Ministério da Educação
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística

UTILIZAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DO COMPORTAMENTO SEDENTÁRIO DE EGRESSOS UNIVERSITÁRIOS EM BASE DE DADOS DESBALANCEADAS

Pedro Augusto Alves Viana Cotta

Ouro Preto MG
2021

Pedro Augusto Alves Viana Cotta

UTILIZAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL PARA CLASSIFICAÇÃO DO COMPORTAMENTO SEDENTÁRIO DE EGRESSOS UNIVERSITÁRIOS EM BASE DE DADOS DESBALANCEADAS

Monografia apresentada ao Curso de Estatística da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau em Bacharelado em Estatística.

Orientador: Prof. Dr. Marcelo Carlos Ribeiro

Ouro Preto - MG
21 de maio de 2021

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C846u Cotta, Pedro Augusto Alves Viana.
Utilização de técnicas de inteligência artificial para classificação do comportamento sedentário de egressos universitários em base de dados desbalanceadas. [manuscrito] / Pedro Augusto Alves Viana Cotta. - 2021.
50 f.

Orientador: Prof. Dr. Marcelo Carlos Ribeiro.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Estatística .

1. Inteligência artificial. 2. Algoritmos. 3. Big data. 4. Aprendizagem de máquina. I. Ribeiro, Marcelo Carlos. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.8

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Pedro Augusto Alves Viana Cotta

Utilização de técnicas de inteligência artificial para classificação do comportamento sedentário de egressos universitários em base de dados desbalanceados

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal

de Ouro Preto como requisito parcial para obtenção do título de Estatístico

Aprovada em 21 de maio de 2021

Membros da banca

Dr. Marcelo Carlos Ribeiro - Orientador- Universidade Federal de Ouro Preto

Dr. Tiago Martins Pereira - Universidade Federal de Ouro Preto

Dr. Fernando Luiz Pereira de Oliveira - Universidade Federal de Ouro Preto

Documento assinado eletronicamente por **Marcelo Carlos Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 26/05/2021, às 16:47, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Fernando Luiz Pereira de Oliveira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 27/05/2021, às 22:06, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0175702** e o código CRC **8362FD15**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

Este trabalho é dedicado a minha família que tanto amo, amigos que sempre acreditaram no meu sucesso e aos professores que me apoiaram.

Agradecimentos

Agradeço a primeiramente a Deus pela saúde e força para nunca desistir.

Agradeço minha mãe Luciana Alves Viana por me apoiar na dedicação aos estudos. Jamais poderei agradecer suficiente por tudo que você fez e faz por mim, te dedico todas as minhas conquistas.

Agradeço todos meus familiares que me apoiaram nessa caminhada, mas em especial a minha irmã Isabella que mesmo com nossas divergências sempre me apoiou e ajudou da melhor maneira possível. Obrigado, João Batista, que ajuda muito em tudo que pode, e ao meu avô, que sempre me apoiou em minhas decisões.

Agradeço aos meus vizinhos, mas em especial ao Claudio e Marilia que me trataram como um filho com muitos conselhos e carinho, também à Irce, com quem tive o prazer de ter aulas, conhecendo o excelente lado profissional dela.

Agradeço aos amigos do futebol que tiveram que aturar meu péssimo temperamento nos jogos, momentos nos quais eu descarregava toda a tensão do dia-a-dia. Por isso peço desculpas. Vocês foram extremamente importantes para minha saúde mental.

Agradeço aos meus amigos de uma vida inteira Phillipe, Claudinho, Virgínio, Eduardo, João Marcelo, William e Rodrigo, que sempre foram como irmãos mais velhos e muito me ensinaram, protegeram-me, preparando-me para os perigos das ruas.

Agradeço aos amigos que pesar do distanciamento geografico criado com o rumo das nossas vidas jamais deixaram de manter contato, Guilherme Andrade, Guilherme Morais, Edilson, Lucas, Felipe, Thiago Augusto, Pedro Henrique e Matheus Gramigna.

Agradeço aos amigos e seus familiares que me aceitaram com todas as minhas loucuras, Amanda Nascimento, Tácila Nascimento, Lázaro Dias, Daianna Lana, Dauberson Mól e Yasminzinha que chegou para somar os nossos motivos para sorrir.

Agradeço aos amigos de Instituto Federal (IF), Thais, Sarah, Thatyanna, Lourrane, Ingrid, Elimara, Viviane, Leidiane e Marcos, que mantiveram conato e amizade após nossa jornada de muitas diversidades.

Agradeço aos amigos e colegas de UFOP que tive o prazer de conhecer, sejam do meu período ou não, Luiz Zanetti, Bruno Oliveira, Gleizer Garrido, Ana, Juliana, Carol, Leticia Marrara, Nathalia, Philipe, Matheus, Debora, Bruno, Victor, Ismael, Jefferson, Joziani, Andressa, Raphaella, Diana, Isadora, Felipe, Leticia Gauna, Yuri, Iara, Milele, Daiane, Brenda, Thiago e Henrique.

Agradeço aos amigos que tive a honra de conhecer em trabalhos ao longo do período

acadêmico na Fundação Gorceix e na Prefeitura de Ouro Preto, essas experiências muito agregaram à minha vida. Em especial gostaria de agradecer ao meu líder e responsável na prefeitura, Samuel Sabino, que sempre me deu uma força no meu período de estágio.

Agradeço a UFOP pelo ensino de qualidade, estrutura e suporte aos alunos e aos excelentes técnicos e professores do departamento de estatística (DEEST), Juliano, Luiz, Diana, Erica, Graziela, Eduardo, Ricardo, Flavio, Spencer, Anderson, Rivert e Ivair. Em especial gostaria de agradecer três professores que me direcionaram para uma área desejada e rica em aprendizados, Marcelo, Fernando e Tiago, sem vocês acho que ainda estaria perdido no curso, muito obrigado.

Finalmente gostaria de agradecer os amigos mais presentes nessa luta, mostrando-me que também temos direito aos dias de glória, Denis, Gabriella, Leilayne, Willian, Marcos, Gulherme (7x1) e Thiago (Pombo). Unidos, sempre vencemos todas as barreiras, e em cada vitória comemoramos da melhor maneira possível, Valdeir, Luiz, Saulo, Breno e Danilo, agradeço por não deixarem a rotina de ir na Komb falhar. Agradeço meu amigo Wellington Ferreira de Souza, não tenho palavras para agradecer-lhe, ele me ajudou em trabalhos, temas e disciplinas, e esteve presente nas comemorações, muito obrigado irmão.

Vocês e outros mais foram essenciais nessa conquista, peço desculpa por não conseguir escrever todos que participaram, mas agradeço muito e de coração.

Você tem que encontrar o que você gosta. E isso é verdade tanto para o seu trabalho quanto para seus companheiros. Seu trabalho vai ocupar uma grande parte da sua vida, e a única maneira de estar verdadeiramente satisfeito é fazendo aquilo que você acredita ser um ótimo trabalho. E a única maneira de fazer um ótimo trabalho é fazendo o que você ama fazer. Se você ainda não encontrou, continue procurando. Não se contente. Assim como com as coisas do coração, você saberá quando encontrar. E, como qualquer ótimo relacionamento, fica melhor e melhor com o passar dos anos. Então continue procurando e você vai encontrar. Não se contente.

Steve Jobs

Resumo

Os recentes avanços nas áreas científicas e tecnológicas possibilitaram o crescimento e armazenamento de grandes volumes de dados. Com a finalidade de se extrair informações, vem surgindo diversas formas de análises, sendo aprimoradas através de ferramentas computacionais apropriadas. O Aprendizado de Máquina vem sendo muito utilizado como ferramenta para análise, mas percebe-se a necessidade de se trabalhar nos dados antes da entrada no algoritmo, pois os algoritmos possuem limitações que podem ser prejudiciais aos resultados, gerando uma predição incorreta, a título de exemplo, dados com classes desbalanceadas podem criar viés para uma determinada classe. Para solucionar esse tipo de problema muitos pesquisadores tem apresentado propostas aos quais nos baseamos e definimos o objetivo desse trabalho.

Palavras-chave: Classificação, Dados desbalanceados, Algoritmo.

Abstract

Recent advances in science and technology have made it possible growing and storing big data. In order to extract information of it, several forms of analysis are emerging and are improved, through appropriate computational tools. Machine learning has been widely used as a tool for analysis, but it is necessary to work the data even before entering'em the algorithm. The algorithms have limitations that could be detrimental to the results, as an example, unbalanced classes could create bias for a especific class. Trying to solve this type of problem, many researchers have presented some answers, witch define the objective of this work.

Keywords: Classification, Unbalanced data, Algorithm.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AUC	Area Under the Curve
CAAE	Certificado de Apresentação de Apreciação Ética
CUME	Coorte de Universidades Mineiras
DEEST	Departamento de Estatística da Universidade Federal de Ouro Preto
IFES	Instituições Federias de Ensino Superior
IMC	Índice de massa corpora
KNN	Knowledge Discovery in Databases
MAP	Maximum A Posteriori
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Oversampling Technique
SUN	Seguimiento Universidad de Navarra
UFJF	Universidade Federal de Juiz de Fora
UFLA	Universidade Federal de Lavras
UFMG	Instituto Federal do Espirito Santo
UFOP	Universidade Federal de Ouro Preto
UFV	Universidade Federal de Viçosa

Lista de símbolos

β	Letra grega minúscula beta
ϵ	Letra grega épsilon
α	Letra grega alfa
$\prod$	Letra grega maiúscula pi
$\sum$	Letra grega maiúscula sigma
$\ln$	Logaritmo natural
π	Letra grega minúscula pi

Sumário

1 Introdução

Os recentes avanços da ciência e tecnologia viabilizaram o crescimento e disponibilidade de grandes volumes de dados. Devido a essa grande massa de dados, faz-se necessário um aumento de diferentes técnicas de analise na extração de informações uteis na tomada de decisão. Logo, áreas de informação, tecnologia e análise vêm ganhando e cristalizando seu espaço. Os campos da Inteligência Artificial (AI) e Aprendizado de Máquina (AM) se dedicam a essa tarefa, desenvolvendo técnicas que permitem que a máquina reconheça padrões, aprenda e execute em novos conjuntos de dados. Mesmo com o avanço dos métodos, é latente a necessidade do tratamento dos dados por meio de diferentes técnicas de mineração de dados, que visam facilitar o processo de aprendizado do algoritmo.

Um dos problemas trabalhados pelo aprendizado de máquina é a classificação, que visa identificar a qual classe uma determinada observação pertence. O algoritmo de classificação é gerado através de um treinamento prévio utilizando um algoritmo de aprendizado de máquina, para que o modelo final seja capaz de identificar padrões em um conjunto de dados. Os algoritmos facilitam o processo de classificação em grandes volumes de dados, entretanto, desenvolver um modelo algoritmo para classificação não é trivial, possuindo limitações que podem prejudicar a avaliação desses modelos.

