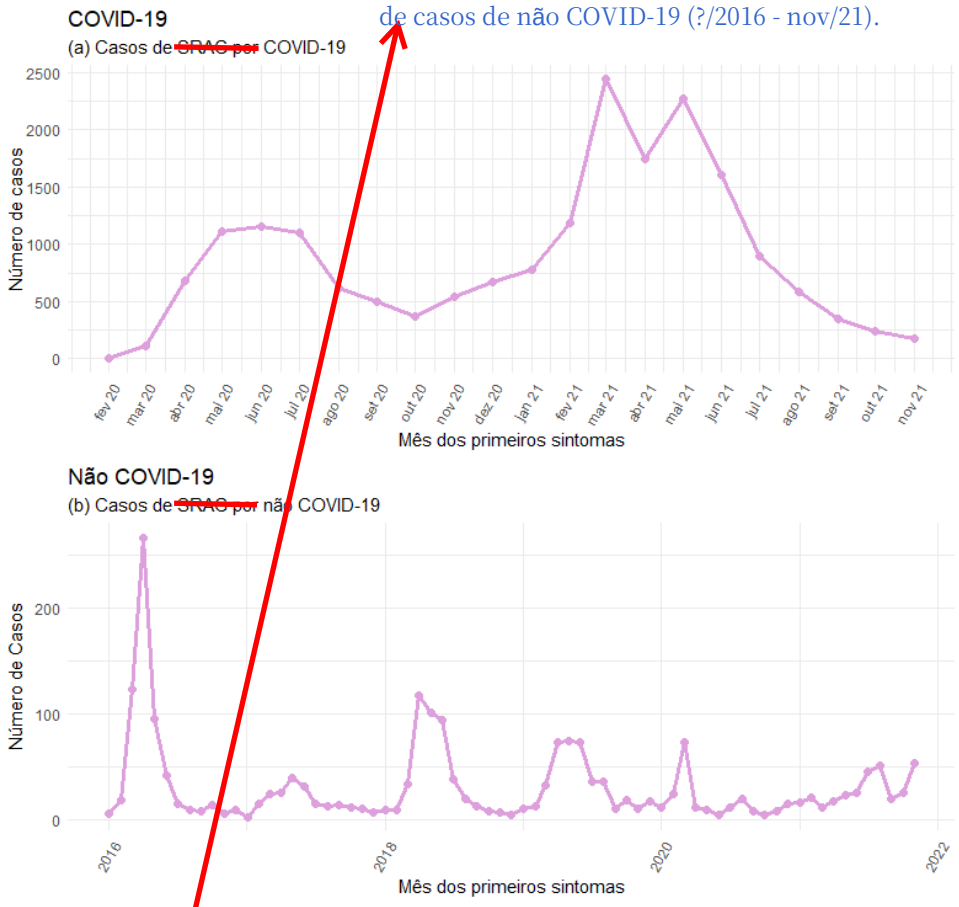

Figura 3 – Gráficos do número de casos de SRAG por COVID-19 e não COVID-19 distribuídos pelo tempo

O número de casos de COVID-19 e não COVID-19 se encontram distribuídos no tempo na Figura 3, em que a data de primeiros sintomas é a data de referência. Podemos observar para os casos de COVID-19, na Figura 3(a), que a quantidade casos tiveram dois aumentos consideráveis, um no início da pandemia, em torno de março para abril de 2020, e outro ainda maior em fevereiro de 2021, alcançando 2.442 casos de COVID-19 em gestantes e puérperas. Ao observar a Figura 3(b), podemos ver a distribuição do número de casos de não COVID-19 desde 2016 até novembro de 2021, podemos notar inclusive um aumento de casos sempre em meados de fevereiro/março/abril. Nota-se também que a frequência de casos de 2020 e 2021 foram

muito abaixo do esperado, levando a presumir que dentro dos casos de SRAG não especificada, também exista bastante casos de SRAG por influenza, outros vírus respiratórios e outros agentes etiológicos.