Dentre alguns entraves, destaca-se o problema de desbalanceamento entre as quantidades de exemplos entre as classes de um conjunto de dados. O problema de classes desbalanceadas aparece quando se tem um grande desequilíbrio na quantidade de indivíduos entre as classes definidas a priori. Isso acarreta um problema para a classificação, pois pode ocorrer uma desigualdade nos padrões entre os grupos, que surgem muitas vezes em informações de classes que são difíceis de se obter, as chamadas classes raras, e que poderiam gerar um grande número de exemplos errôneos, resultando em uma qualidade ruim do classificador.

Muitos dos conjuntos de dados utilizados apresentam desequilíbrio entre as classes, e muitas das vezes a classe em menor quantidade é a classe de interesse. Têm-se como exemplo, um estudo clinico que busca classificar a presença de uma doença rara que ocorre na minoria dos indivíduos de uma determinada população. O classificador, através dos critérios pré-estabelecidos, pode ter dificuldades em generalizar os indivíduos da classe de doença rara, gerando falsos resultados.

Diante do exposto, este trabalho apresenta técnicas que buscam solucionar o problema em questão, aplicado em um conjunto de dados desbalanceados do projeto CUME, estudo de coorte aplicado em alunos egressos de universidades mineiras. O conjunto de dados da pesquisa possui 2609 observações de graduados e pós-graduados, e através das características, é possível prever o comportamento sedentário do indivíduo. Definido a princípio, o comportamento sedentário

representa a classe minoritária de interesse, prejudicada na classificação pela sobreposição da classe dos indivíduos que não apresentam comportamento sedentário. No intuito de minimizar a quantidade de erros dos dados na discriminação das classes e, otimizando a classe de interesse no processo de classificação, define-se um modelo de análise através de métricas, comparado-o com os outros modelos, para que seja capaz de prever o comportamento sedentário pelo menos em metade dos indivíduos que possuem essa característica.

Objetiva-se, por conseguinte, pesquisar sobre os principais conceitos e técnicas de inteligência artificial relacionadas ao aprendizado em classes desbalanceadas e especificamente, pesquisar e aplicar métodos de classificação para um conjunto de dados desbalanceados. A revisão de literatura explora os conceitos de ..... O capítulo três explora os materiais e métodos utilizados, em seguida, o quarto apresenta e discute os resultados obtidos. O capítulo 5 encerra o estudo com as considerações finais.

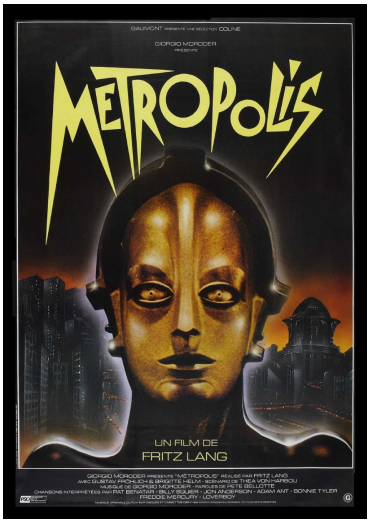
2 REVISÃO BIBLIOGRÁFICA

Neste capítulo apresentamos alguns dos principais conceitos de técnicas e métodos fundamentados em outros trabalhos de pesquisa publicados na literatura, ao qual nos baseamos para as análises presentes neste trabalho.

2.1 Inteligência artificial e Aprendizado de Máquina

Inteligência artificial (IA) instiga a curiosidade humana a partir de diversas perspectivas. No século passado, por exemplo, o cinema ainda mudo e em preto e branco, robôs, *softwares* e androides foram protagonistas de clássicos para admiradores de ficção científica e IA. Filmes como *Metropolis - 1927* e *2001: a space odyssey - 1968* Figura 1 remetem aos dias atuais uma compreensão do desenvolvimento de tecnologias avançadas. *Metropolis – 1927*, por exemplo, provoca nos telespectadores o seguinte questionamento: é possível um robô tomar a forma de um ser humano? Após 94 anos do lançamento do filme isso ainda é incerto.

Figura 1 – Filmes sobre inteligência artificial



[Metropolis]

[2001: a space odyssey]

Fonte: Google, 2021

Em *2001: a space odyssey - 1968*, um dos protagonistas do clássico é o computador **HAL9000**, ou simplesmente "Hal", uma inteligência artificial capaz de controlar uma nave, interagir com a tripulação e tomar decisões autônomas. Certamente, no ano de lançamento do filme toda essa capacidade de processamento que Hal apresentava era considerada como conceitos ficcionais e imaginativos. No entanto, após algumas décadas de lançamento, as máquinas estão aprendendo a dirigir automóveis. Já existem carros que podem se mover sem motorista, interagir com os usuários e tomar decisões autônomas, como os carros das empresas Google e Tesla. A sensação transmitida é de que a ficção científica torna-se realidade, e vice-versa.

Figura 2 – Carros que dirigem sozinhos

[Tesla modelo 3.]  [Waymo.] 

Fonte: Google, 2021

O termo Inteligência Artificial foi utilizado pela primeira vez pelo professor John McCarthy em 1955, numa proposta de conferência de pesquisa de verão de dois meses para 10 pessoas, intitulada como "inteligência artificial". McCarthy escreveu: "propomos que um estudo de inteligência artificial seja realizado durante o verão de 1956 no *Dartmouth College*[1]. O estudo deve prosseguir com base na conjectura de que todos os aspectos da aprendizagem e qualquer outra característica da inteligência podem, em princípio, ser descritos com tanta precisão que uma máquina pode ser feita para simulá-los. Será feita uma tentativa de descobrir como fazer as máquinas usarem a linguagem, formar abstrações e conceitos, resolver tipos de problemas agora reservados aos humanos e melhorar a si mesmas. Achamos que um avanço significativo pode ser feito em um ou mais desses problemas se um grupo cuidadosamente selecionado de cientistas trabalharem juntos durante um verão" (RAJARAMAN, 2014).

Além de cunhar o termo, John McCarthy contribui para o campo por mais de cinco décadas. Essa contribuição é responsável por atribuir ao professor o título de "*Father of Artificial Intelligence*"[2]. Segundo Rajaraman (2014), John McCarthy utilizou o termo para descrever programas de computador que aparentemente exibem inteligência, ou seja, os computadores realizam tarefas que, quando realizadas por humanos, requerem que sejam inteligentes. Ele acreditou ao longo da vida no uso da lógica matemática para descrever o conhecimento, incluindo o conhecimento do senso comum, o que levou ao desenvolvimento do assunto da representação do conhecimento.

Em 1958, McCarthy inventou a linguagem de programação de computador LISP – linguagem de programação que perdurou por mais de cinquenta anos – para resolver problemas de Inteligência Artificial. Rajaraman (2014) afirma que a linguagem LISP é a segunda linguagem de programação mais antiga depois do FORTRAN. Além de suas contribuições técnicas, suas contribuições como professor foram fundamentais na criação de duas escolas famosas em Inteligência Artificial: uma no Instituto de Tecnologia de Massachusetts (MIT) e outra na Universidade de

[1] A Faculdade de Dartmouth é uma universidade estadunidense fundada em 1769, localizada na região nordeste dos Estados Unidos, na cidade de Hanover, no estado de New Hampshire.

[2] Pai da Inteligência Artificial

Stanford, Califórnia.

2.2 Classificação

Para a tarefa de classificação existem diversos métodos, com abordagens e técnicas distintas para a resolução desse problema, e não há uma regra geral que defina qual é o melhor método, pois os métodos são implementados de acordo com os diversos tipos de dados, dependendo do método pode ter vantagem em desempenho, em qualidade tempo de processamento, entre outras características. Apresentaremos alguns tipos de classificadores.

Em problemas que utilizam aprendizado supervisionado, cada exemplo é descrito por um vetor de atributos de valores de características e por um especial, que descreve uma característica de interessados em criar um modelo (HARRINGTON, 2012). Tais atributos podem ser discretos, ordinais ou contínuos. No caso de discreto, a problemática é conhecida como classificação, que segundo (MOREIRA, 2005), consiste no problema de identificar as características mais significativas de uma nova observação de um conjunto (subpopulação), e através de uma hipótese de classificação ser alocada para a categoria já definida.

Classificação é uma tarefa frequentemente utilizada e pode ser encontrada em todas as áreas do conhecimento humano. Na medicina por exemplo, essa tarefa é comum para predizer se um tumor é benigno ou maligno, na área financeira a previsão se uma pessoa é boa pagadora ou não, se é passível de concessão de crédito no banco, se uma operação é fraudulenta ou não. Classificar é comum e fundamental para a atividade humana, e o desenvolvimento de sistema computacional permite realizar essas tarefas de forma automática e imprescindível.

Ferramentas computacionais são muito utilizadas em processos de classificação com grande volume de dados. Algoritmos são funções que, na etapa de aprendizado supervisionado, atuam em duas etapas, treinamento, onde o algoritmo é capaz de aprender com os dados de entrada através de registros de atributos específicos da classe. A outra etapa é a de predição ou dados de teste, na qual o algoritmo é capaz de reconhecer os padrões estabelecidos no treinamento, sendo capaz de reproduzi-los em um novo conjunto de dados e alocar os indivíduos nas classes pré-estabelecidos. Os algoritmos facilitam o processo de classificação em grande volume de dados, mas desenvolver um modelo algoritmo para classificação não é trivial, possuem limitações que podem prejudicar a avaliação desses modelos. Entre eles está o problema de desbalanceamento entre as quantidades de exemplos entre as classes de um conjunto de dados.

Os algoritmos tradicionais de classificação apresentam dificuldades em discriminar as classes com poucos representantes (classes minoritárias), um exemplo é o campo da medicina, no problema de diagnóstico de doenças raras, o qual ocorrem em quantidades bem menores do que a quantidade da população total, o objetivo do reconhecimento é detectar a parte doente da população.

Algoritmos de classificação são definidos classificadores, que trabalham de acordo com seus dados de entrada, sejam de duas classes no formato simples “sim/não” ou “pertencente/não pertence” definido de classes binárias ou “multiclasses”, que são definidas pelo número finito e maior que duas categorias ou classes, lineares ou não lineares, enfim, é importante definir os tipos de classificadores de acordo com a distribuição dos dados de entrada.

2.2.1 Classificadores

A classificação é uma tarefa de análise de dados e identificação de padrões, que em grandes volumes de dados necessita de auxilio de ferramentas computacionais para construir um classificador, ou seja, que atribua uma classe a certas observações descritas por um conjunto de atributos. Considerando que existem diferente métodos, abordagens e técnicas para a resolução de problema, e que não há regra geral que define qual é o melhor método, visto que os métodos são implementados de acordo com diferentes distribuições e características, apresenta-se diferentes tipos de classificadores.