No que se referem as variáveis numéricas, temos a variável idade e as variáveis de datas, bens como, data dos primeiros sintomas, data da notificação do caso e data da internação. Sequencialmente, foi criada uma variável que conta os dias entre a data dos primeiros sintomas até a data de notificação. Posto isso, foram feitas descritivas das variáveis idade e tempo de sintomas até notificação, presentes respectivamente nas Tabelas 2 e 3.

Classificação	freq	media	DP	mediana	q25	q75
COVID-19	19136	29,90	7,23	30,00	25,00	35,00
não COVID-19	2230	27,44	7,05	27,00	22,00	32,00

Tabela 2 – Medidas descritivas da variável de idade nos grupos COVID-19 e não COVID-19, sendo elas a frequência (freq), média (media), desvio padrão (DP), mediana (mediana), quartil 25 (q25) e quartil 75 (q75).

Classificação	freq	media	DP	mediana	q25	q75
COVID-19	19136	10,26	21,60	7,00	4,00	11,00
não COVID-19	2230	7,32	20,33	3,00	2,00	6,00

Tabela 3 – Medidas descritivas da variável de tempo dos sintomas até a notificação nos grupos COVID-19 e não COVID-19, sendo elas a frequência (freq), média (media), desvio padrão (DP), mediana (mediana), quartil 25 (q25) e quartil 75 (q75).

5.2 Reamostragem dos dados com variáveis contínuas

Os métodos explorados nessa seção, foram definidos na seção 3.8. Como a maioria dos métodos propostos são métodos baseados em dados contínuos, para exemplificação vamos utilizar duas variáveis numéricas do banco de dados, a idade e o tempo dos primeiros sintomas até a notificação.

Podemos observar na Figura 4 a distribuição dos dados após a aplicação dos métodos Down-Sample (3.8.7) e Up-sample (3.8.1). Como comentado anteriormente, podemos observar na Figura 4(b) que o método Down-Sample removeu observações aleatórias da classe "covid-19", equiparando os dados, ao contrário do método Up-Sample, Figura 4(c), em que ele replicou os registros alcatóriamente. Ambos os métodos foram aplicados utilizando o pacote *themis*.

Aplicando o método SMOTE, podemos visualizar a sobreamostragem que ele realizou na Figura 5, reamostrando a classe minoritária não COVID-19 a fim de equilibrar os dados. O pacote utilizado para aplicação do método SMOTE foi o pacote *themis*, mas existem os pacotes *smotefamily* e *UBL* que também realizam o método SMOTE.

Os métodos Boderline-SMOTE1 e Boderline-SMOTE2 são generalizações do método SMOTE, definidos na seção 3.8.4, porém eles realizam uma sobreamostragem apenas nas