Por exemplo, a classificação linear tem como objetivo encontrar uma função linear capaz de separar as amostras em suas respectivas classes, construindo assim o classificador com a menor taxa de erro possível. Tem-se o exemplo de uma base de dados bidimensional e com duas classes separáveis linearmente. Com o classificador, encontra-se uma função linear capaz de produzir uma reta separando as observações das duas classes. Observa-se que, na Figura 3, os classificadores lineares em duas bases de dados. Em (a) uma reta separa as amostras e, em (b) um hiperplano é projetado separando as amostras em mais de uma dimensão.

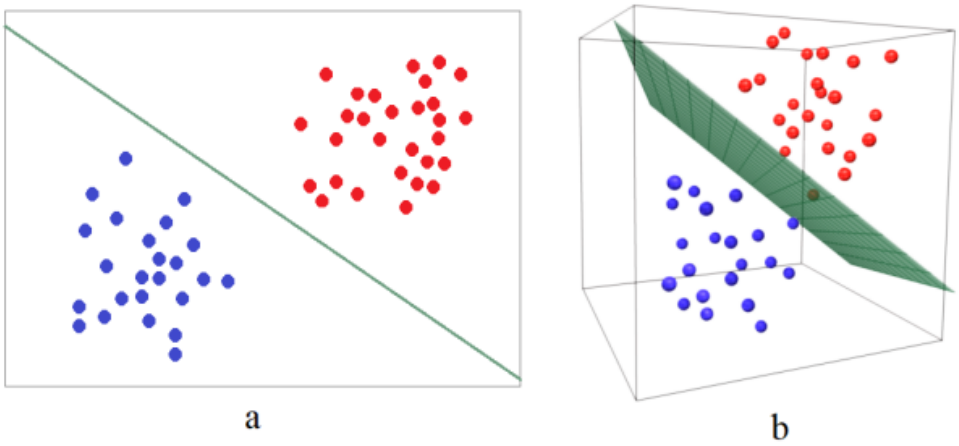


Figura 3 – Classificadores lineares

Já os classificadores não lineares são dedicados às amostras que não sejam separáveis em classes por uma função linear, as quais teriam uma taxa de erro alta caso se utilizasse um classificador linear. Diante disso, é necessário um classificador mais robusto com uma função não linear, que tenha um desempenho apropriado para separar as amostras em suas devidas classes. Na Figura 4 observa-se que a base de dados não é de classes separáveis linearmente. Em (a) o

classificador linear não consegue separar as amostras corretamente, mas em (b) o classificador não-linear é capaz de separar as amostras de cada classe com maior precisão.

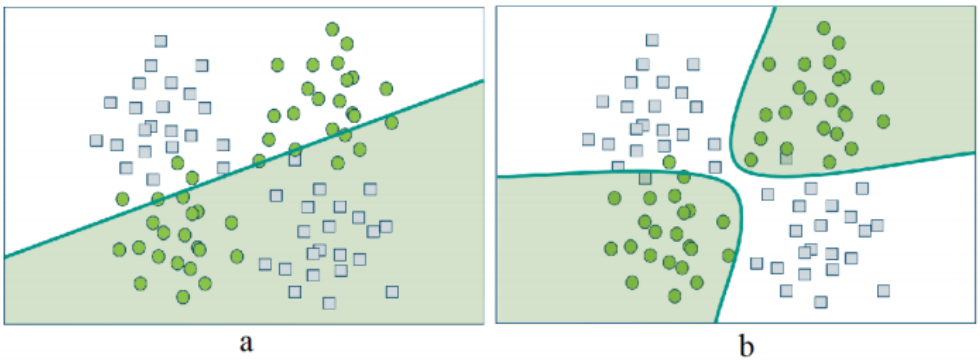


Figura 4 – Separação linear e não linear

2.2.2 Naïve Bayes

O classificador de Naive Bayes ou Bayes ingênuo é um algoritmo de aprendizado supervisionado linear, fortemente baseado na teoria de probabilidade condicional de Thomas Bayes, que através de fórmula matemática, calcula a probabilidade de ocorrência de um evento dado que outro já ocorreu – conhecida como probabilidade condicional e com suposição de independência entre os preditores. Uma vez calculado, o modelo de probabilidade pode ser usado para fazer previsões para novos dados usando o Teorema de Bayes. É frequentemente usado na análise e na classificação de textos e utilizados, por exemplo, em filtros de *spams* (KELLEHER; NAMEE; D'ARCY, 2015).

O classificador Naive Bayes calcula a probabilidade de que um determinado indivíduo pertença a uma determinada classe, sendo X uma observação ou indivíduo, descrito por seu vetor $(x_1 , ..., x_n)$, e uma classe de destino y. O teorema de Bayes nos permite expressar a probabilidade condicional $P (y \mid X)$ como um produto de probabilidades mais simples usando a suposição de independência. "A vantagem de assumir que todos os preditores são independentes é que isso reduz drasticamente a complexidade dos cálculos a serem realizados" (BURGER, 2018).

"Apesar de se assumir independência condicional entre os preditores, o algoritmo Naive Bayes se mostra eficaz na classificação" (HARRINGTON, 2012), "geralmente fornecendo modelos que possuem boa acurácia" (KELLEHER J.D.; NAMEE, 2015.).

Através do teorema de Bayes pode se afirmar a seguinte relação, dada variável de classe y e vetor de recurso dependente x_1 através x_n:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\, P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)} \tag{2.1}$$

Diante da suposição de independência condicional de que

$$P\left(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right) = P\left(x_i|y\right), \tag{2.2}$$

para todos i, esta relação é simplificada para

$$P\left(y \mid x_1, \ldots, x_n\right) = \frac{P\left(y\right) \prod_{i=1}^{n} P\left(x_i \mid y\right)}{P\left(x_1, \ldots, x_n\right)} \tag{2.3}$$

Onde $P(x_1, \ldots, x_n)$ é constante dada a entrada, podemos usar a seguinte regra de classificação:

$$P\left(y \mid x_1, \ldots, x_n\right) \propto \ P\left(y\right) \prod_{i=1}^{n} P\left(x_i \mid y\right), \tag{2.4}$$

$$\hat{y} = argmax P\left(y\right) \prod_{i=1}^{n} P\left(x_i \mid y\right) \tag{2.5}$$

O algoritmo Naive Bayes elenca a observação na classe que tem a maior probabilidade (KELLEHER J.D.; NAMEE, 2015.). Esse método é denominado estimativa MAP (do inglês, *Maximum A Posteriori*), que pode usar como estimativa para $P(y)$ e $P\left(x_i \mid y\right)$.

O classificador Naïve Bayes se mostra simples, eficiente e rápido, segundo (BEERAVELLI et al., 2018), e se adapta para lidar com valores ausentes (KELLEHER J.D.; NAMEE, 2015.). Ademais, a característica de estimação independentemente das distribuições com propriedades condicionais de classe, corroboram a eficiência do classificador. Considerando, por exemplo, um classificador para avaliar o risco em pedidos de crédito, parece contraintuitivo ignorar as correlações entre idade, nível de educação e rendimentos.

O classificador também apresenta desvantagens, tais como não ter bom desempenho para variáveis explicativas contínuas e, para observações faltantes de variáveis categóricas ele atribui probabilidade zero e exclui da previsão. Contudo, o classificador de Naïve Bayes se mostra eficiente, simples e pode ser aplicado em múltiplas distribuições de probabilidades, logo, para diferentes suposições de distribuição pode-se trabalhar uma variação de classificador do tipo Naïve Bayes, buscando o melhor modelo de acordo com a distribuição de probabilidade dos dados de entrada.

2.3 Dados desbalanceados

No aprendizado de máquina supervisionado o algoritmo pode apresentar limitações, que podem prejudicar a predição do classificador, dentre elas está o desbalanceamento da quantidade de observações entre as classes de um conjunto de dados.

Um conjunto de dados é dito como desbalanceado quando há uma desproporção muito alta entre uma ou mais classes (CHAWLA; JAPKOWICZ; KOTCZ, 2004). Por exemplo, observando-se o caso de um estudo populacional de um determinado tipo de doença rara a existência de um grande desequilíbrio entre o número de não portadores, ou seja, o número de ocorrências na categoria. O desequilíbrio ocorre por diversos fatores, como por exemplo, pela impossibilidade de se coletar quantidades iguais de amostras pertencentes às classes, ou, pode ser característica da própria amostra que realmente reflita uma população desbalanceada.

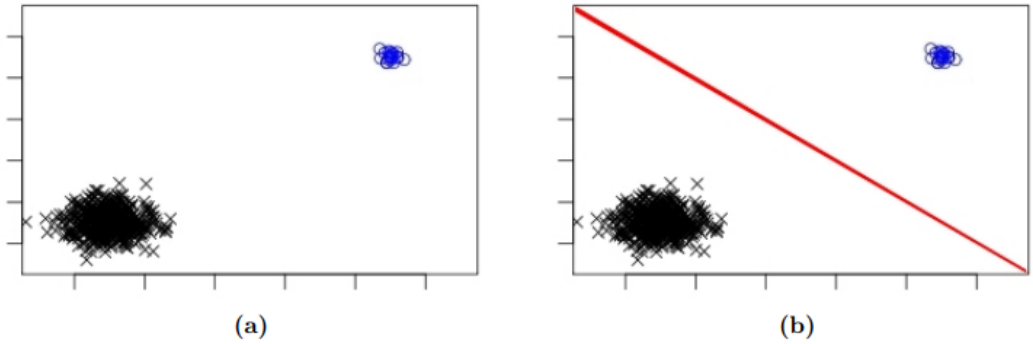


Figura 5 – Amostras com classes separáveis

Assim, o desbalanceamento das classes passa a ser um problema, pois os classificadores tradicionais partem da suposição das classes serem equilibradas, o que pode resultar em uma discriminação ineficiente (MONARD; BATISTA, 2002.). O algoritmo classificador, ao receber o conjunto de dados, discrimina com grande taxa de acerto a classe dos exemplos majoritários ou de maior proporção. Normalmente, os indivíduos da classe minoritária ou de menor proporção não apresentam uma boa taxa de classificação. Ou seja, as classes majoritárias são favorecidas enquanto as classes minoritárias possuem baixa taxa de reconhecimento, e que pode incorrer das classes de menor quantidade serem a classe de interesse.

Porém, o problema em questão não é somente o desequilíbrio das quantidades de indivíduos em cada classe, mas também na separabilidade das classes em questão (PRATI; BATISTA; MONARD, 2004), já que as classes desequilibradas podem se encontrar agrupadas e separáveis em um espaço, como é apresentado na Figura 5, que explicita que (a) as classes desbalanceadas e agrupadas no espaço em (b) são facilmente separadas por uma reta, simulando um funcionamento de um classificador linear em conjunto de dados onde as classes são desbalanceadas e separáveis.

O problema de dados desbalanceados ocorre quando as amostras das classes estão misturadas em um mesmo espaço, dificultando a separabilidade. A classe majoritária, representada pela maior quantidade, referente à outra categoria e misturada no mesmo espaço, resulta na sobreposição dos dados desbalanceados, prejudicando a discriminação da classe minoritária como pode ser observado na Figura 6.

Diante desse problema, é necessário um tratamento para que a classificação não sofra essas induções na discriminação, causadas pela sobreposição de classes, viabilizando um processo

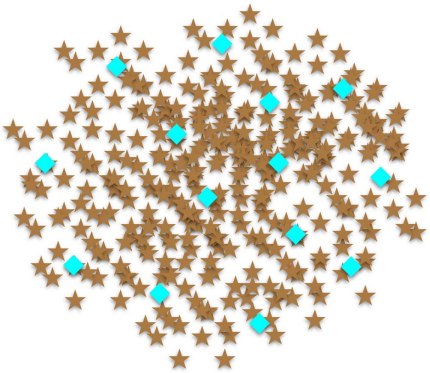


Figura 6 – Classes sobrepostas

de classificação eficiente e preciso. A literatura apresenta alguns métodos, tal como técnicas de amostragem que atuam na quantidade de observações da amostra e equilibrando as classes, seja na retirada de indivíduos da classe majoritária (subamostragem) ou na replicação dos indivíduos de classe minoritária (sobreamostragem), ou ainda, na adaptação de algoritmos, que através do aprendizado de outros algoritmos fracos teriam um algoritmo final forte e mais preciso.

2.4 Adaboost

O método *boosting* consiste em melhoraramento de um algoritmo base, ajustando o seu funcionamento de acordo com os erros cometidos pelo classificador anterior, de acordo com Freund e Shapire, sendo definido como um método para aprimorar o desempenho de qualquer algoritmo de aprendizado (FREUND; SCHAPIRE et al., 1996). O Adaboost, por sua vez, é da família do método *Boosting*, que faz uma adaptação no algoritmo, atribuindo pesos a todas as instâncias, considerando os erros dos classificadores base e na tentativa de obter alta precisão.

O Adaboost combina vários classificadores para gerar respostas melhores no classificador base, o adaboost atribui pesos mais altos nos indivíduos classificados erroneamente nos dados de treinamento para minimizar o erro, que é calculado e atualizado no conjunto de treinamento ao qual foram incorretamente classificados, esse processo ocorre de forma iterativa, ou seja, o processo ocorre T vezes, até que o erro seja minimizado.

Os pesos são atribuídos aos dados de treinamento, onde $D_1\left(i\right) = \frac{1}{m}$ para cada m entrada, logo, para cada classificador, h_t é definido erro ponderado ϵ_t, que é o somatório dos pesos D_t dos dados classificados errados.

$$D_{t+1}(i) = \frac{D_t\left(i\right)}{Z_t} e^{-\alpha_t y_i h_t(x_i)} \tag{2.6}$$

O Z_t é uma constante de normalização que garante a soma do vetor D_{t+1} seja 1, e y_i é a saída desejada igual a 1 para entradas positivas, e -1 para negativas. $h_t(x_i)$ é o resultado do classificador para o a entrada x_i, e α_t é o peso dado para o classificador fraco h_t na composição do classificador forte.

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{2.7}$$

A fórmula acima mostra como é calculado o peso dado ao classificador fraco. A ideia do algoritmo adaboost é buscar em cada iteração o classificador fraco que melhor separa os indivíduos, dando um maior peso aos indivíduos que poucos classificadores fracos acertaram em iterações anteriores.

Contudo, o método do Adaboost, por ser flexível e possuir simplicidade na implementação, é utilizado para o problema de classes desbalanceadas, fazendo com que o classificador considere somente exemplos positivos na fase de treinamento, visando otimizar a precisão e sensibilidade do classificador. Diversas variações do método foram desenvolvidas, sendo o Adaboost o padrão para problemas de classes binárias, e o AdaboostM1 para problemas multiclasses, por exemplo.

2.5 SMOTE

Técnicas de sobreamostragem são baseadas em replicação ou reprodução sintética de dados preexistentes da classe minoritária, representada pela classe de menor quantidade de instâncias em relação a uma classe de maior quantidade de instâncias, também denominada classe majoritária. Com a finalidade de equilibrar as classes para ter uma melhor discriminação dos classificadores do que na distribuição original (JO; JAPKOWICZ, 2004), não há garantia de que a distribuição original dos dados seja a mais adequada para a construção de classificadores. Contudo, apenas replicar casos existentes das classes de interesse, aumenta o viés do classificador para essa mesma classe, prejudicando a generalização do classificador (CHAWLA; JAPKOWICZ; KOTCZ, 2004), o que pode gerar problemas de termos de região de decisão no espaço de atributos.

Diante dessa problemática, (CHAWLA; JAPKOWICZ; KOTCZ, 2004) desenvolveram um método, definido como SMOTE (*Synthetic Minority Oversampling Technique*), técnica de sobreamostragem que consiste em balancear as classes através da replicação ou da geração de dados sintéticos para a classe minoritária, a partir dos dados já existentes, baseado na vizinhança de cada caso da classe.

O SMOTE consiste na replicação de dados sintéticos na vizinhança de cada instância a partir dos indivíduos preexistentes que são aleatoriamente selecionados na classe de minoritária, que muitas das vezes é a classe de interesse. O método ignora os indivíduos da classe majoritária, fazendo com que as classes fiquem equilibradas, o que pode resultar em exemplos ambíguos se houver uma forte sobreposição. As atribuições dos dados sintéticos são criadas a partir da

diferença entre o vetor de características (amostra) em consideração ao seu vizinho mais próximo, que são interpoladas no espaço amostral e inseridos aleatoriamente ao longo de cada um do segmento de reta que une o caso da classe de interesse a um de seus k vizinhos mais próximos de forma aleatória. Essa diferença é multiplicada por um número aleatório entre 0 e 1 e adicionado ao vetor de característica em consideração, o algoritmo de KNN, para criar seus representantes e os k vizinhos mais próximos, escolhe os dados de forma aleatória da classe minoritária, em seguida, são definidos k-vizinhos mais próximos dos dados, assim as observações sintéticas seriam então feitas entre os dados aleatórios e o vizinho mais próximo k selecionado aleatoriamente. O procedimento é repetido n vezes até que a classe minoritária tenha a mesma proporção que a classe majoritária ou da quantidade desejada.

O método SMOTE funciona para que o classificador construa regiões de decisão maiores que contenham pontos de classe minoritária próximos, no entanto, pode ter um efeito adverso, ou seja, criar casos positivos que violam o espaço de decisão das categorias negativas. Esse efeito é chamado de sobreposição de classes, consoante (BATISTA; ENUMO, 2004), pode atrapalhar o desempenho do classificador por tais dados, mas que também pode ser uma característica natural dos dados apurados pelo SMOTE.

2.6 Pré-Processamento

O pré-processamento engloba um conjunto de análises iniciais na base de dados para se ter um entendimento do comportamento dos dados, além de fazer tratamentos, caso necessários. De acordo com (KLEMETTINEN; MANNILA; TOIVONEN, 1999), a fase de pré-processamento é a mais complexa, podendo tomar até 80 % do tempo do processo.

De outra forma, o pré-processamento envolve um conjunto de análises iniciais dos dados com a finalidade de tirar informações de eventuais problemas estruturais da base de dados, e com um conjunto de operações de limpeza para poder minimizar ou eliminar erros, vieses ou falhas, que podem aparecer como inconsistências, poluição, atributos duplicados ou redundantes, valores em branco ou *defaults* e dados com classes desbalanceadas. Estes possíveis erros podem ser trabalhados por funções do pré-processamento, sendo elas:

- Seleção de dados: idealmente, seria interessante considerar todas as variáveis do conjunto, por ter uma boa resposta dos modelos, porém torna-se necessário quando o conjunto de dados é de grande volume. Logo, é recomendada a seleção dos atributos que realmente serão usados no processo.

- Limpeza dos Dados: a limpeza consiste em garantir a qualidade dos dados, eliminando fatores desfavoráveis de forma que não prejudique o resultado final. A retirada das observações problemáticas, a atribuição de valores padrões e a aplicação de técnicas de agrupamento,

são algumas das técnicas utilizadas nessa etapa que auxiliam a descoberta dos melhores valores.

- Integração dos Dados: analisar profundamente os dados coletados para que se possa integrá-los de forma consistente e apta à obtenção de resultados satisfatórios.
- Transformação dos Dados: através dela constitui-se um padrão para os dados, facilitando o uso das técnicas computacionais de análises. Geralmente alguns algoritmos requerem um formato apropriado, existindo assim a necessidade de aplicação de transformação destes dados.
- Redução dos Dados: utilizar técnicas para que o conjunto de dados original seja convertido em um conjunto menor, porém sem perder o valor dos dados originais.

2.7 Métricas de avaliação para modelo de classificação

Diversos métodos e métricas são usados na avaliação da qualidade do modelo de classificação, ao pelos quais podem-se fazer comparações, ver suas estabilidades e qualificar sua classificação. Alguns conceitos importantes são apresentados para entender algumas dessas métricas:

- Matriz de confusão é uma matriz quadrada $q\ x\ q$, onde q representa o número de rótulos de classe envolvidos no problema. Apresenta a frequência de indivíduos distribuídos nas classes no processo de classificação, armazenando os erros e acertos realizados pelo classificador ao alocar os indivíduos no conjunto de dados de teste. A célula c_{ij} aponta o número de indivíduos de teste que o classificador associou à classe i e que, de fato, pertencem à classe j. Desta forma, as células da diagonal principal sempre irão conter o número de objetos corretamente classificados pelo modelo. As frequências são denotadas da seguinte maneira:
  - Verdadeiro Positivo (VP): ocorre quando um indivíduo i é classificado corretamente pela classe k, que é a sua correspondente, ou seja, sua classe real.
  - • Verdadeiro Negativo (VN): ocorre quando um indivíduo i é classificado corretamente na sua classe, sendo ela diferente de k.
  - Falso Positivo (FP): ocorre quando um indivíduo i é classificado incorretamente na classe k, sendo k a classe real do indivíduo.
  - Falso Negativo (FN): ocorre quando um indivíduo i é classificado incorretamente na sua classe, sendo sua classe diferente de k.