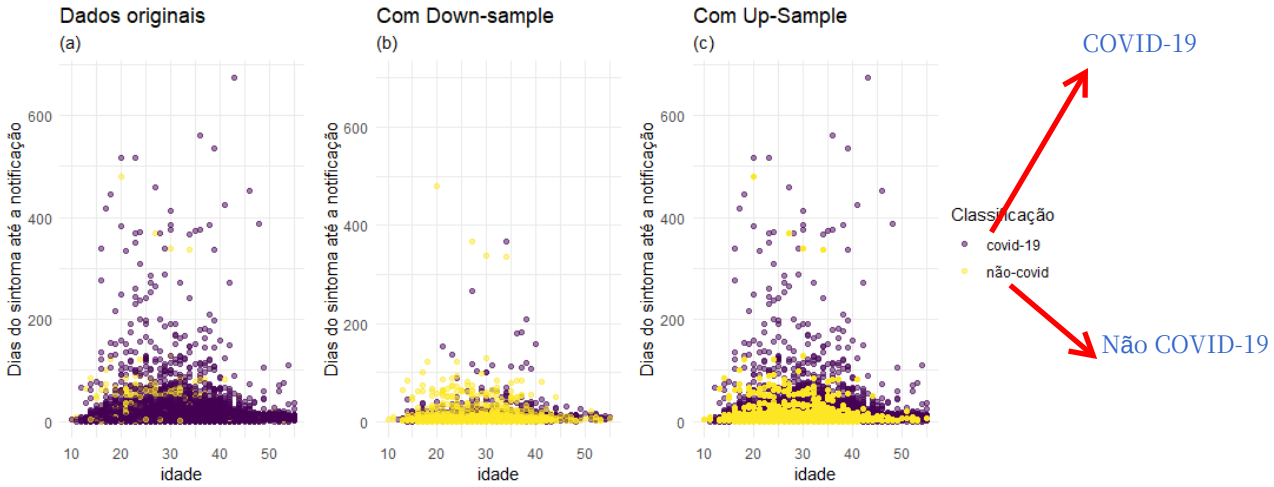

Figura 4 – Gráficos da distribuição dos dados com os métodos de reamostragem Down-Sample e UP-Sample para cada classificação.

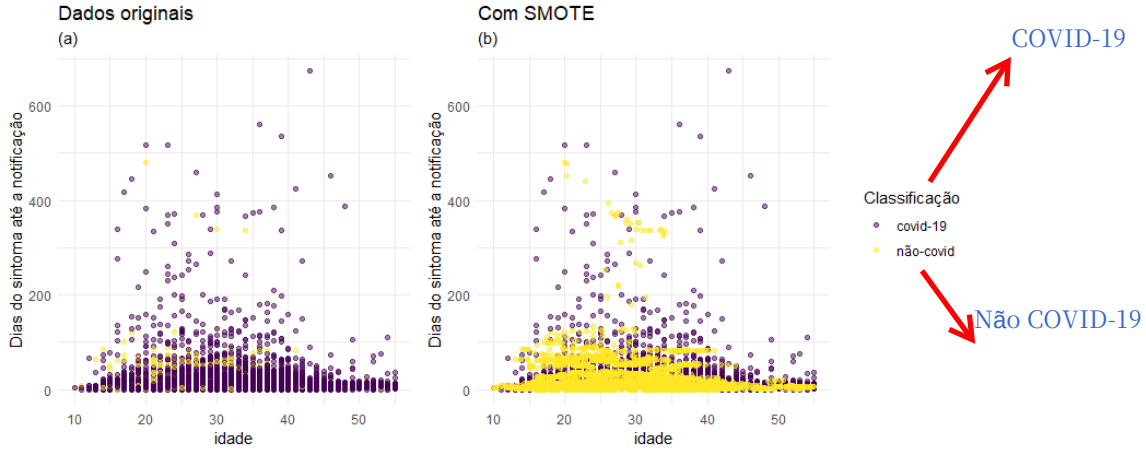

Figura 5 – Gráficos da distribuição dos dados com o método de reamostragem SMOTE para cada classificação.

fronteiras entre as classes de interesse. A Figura 6 mostra a distribuição dos dados após a aplicação do método, podemos observar claramente que o método só reamostrou realmente as observações da classe minoritária não COVID-19 nos lugares onde se concentram as maiores fronteiras entre as classes COVID-19 e não COVID-19. A diferença entre os dois métodos não parece ter muita relevância nos dados, sendo que o primeiro ~~amostra~~ uma parte ~~específica~~ e o segundo abrange uma fronteira maior, ja que considera todos os k-vizinhos mais próximos em que a classe majoritária é maior ou menor porém com menos peso, como citados anteriormente na ~~seção~~ 3.8.4. O pacote utilizado para aplicação dos métodos foi o pacote *themis*, mas existe o pacote *smotefamily* que também realizam os métodos.