		Valor Predito	
		Positivo	Negativo
Real	Positivo	Verdadeiro positivo ((VP)	Falso Negativo (FN)
	Negativo	Falso positivo (FP)	Verdadeiro negativo (VN)

- Acurácia é a métrica mais intuitiva e simples usada para avaliar modelos de classificação. É basicamente a porcentagem de classificados corretamente pelo modelo. Mesmo sendo a métrica bem simples, ela pode não dar uma boa diretriz para o modelo, tal como para classes desbalanceadas, acarretando na ocorrência de de mais verdadeiros positivos e verdadeiros negativos do que falsos negativos positivos, sendo ainda mais se as proporções forem bem altas com de 80 %. Logo é salutar basear-se em mais de uma métrica para avaliar o modelo.

$$Acurácia = \frac{VP + VN}{P + N} \tag{2.8}$$

- A precisão é uma métrica que é dada a partir da razão entre os valores positivos (VP) e os totais preditos (VP+FP), isso nos diz o quanto dos verdadeiros positivos realmente são positivos.

$$precisão = \frac{VP}{VP + FP} \tag{2.9}$$

- *Recall*, Revocação ou Sensibilidade: essa métrica é a frequência na qual o modelo prevê corretamente cada indivíduo a sua classe correta, ou seja, o número de vezes que uma determinada classe foi prevista corretamente sobre o número de observações dessa classe no conjunto de dados, quando considerada a classe real.

$$recall \ = \frac{VP}{VP + FN} \tag{2.10}$$

- F1-Score é a métrica da média harmônica entre a precisão e o *recall*, com valor de 0 a 1.

$$F1Score = \frac{2 * Precisiã * Recall}{Precisião + Recall} \tag{2.11}$$

- Especificidade é a probabilidade do indivíduo i ser classificado como diferente da classe k e não ser alocado para sua classe real.

$$Especificidade = \frac{VN}{FP + VN} \tag{2.12}$$

- ROC (*Receiver Operating Characteristics*) e AUC (*Area Under the Curve*) são duas métricas utilizadas para medir o desempenho de classificadores. ROC é uma curva de probabilidades da taxa de Verdadeiros Positivos, representada pelo eixo vertical sobre a taxa de Falsos positivos, representada pelo eixo horizontal. A área abaixo da curva gerada é conhecida como o AUC, que é medida de 0 a 1, quanto mais próxima de 1, melhor a capacidade de classificação do seu modelo.

3 Material e Métodos

3.1 Obtenção dos dados

Este trabalho utiliza dados do projeto CUME, estudo de coorte aberto realizado com alunos de graduação ou pós-graduação de Instituições Federais de Ensino Superior (IFES), de Minas Gerais, desde 1994. A característica deste estudo é o recrutamento permanente de participantes com base na coleta de dados *online*, de forma que a amostra continua a crescer a cada dois anos.

A primeira coleta foi realizada entre março de 2016 e agosto de 2016. Participaram desses estudos, graduados da UFMG e da UFV, que concluíram a graduação ou pós-graduação entre 1994 e 2014. Em 2018, o questionário de linha de base Q 0 foi enviado de volta para a UFMG e UFV que não responderam em 2016, bem como para recém-formados ou pós-graduados de 2015 a 2017. Além da UFOP, aos da UFLA e da UFJF, que se formaram de 1994 a 2017.

Os dados são coletados em plataforma online a cada dois anos, o que contribui para o crescimento da amostra. Os critérios de inclusão dos participantes são: idade acima de 18 anos e ter concluído o curso de graduação ou pós-graduação nas instituições já mencionadas anteriormente. O questionário de coleta de dados do projeto CUME pode ser acessado no seguinte sítio eletrônico: (WWW.PROJETOCUME.COM.BR, 2021). Os dados coletados por meio do questionário foram automaticamente registrados e exportados para uma planilha Excel.

O estudo "Coorte das Universidades Mineiras (CUME): impacto do padrão alimentar brasileiro e da transição nutricional sobre as doenças e agravos não transmissíveis" foi aprovado pelos Comitês de Ética em Pesquisa com Seres Humanos da UFMG, da UFV, UFOP, UFJF e UFLA sob os pareceres CAAE: 07223812.3.3001.5153, 44483415.5.1001.5149, 44483415.5.2003.5150, 44483415.5.2004.5133; 44483415520045133 (pareceres nos anexos A; B; C e D).

O intervalo de amostragem da linha de base consiste em todos os participantes da coorte CUME, todos os participantes incluindo gênero. Esses participantes responderam totalmente ao questionário Q_0 e atenderam aos seguintes critérios de seleção, a saber: $i)$ ter concluído a graduação ou curso de pós-graduação na instituição CUME participante; e $ii)$ ser brasileiro e residir no Brasil.

A coleta dos dados foi realizada através de questionários online autopreenchidos que foram enviados por *e-mail* a cada participante. Foram excluídos os participantes que não preencheram o questionário completa ou incorretamente e os que não eram naturalmente brasileiros ou residentes do Brasil no ano anterior à coleta. A partir dos dados coletados no projeto CUME, foi utilizada amostra de 2609 indivíduos e 6 variáveis de estudo.

3.2 Variáveis do estudo

No $Q - 0$, dados sociodemográficos, estilo de vida e antropométricos também foram coletados, sendo eles, idade (< 40 anos ou $\geq$ 40 anos - (OLIVEIRA; PEREIRA, 2019)) Para avaliação da renda familiar ($<$ 5 salários-mínimos ou $\geq$ 5 salários - mínimos), foram considerados o recebimento de menos que cinco salários-mínimos ou mais que cinco salários-mínimos ($<$ 5 salários-mínimos ou $\geq$ 5 salários-mínimos), baseados no valor proposto em 2020 (DOMINGOS et al., 2018).

A prática de exercício físico nas horas vagas baseia-se na verificação de um questionário para indivíduos com ensino superior por investigadores espanhóis da coorte do SUN. Além da segunda parte do questionário também incluir questões sobre o tempo despendido em atividades sedentárias, bem como uma série de atividades esportivas e de lazer e seu tempo / frequência (MARTÍNEZ-GONZÁLEZ et al., 2005). O exercício físico durante o lazer é considerado uma atividade física realizada pelo indivíduo não-essencial à vida diária, mas é realizada pelo indivíduo a seu critério. Essas atividades incluem a participação em esportes, preparação física ou treinamento e atividades recreativas, como caminhada, dança e jardinagem (BOSMAN et al., 2010).

As perguntas incluem o número de dias por semana, a duração média e a intensidade do exercício físico. Ou seja: "Nos últimos 12 meses, quantos dias você se exercitou ou se exercitou em média por semana?": menos de 10 minutos; entre 10 e 19 minutos; entre 20 e 29 minutos; entre 30 e 39 minutos; entre 40 e 49 minutos; entre 50 e 59 minutos; mais de 60 minutos. Essas faixas de tempo foram transformadas em valores absolutos por meio do cálculo da média entre aos minutos indicados por elas.

Dessa forma, de acordo com as normas da Organização Mundial da Saúde (BOSMAN et al., 2010), esses valores foram encontrados e os participantes foram classificados como fisicamente ativos ou insuficientemente ativos fisicamente. Indivíduos que realizam $\geq$ 150 minutos de exercício de intensidade moderada ou $\geq$ 75 minutos de exercício de intensidade vigorosa por semana são considerados exercícios físicos. (BOSMAN et al., 2010).

Em relação à antropometria, os dados auto referidos de peso e estatura são utilizados para classificar os indivíduos com sobrepeso e com excesso de peso a partir do cálculo do índice de massa corporal IMC, que é composto pelo peso (kg) dividido pelo quadrado da altura (metros quadrados). Conforme declarado por Miranda dentre outros, o peso e a altura auto-relatados foram verificados por uma subamostra de participantes do projeto CUME (2017).

A variável da classe de interesse definida pelo comportamento sedentário, tem o termo direcionado para as atividades que são realizadas na posição deitada ou sentada e que não aumentam o dispêndio energético acima dos níveis de repouso (AINSWORTH et al., 2000). Com isso, foi utilizada para definir o comportamento sedentário, a mediana do tempo sentado total. (BAUMAN et al., 2011).

3.3 Seleção do método

Inicialmente, antes da entrada dos dados no modelo foi feita uma análise inicial dos dados buscando descrever o comportamento dos dados e verificar possíveis incoerências e erros que podem prejudicar a etapa de aprendizagem do modelo.

Após a análise inicial, o conjunto de dados foi dividido aleatoriamente em dois conjuntos, sendo eles, treinamento e teste. O conjunto de treinamento forneceu o processo de aprendizagem do modelo nas etapas de classificação, constituído de 70 % dos dados de entrada; os 30 % das observações restantes dos dados de entrada foram destinadas ao conjunto de teste, pelo qual seria avaliada a resposta de aprendizado do classificador diante de novas observações. Esse processo se repetiu para cada método de classificação aplicado ao conjunto de dados mantendo as mesmas condições.

Nessa etapa, a classificação dos dados foi obtida através do aprendizado do modelo de classificação Naïve Bayes. O modelo mapeou o conjunto de dados de entrada em um número finito de classes de interesse definidas a priori, sendo elas, "não sedentário" e "sedentário, e rotuladas com 0 e 1, respectivamente.

Com intuito de se ter um classificador mais adequado, aplicou-se o método de aprendizagem supervisionado Adaboost para esse conjunto de classes binária. O conjunto de dados passa por uma etapa de treinamento e de teste para a validação do modelo, os dados são divididos em quantidades para treino e teste sendo 70 % e 30 % dos dados para cada conjunto respectivamente e, dentro desses conjuntos ocorre a classificação, sendo avaliada através das métricas.

Com a finalidade de selecionar um método diferente e o mais adequado ao nosso problema em questão, foi utilizado o método SMOTE, que foi aplicado no conjunto de dados deixando as classes em quantidades equilibradas para uma classificação mais efetiva.

Após a classificação, os resultados de cada método foram analisados e comparados às métricas de avaliação de cada método aplicado, e através da comparação entre as métricas foi selecionado o modelo mais adequado para o problema de pesquisa.

Utilizou-se o RStudio para as análises, que se trata de software estatístico livre e de ambiente de desenvolvimento integrado para R-program (R, 2021), uma linguagem de programação para gráficos e cálculos e análises estatísticas. Nessas análises iniciais foi necessária utilização de alguns pacotes para otimizar o processo em velocidade e funções para adequações da base.

4 Resultados e Discussões

Nesta seção são apresentados os resultados das análises feitas na pesquisa acerca do problema de classificação em conjunto de dados desbalanceados, sendo de maneira organizada, respeitando todo o processo, com as discussões e evidências necessárias para as conclusões finais.

4.1 Observação e limpeza da base de dados

Foi feito um pré-processamento para observar a estrutura da base de dados, conhecer as variáveis e verificar a possibilidade de valores faltantes do conjunto de dados, sendo localizados e feitos os devidos tratamentos quando necessários, para a entrada dos dados no modelo.

Na Figura 7, observa-se que, as variáveis não possuem nenhum valor faltante, logo, não foi necessário nenhum tratamento para dados faltantes.

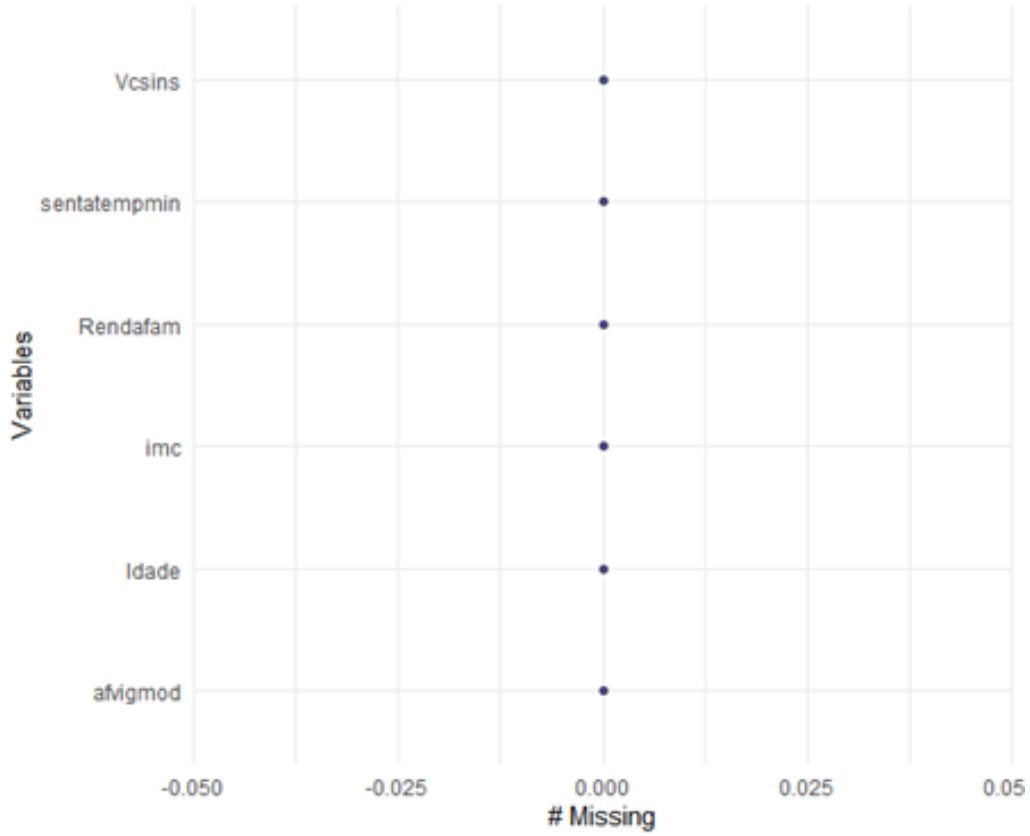


Figura 7 – Valores faltantes em cada variável

Na Figura 8, com a finalidade de observar as possíveis relações entre as variáveis, apresenta-se um correlograma. Verifica-se, através do gráfico, a gradação das cores em relação às correlações positivas e negativas. A baixa correlação entre as variáveis, bem próximas

de 0, quase nula, torna a relação desprezível entre as variáveis, e assim verifica a independência entre elas. (Verifica-se através do gráfico as baixas relações entre as variáveis do conjunto, sendo as que apresentam uma maior relação associam-se à variável Idade com Rendafam e imc, porém assumem um valor máximo de 0,24 e 0,27 respectivamente, sendo elas inferiores a 0,3, sendo uma relação desprezível, verificando a característica de independência entre as variáveis).

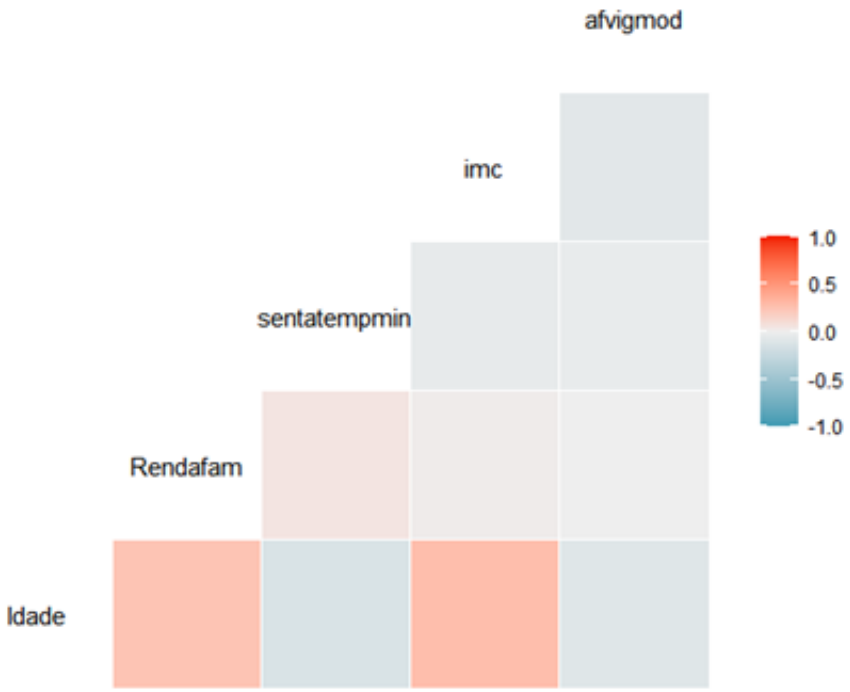


Figura 8 – Correlograma das variáveis

A Figura 9 apresenta a disparidade do tamanho amostral em relação às classes “não sedentário” e “sedentário”, nas quais 85 % das observações são da primeira categoria, rotuladas por 0 e 1, respectivamente. Na classe 0 estão alocados 2269 indivíduos, enquanto a classe 1 possui 340 indivíduos.

A Figura 10 apresenta a distribuição de frequência de cada variável e sua representatividade para cada classe. Nota-se desproporção das classes, sendo interessante observar a variável “sentatempmin”, que representa o tempo que o indivíduo trabalha sentado, quanto maior o tempo que a pessoa trabalha sentado, maior é a frequência de indivíduos sedentários.

Na Figura 11 apresenta-se o desequilíbrio da quantidade de indivíduos representados pelas classes. O principal fator apresentado é a dificuldade em discriminar as classes, visto que os dados das classes encontram-se sobrepostos. Para o problema de classificação, isso ocorre devido a indução dos os algoritmos tradicionais, que tendem a favorecer a classe majoritária quando as classes estão desequilibradas. Em um cenário de sobreposição, a classe minoritária acaba sendo prejudicada por falsas induções implicando em erros de classificação. Diante disso, nota-se a

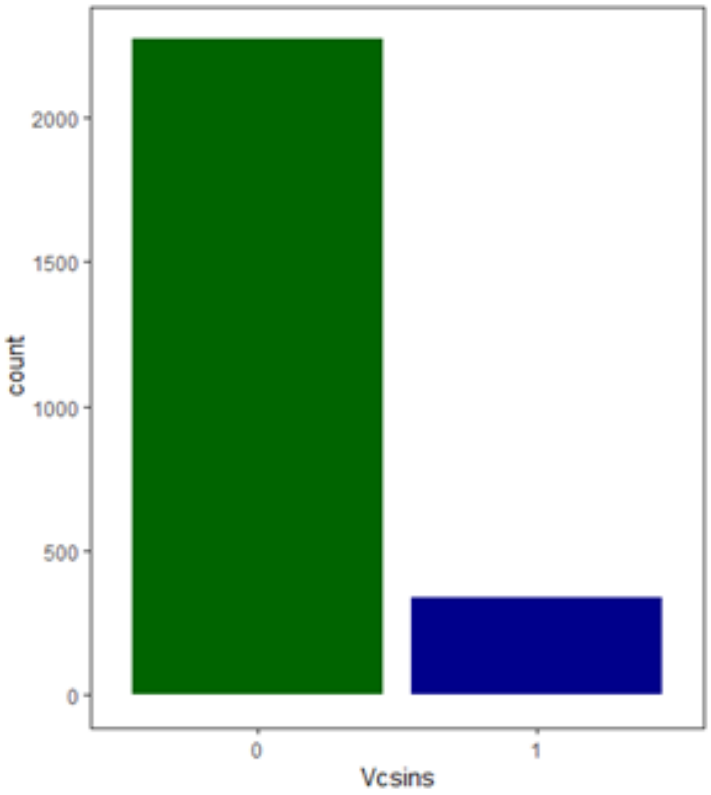


Figura 9 – Desbalanceamento entre as classes

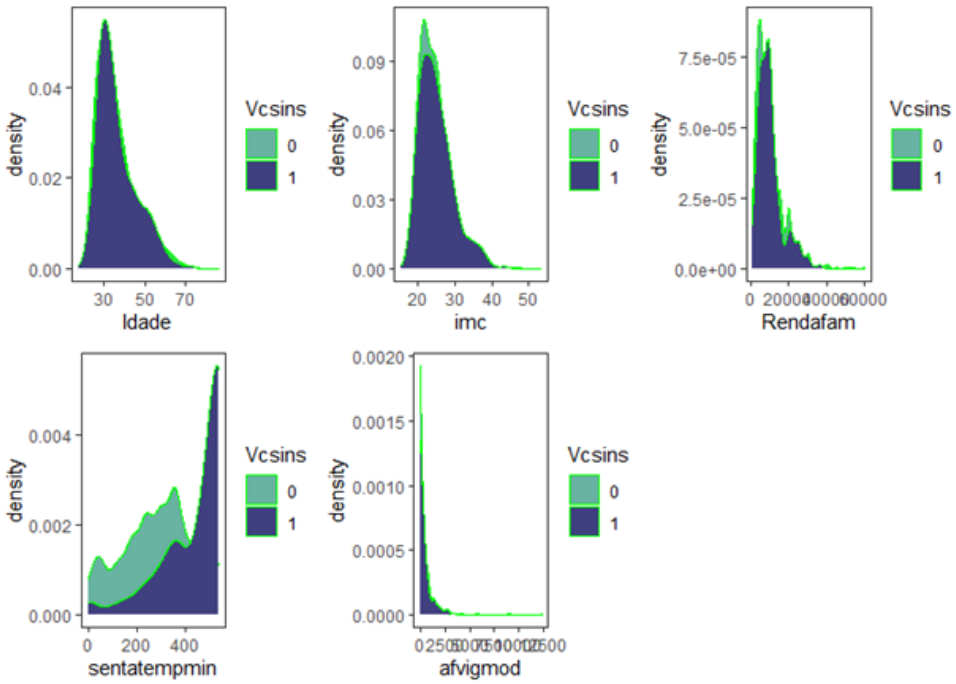


Figura 10 – Distribuição de frequência dos indivíduos

necessidade de um método de separação mais adequado que os classificadores tradicionais.