método reamostra

específica

Seção

itálico

Como discutido anteriormente, o método ~~Boderline~~-SMOTE tem sua desvantagem por dar mais atenção às observações entre as fronteiras, fazendo com que o resto dos dados permaneçam os mesmos, podendo causar até mesmo um viés nos dados no espaço onde o método sintetiza mais registros. Pensando em resolver esse problema, foram desenvolvidos os métodos

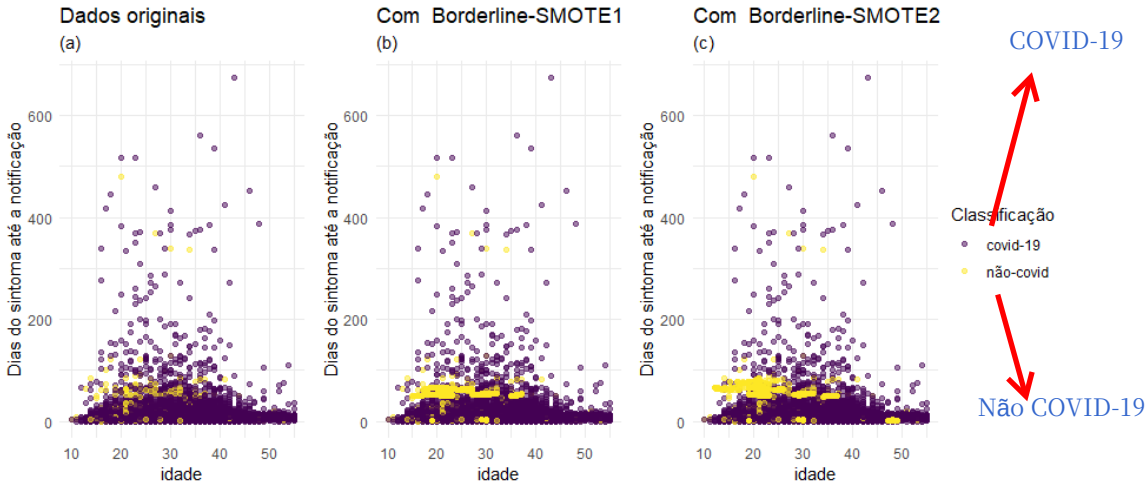

Figura 6 – Gráficos da distribuição dos dados com os métodos de reamostragem ~~Boderline~~-SMOTE1 e ~~Boderline-Smote~~2 para cada classificação.

SMOTE

me parece muito similar ao original. Verificar os n em cada grupo.

ADASYN (3.8.5) e DBSMOTE (3.8.6). As aplicações desses métodos a base de dados pode ser visualizado na Figura 7, podemos notar que o ADASYN se assemelha bastante ao SMOTE, porém agora dando atenção também as fronteiras. Observamos que o método DBSMOTE não foi similar ao ADASYN e nem ao SMOTE, realizando um equilíbrio dos dados mais suave. O pacote utilizado para aplicação dos método ADASYN foi o pacote *themis*, mas existe o pacote *smotefamily* que também realizam os métodos. Para aplicação do método DBSMOTE foi utilizado o pacote *smotefamily*.

Informação repetitiva, talvez colocar no final do primeiro parágrafo da Seção 5.2.

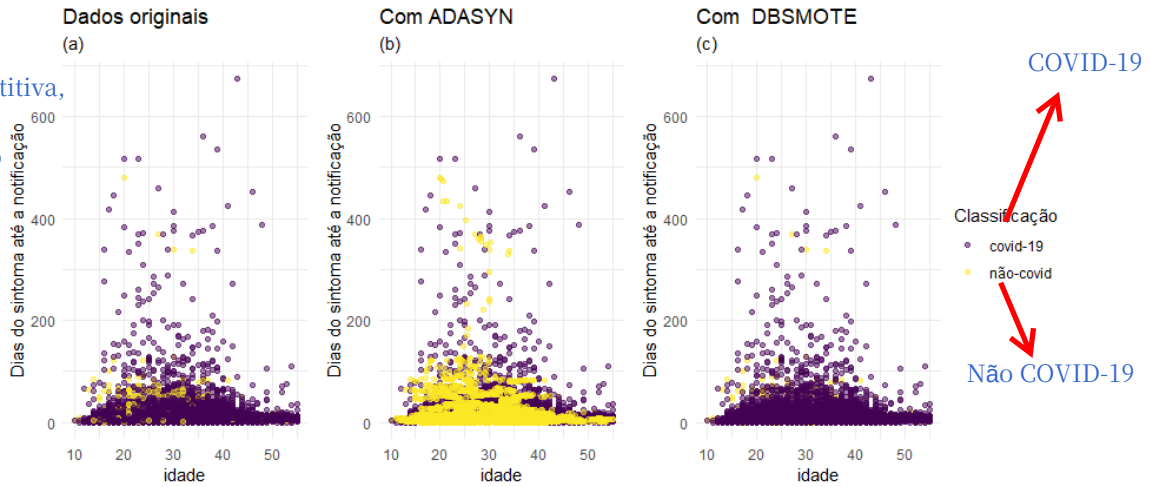

Figura 7 – Gráficos da distribuição dos dados com os métodos de reamostragem ADASYN e DBSMOTE para cada classificação.

No que se referem aos métodos de subamostragem Near-miss e Links de Tomek, podemos observar a distribuição dos dados nas Figura 8, podemos notar que o método Near-miss realizou a subamostragem equilibrando totalmente os dados. Já o links de tomek, por sua definição, ele só remove os registros majoritários próximos as fronteiras com a classe minoritária, similar a idéia do ~~Boderline~~-SMOTE, mas de forma a realizar uma subamostragem, removendo assim poucos registros.

Citar quantos n perdeu e seu percentual em relação ao grupo.

Links de Tomek

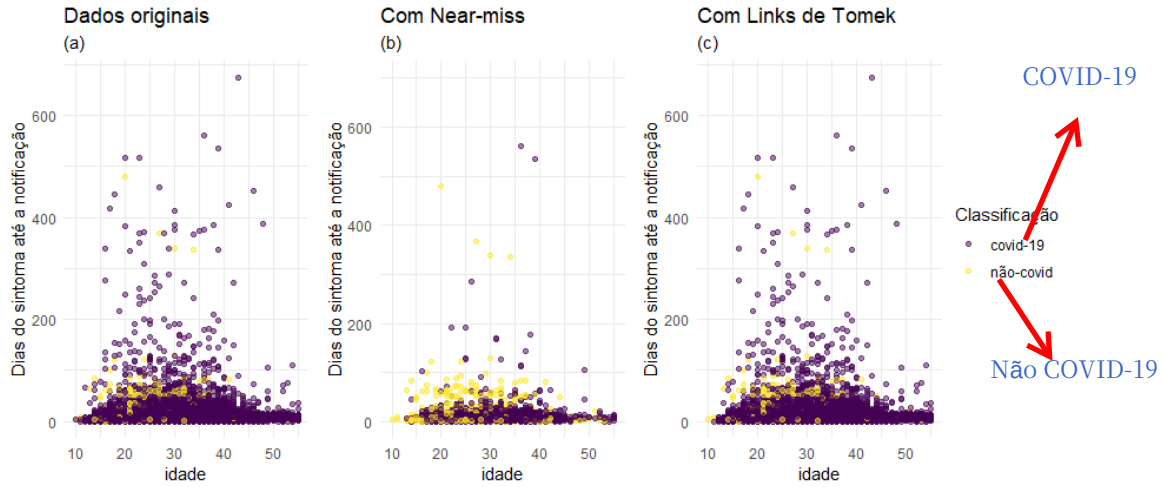

Figura 8 – Gráficos da distribuição dos dados com os métodos de reamostragem Near-miss e Links de Tomek para cada classificação.