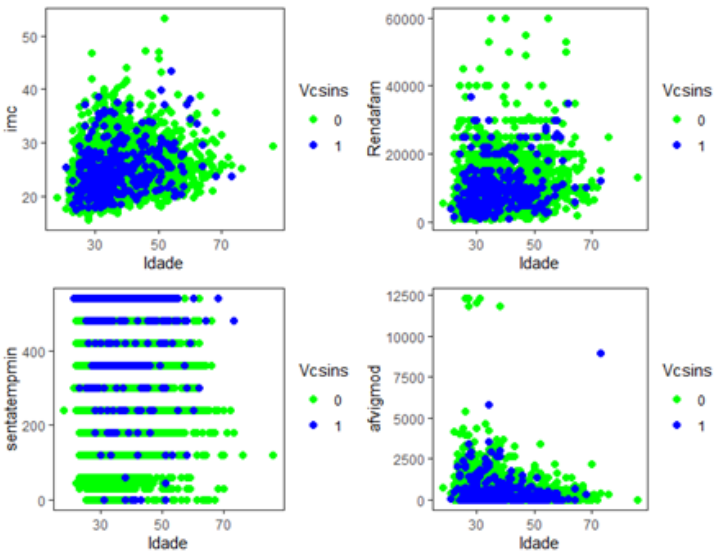


Figura 11 – Dispersão dos dados

4.2 Classificação

No treinamento do modelo, a classe alvo obteve 1589 observações para a classe de não sedentários e 238 para a classe de sedentários, conforme a partição do conjunto de dados para a etapa de aprendizado do modelo.

Para o conjunto de dados de teste o modelo contou com 680 indivíduos não sedentários e 102 sedentários.

Após treinamento e teste do modelo, avaliou-se através das métricas as respostas de classificação geradas pelo modelo, contudo, fez-se necessária a testagem da eficiência do modelo, visto que os dados são desbalanceados, criando viés indesejado.

4.3 Avaliação do modelo

Para a avaliação de um modelo existe uma grande quantidade de métricas, cada qual avalia o modelo de uma forma. Apresenta-se algumas, com base em estudos feitos. Inicialmente, apresenta-se o modelo de Naïve Bayes.

Através da matriz de confusão, foi feita a avaliação da proporção de testes do conjunto de dados, fazendo a comparação da classe verdadeira do dado com a classe predita pelo classificador gerado. Os valores gerados são as bases para as próximas métricas:

Observa-se pela matriz de confusão que, dos 680 indivíduos não sedentários, 668 foram classificados em sua classe correta e 12 indivíduos não sedentários restantes foram classificados incorretamente como sedentários. Dos 102 indivíduos sedentários, 5 foram classificados corretamente como sedentários e 97 indivíduos restantes do grupo dos sedentários foram classificados incorretamente como não sedentários.

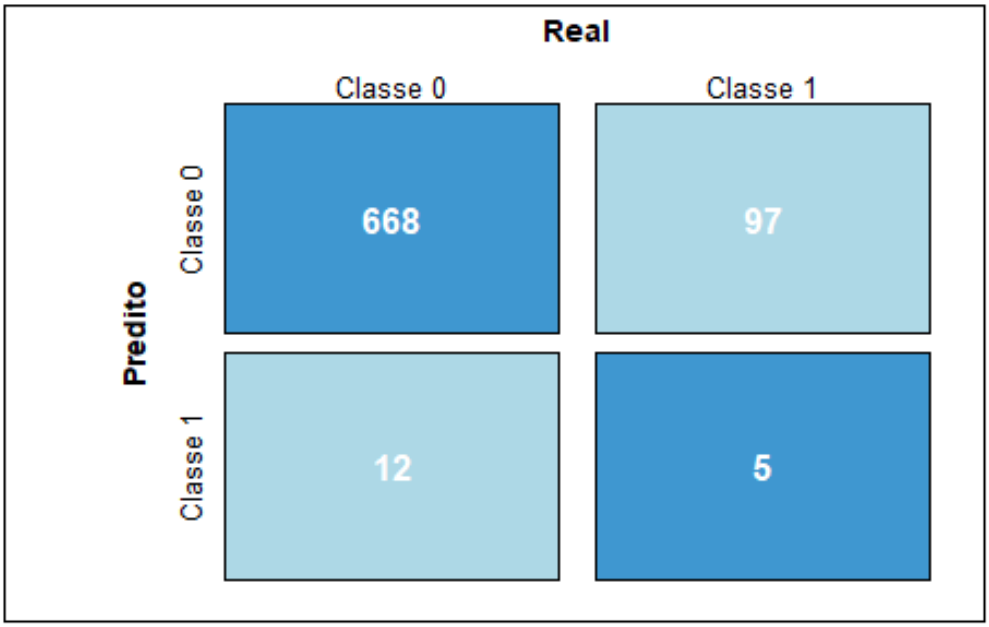


Figura 12 – Matriz de confusão Naïve Bayes

A classificação apresentou acurácia de 86,06 %, com isso pode-se afirmar que a proporção de casos que foram corretamente previstos é de 86,06 %.

A precisão de 98,23 % indica que o modelo elencou corretamente aqueles classificados como positivos, que foi a proporção na qual o modelo pode indicar os que foram classificados em não sedentários e realmente pertencem a essa classe.

Com o *recall* de 87,32 % pode-se afirmar que, da proporção de casos positivos que foram identificados corretamente, dos que eram não sedentários o modelo acertou em uma proporção de 87,32 %.

Através da especificidade pode-se afirmar que, a proporção de casos negativos que foram identificados corretamente é de 29,41 %, sendo essa a proporção de indivíduos classificados corretamente como sedentários.

O Naïve Bayes é considerado um classificador bom, por ser simples e ter um bom tempo de resposta, apresentando boas métricas e de modo geral, bons resultados avaliativos. Contudo, apresenta ineficiência para a classe de interesse, sendo influenciado pela sobreposição de classes, e resultando em uma quantidade grande de falsos positivos, prejudicando a classificação dos indivíduos sedentários.

4.4 Método AdaBoost

Foi aplicado o classificador de adaptação, Adabosst, para que através do conjunto de treinamento nos modelos fracos se chegasse em um classificador forte para o conjunto de dados desbalanceados.

Através dos resultados obtidos nas métricas, avaliou-se as respostas de classificação geradas pelo modelo, contudo, pode-se observar o seguinte comportamento preditivo do modelo para as classes desproporcionais.

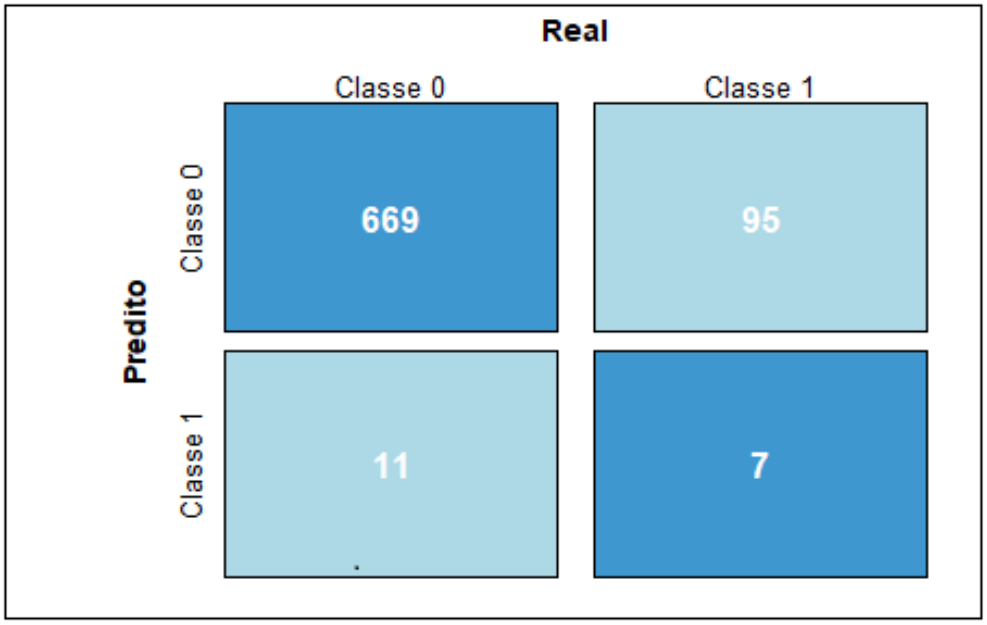


Figura 13 – Matriz de confusão do método Adaboost

Observou-se que, dos 680 indivíduos não sedentários, 669 foram classificados corretamente 11 indivíduos não sedentários restantes foram classificados incorretamente como sedentários. Dos 102 indivíduos sedentários, 7 foram classificados corretamente como sedentários e 95 indivíduos restantes do grupo dos sedentários foram classificados incorretamente como não sedentários.

A classificação apresentou acurácia de 86,44 %, com isso pode-se afirmar que, a proporção de casos que foram corretamente previstos é de 86,44 %.

A precisão de 98,38 % do modelo, acertando corretamente aqueles classificados como positivos, define a proporção a qual o modelo indicando que foram os classificados em não sedentários, realmente são pertencentes a essa classe.

Com o *recall* de 87,56 % pode-se afirmar que, da proporção de casos positivos que foram identificados corretamente, dos que eram não sedentários o modelo acertou em uma proporção

de 87,56 %.

Através da especificidade pode-se afirmar que, a proporção de casos negativos que foram identificados corretamente é de 38,89 %, sendo essa a proporção de indivíduos classificados corretamente como sedentários.

O método Adaboost apresentou boas métricas e de modo geral, bons resultados avaliativos até quando comparados com um classificador forte, como exemplo o de Naïve Bayes, que apresentou resultados semelhantes. Contudo, ele apresenta ineficiência para a classe de interesse, ainda que os pesos atribuídos resultem em uma melhor classificação da classe de indivíduos sedentários, a quantidade de acertos para a classe de interesse ainda é insatisfatória.

4.5 Método SMOTE

Aplicando o método do SMOTE foi possível gerar dados sintéticos para a classe minoritária, a classe de interesse, a partir de observações já existentes, deixando as classes equilibradas, para uma boa predição do modelo.

Para o conjunto de dados de teste o modelo classificou 680 indivíduos não sedentários e 680 sedentários.

A avaliação foi feita na proporção de acertos do conjunto de dados de testes, fazendo a comparação da classe real do dado com a classe predita pelo classificador.