itálico

Por fim, o método ROSE foi aplicado aos dados, e sua distribuição pode ser visualizada na Figura 9. Podemos observar que o método realizou uma subamostragem da classe majoritária e uma sobreamostragem da classe minoritária, reamostrando todas os registros através do método bootstrap ~~suaviado~~ (3.7). Apresentar os valores amostrais ao final.

suavizado

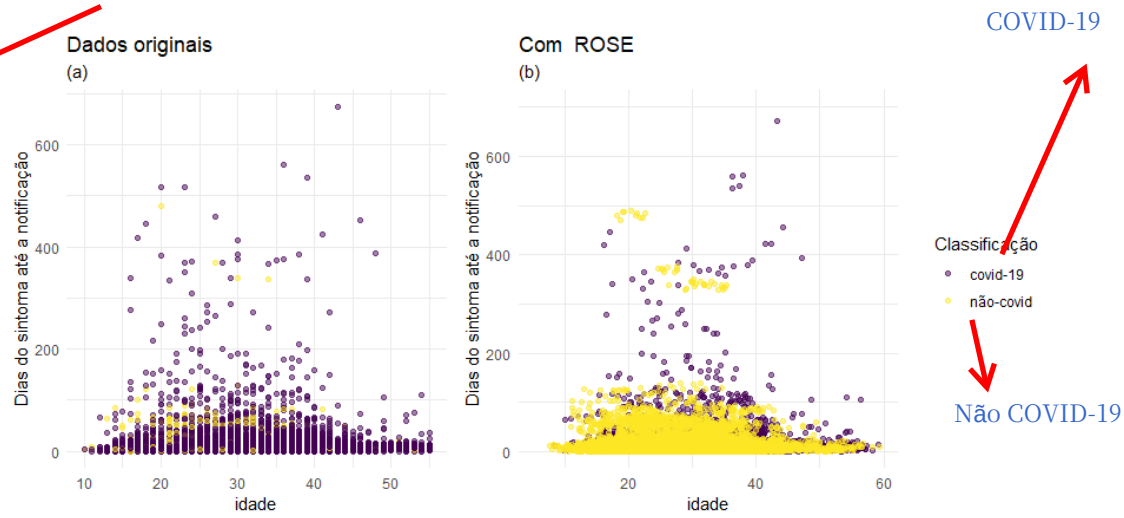

Figura 9 – Gráficos da distribuição dos dados com o métodos de reamostragem ROSE para cada classificação.

Os próximos passos deste estudo se encontram no Capítulo 4.

REFERÊNCIAS

AL-DOSARY, N. M. N. A.-H. S. A. A. A. M. K-nearest neighbors method for prediction of fuel consumption in tractor-chisel plow systems. Associação Brasileira de Engenharia Agrícola, 2021. Citado na página 10.

AMORIM, W. N. d. **Ciência de dados, poluição do ar e saúde**. [S.l.]: PhD thesis, Universidade de São Paulo, 2019. Citado na página 15.

BUNKHUMPORNPAT, C. S. K. L. C. **DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique**. 36. ed. [S.l.]: Applied Intelligence, 2012. 664–684 p. Citado 2 vezes nas páginas 4 e 13.

CERNUDA, C. R.-B. D. A.-A. I. M. G. I. Z. U. **Generalized SMOTE: A universal generation oversampling technique for all data types in imbalanced learning**. [S.l.: s.n.], 2021. Citado na página 12.

CHAWLA, V. N. B. W. K. H. O. L. K. P. W. **SMOTE: Synthetic Minority Over-sampling Technique**. [S.l.]: Journal of Artificial Intelligence Research 16, 2002. Citado 3 vezes nas páginas 4, 11 e 12.

ESTER, M. K. H.-P. S. J. X. X. **A density-based algorithm for discovering clusters in large spatial databases with noise**. [S.l.]: The 2nd international conference on knowledge discovery and data mining, 1996. 226-231 p. Citado na página 13.

FERNANDEZ, A. G. S. G. M. P. R. C. K. B. H. F. **Learning from Imbalanced Data Sets**. [S.l.]: Springer, pg.83, 2018. v. 1. Citado na página 4.

FRANCISCO, R. P. V. L. L. R. A. S. **Obstetric Observatory BRAZIL - COVID-19: 1031 maternal deaths because of COVID-19 and the unequal access to health care services**. [S.l.]: Clinics (São Paulo), 2021. Citado na página 5.

FRIEDMAN, J. H. **Greedy function approximation: a gradient boosting machine**. [S.l.]: Annals of statistics, 2001. Citado na página 15.

GOSWAMI, S. **Class Imbalance, SMOTE, ~~borderline~~ Borderline SMOTE, ADASYN**. [s.n.], 2020. Disponível em: <https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804>. Citado na página 12.

GOWER, J. C. **A general coefficient of similarity and some of its properties**. [S.l.]: Biometrics, 1971. Citado 2 vezes nas páginas 8 e 12.

HAN, H. W.-Y. W. B.-H. M. **Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning**. [S.l.]: Advances in intelligent computing, 2005. 878-887 p. Citado 2 vezes nas páginas 4 e 12.

HAND, D. H. M. P. S. **Principles of Data Mining. The MIT Press**. [S.l.: s.n.], 2001. Citado na página 9.

HART, P. **The condensed nearest neighbor rule**. 4. ed. [S.l.]: IEEE Transactions on Information Theory, 1968. 515-516 p. Citado na página 14.

HE HAIBO, B. Y. **ADASYN: Adaptive synthetic sampling approach for imbalanced learning**. [S.l.]: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008. Citado 2 vezes nas páginas 4 e 13.

IZBICKI, R. S. T. M. d. Aprendizado de maquina: uma abordagem estatística. Câmara Brasileira do Livro, SP, Brasil, 2020. Citado na página 9.

JAMES G., W. D. . H.-T. . T. R. **An Introduction to Statistical Learning with aplications in R**. [S.l.: s.n.], 2013. Citado na página 15.

JAPKOWICZ, N. **The Class Imbalance Problem: Significance and Strategies**. [S.l.]: Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning, 2000. Citado na página 4.

MENARDI, G. T. N. **Training and assessing classification rules with imbalanced data.** [S.l.]: Data Mining and Knowledge Discovery, 28(1):92–122,, 2014. Citado 2 vezes nas páginas 4 e 14.

MORETTIN, P. A. S. J. M. **Introdução a Ciência de Dados - Fundamentos e Aplicações .** [S.l.: s.n.], 2020. Citado na página 10.

SULC, M. Z. R. H. **Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering**. [S.l.: s.n.], 2019. Citado na página 12.

TAKEMOTO, M. M. M. d. O. d. O. A. C. e. a. **Higher case fatality rate among obstetric patients with COVID-19 in the second year of pandemic in Brazil: do new genetic variants play a role?** [S.l.]: medRxiv, 2021. Citado na página 5.

TOMEK, I. **Two Modifications of CNN**. 6. ed. [S.l.]: Systems, Man and Cybernetics, IEEE Transactions on, 1976. 769-772 p. Citado 2 vezes nas páginas 4 e 14.

UDDIN, S. H. I. H. L. H. M. M. A. G. E. Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction. Scientific Reports , 6256, 2022. Citado na página 10.

WANG, S. **OPTIMIZING THE SMOOTHED BOOTSTRAP**. [S.l.]: Inst. Statist. Math., 1995. Citado na página 10.

YANG H., L. M. **Software Defect Prediction Based on SMOTE-tomek and XGBoost**. [S.l.]: Bio-Inspired Computing: Theories and Applications, 2022. Citado na página 4.

ZHANG, J. M. I. **KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction**. [S.l.]: Proceeding of International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets, 2003. Citado na página 4.