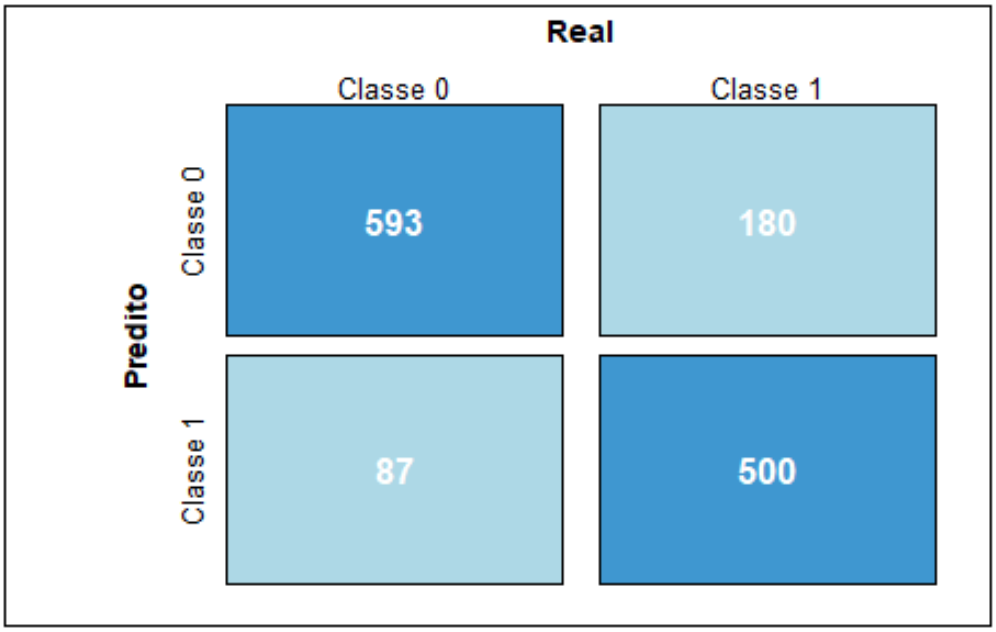


Figura 14 – Matriz de confusão do método SMOTE

Observa-se que, dos 680 indivíduos não sedentários, 593 foram classificados à sua classe

correta e 87 indivíduos não sedentários restantes foram classificados incorretamente como sedentários. Dos 680 indivíduos sedentários, 500 foram classificados corretamente como sedentários e 180 indivíduos restantes do grupo dos sedentários foram classificados incorretamente como não sedentários.

A classificação apresentou acurácia de 80,37 %, com isso, pode-se afirmar que a proporção de casos que foi corretamente prevista é de 80,37 %.

A precisão 76,71 % indica que o modelo acertou corretamente aqueles classificados como positivos, proporção a qual o modelo explicita os que foram classificados em não sedentários e realmente eram pertencentes a essa classe.

Com o *recall* de 87,20 % pode-se afirmar a proporção de casos positivos que foram identificados corretamente, logo, observa-se que dos que eram não sedentários o modelo acertou em uma proporção de 87,20 %.

Por meio da especificidade, afirma-se que a proporção de casos negativos que foram identificados corretamente é de 73,53 %, sendo essa a proporção de indivíduos classificados corretamente como sedentários.

Por meio da classificação do modelo Naïve Bayes a partir dos dados gerados pelo método SMOTE, boas métricas foram geradas e bons resultados avaliativos. Contudo, identifica-se ineficiência em algumas métricas, tal como a precisão, por exemplo, que quando comparada com os demais métodos utilizados e que podem ser observados na Tabela 1. Apesar disso, para a classe de interesse foi satisfatória e mostrou-se eficiente para o problema proposto.

Na Tabela 1, pode ser observado, de maneira geral, o desempenho de cada método para problema de desequilibro das classes. Através da comparação dos resultados, o modelo que mais se adequou ao objetivo de prever a presença do comportamento sedentário dos indivíduos foi o métodoSMOTE.

Tabela 1 – Resultado das métricas de avaliação dos modelos

Métrica	Naïve Bayes	Adaboost	SMOTE
Acurácia	86 %	86 %	80 %
Precisão	98 %	98 %	77 %
Recall	87 %	88 %	87 %
F1-score	92 %	93 %	82 %
Especificidade	29 %	39 %	74 %

Na sequência são apresentadas as considerações finais da pesquisa. Para além de sintetizar o que foi exposto, busca abrir novas janelas de pesquisa associadas ao tema.

5 Considerações Finais

Visto o grande volume de dados armazenados e a necessidade de ferramentas computacionais para auxiliar nas análises, este trabalho apresentou o conceito de aprendizado de máquina e a aplicação, pela tarefa de classificação, abordagem muito frequente em conjuntos de dados desbalanceados.

Diante do problema de classificação binária, aplicado em um conjunto de dados reais com as classes desequilibradas, apresentou-se sobreposição dos dados que dificultou a generalização. Neste trabalho, foram propostos três métodos diferentes, sendo eles, Naïve Bayes, classificador relativamente forte e simples, Adaboost, método de reforço de algoritmo considerado apropriado para conjunto de dados desequilibrados, e SMOTE, técnica de sobreamostragem.

O SMOTE, através da replicação dos indivíduos já existentes com base na técnica k vizinho mais próximo, equilibrou as classes permitindo usar o classificador de Naive Bayes, tal qual, já utilizado na base para cenários diferentes, teve desempenho inferior para prever o comportamento não sedentário, o que pode ter ocorrido devido aos exemplos sintéticos criados que desconsideram a classe majoritária, possivelmente resultando em exemplos ambíguos para uma forte sobreposição para as classes. Contudo, a previsão foi satisfatória para o comportamento sedentário, acertando 74 % dos indivíduos sedentários, atendendo os objetivos propostos no início da pesquisa, e garantindo bons resultados avaliativos.

O método analisado e avaliado positivamente sugere alternativas na área da saúde como técnica pré-diagnóstica, não somente para o comportamento sedentário, mas também para doenças raras, por exemplo. O fato de ter obtido 75 % de acerto para classe minoritária elenca opção confiável para um pré-diagnostico.

A pesquisa demonstra o impacto na generalização das classes quando apresenta-se um desequilíbrio entre as classes, sendo necessário uma tratativa diferente dos métodos tradicionais. Entretanto, há aspectos a serem otimizados, tal como o efeito negativo causado em exemplos ambíguos, que podem aparecer se a sobreposição for muito forte. Diante disso, o estudo viabiliza proposta para trabalhos futuros e replicações, configurando-se em tratativa eficiente em dados com forte sobreposição ou um outro método mais eficiente.

Referências

AINSWORTH, B. E. et al. Compendium of physical activities: an update of activity codes and met intensities. *Medicine and science in sports and exercise*, WILLIAMS & WILKINS, v. 32, n. 9; SUPP/1, p. S498–S504, 2000. 17

BATISTA, M. W.; ENUMO, S. R. F. Inclusão escolar e deficiência mental: análise da interação social entre companheiros. *Estudos de psicologia*, SciELO Brasil, v. 9, n. 1, p. 101–111, 2004. 12

BAUMAN, A. et al. The descriptive epidemiology of sitting: a 20-country comparison using the international physical activity questionnaire (ipaq). *American journal of preventive medicine*, Elsevier, v. 41, n. 2, p. 228–235, 2011. 17

BEERAVELLI, V. et al. An artificial neural network and taguchi integrated approach to the optimization of performance and emissions of direct injection diesel engine. *Eur J Sustain Dev Res*, v. 2, 2018. 8

BOSMAN, F. T. et al. *WHO classification of tumours of the digestive system.* [S.l.]: World Health Organization, 2010. 17

BURGER, S. V. *Introdução ao aprendizado de máquina com R: análise matemática rigorosa*. [S.l.]: "O'Reilly Media, Inc.", 2018. 7

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v. 6, n. 1, p. 1–6, 2004. 9, 11

DOMINGOS, A. L. G. et al. Cohort profile: the cohort of universities of minas gerais (cume). *International journal of epidemiology*, v. 1, p. 10, 2018. 17

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156. 10

HARRINGTON, P. *Machine learning in action*. [S.l.]: Manning Publications Co., 2012. 5, 7

JO, T.; JAPKOWICZ, N. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, ACM New York, NY, USA, v. 6, n. 1, p. 40–49, 2004. 11

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. Fundamentals of machine learning for predictive data analytics: algorithms. *Worked Examples, and Case Studies.*, The MIT Press, 2015. 7

KELLEHER J.D.; NAMEE, B. D. A. *Fundamentals of machine learning for predictive data analysis*. [S.l.]: Cambridge: The MIT Press, 2015. 7, 8

KLEMETTINEN, M.; MANNILA, H.; TOIVONEN, H. Rule discovery in telecommunication alarm data. *Journal of Network and Systems Management*, Springer, v. 7, n. 4, p. 395–423, 1999. 12

MARTÍNEZ-GONZÁLEZ, M. A. et al. Validation of the spanish version of the physical activity questionnaire used in the nurses' health study and the health professionals' follow-up study. *Public health nutrition*, Cambridge University Press, v. 8, n. 7, p. 920–927, 2005. 17

MONARD, M.; BATISTA, G. *Learning with skewed class distribution, Advances in Logic, Artificial Intelligence and Robotics*. [S.l.]: IOS Press, pp. 173-180., 2002. 9

MOREIRA, M. A. *Fundamentos do sensoriamento remoto e metodologias de aplicação*. [S.l.]: UFV, 2005. 5

OLIVEIRA, F. W. S.; PEREIRA, A. C. C. Elementos iniciais da relação entre o instrumento de pedro nunes, jacente no plano, e o cálculo da latitude no século xvi. *História da Ciência e Ensino: construindo interfaces*, v. 19, p. 39–53, 2019. 17

PRATI, R. C.; BATISTA, G. E.; MONARD, M. C. Learning with class skews and small disjuncts. In: SPRINGER. *Brazilian Symposium on Artificial Intelligence*. [S.l.], 2004. p. 296–306. 9

R, R. C. T. *A Language and Environment for Statistical Computing*. 2021. Vienna,Austria, 2021. Disponível em: <<https://www.r-project.org/>. 18

RAJARAMAN, V. Johnmccarthy—father of artificial intelligence. *Resonance*, Springer, v. 19, n. 3, p. 198–207, 2014. 4

WWW.PROJETOCUME.COM.BR. *Projeto CUME Transforming XML documents*. 2021. Abr, 2021. Disponível em: <<https://www.projetocume.com.br/>. 16

Anexos

ANEXO A – PARECER DA UFV

UNIVERSIDADE FEDERAL DE VIÇOSA - UFV

PARECER CONSUBSTANCIADO DO CEP

Elaborado pela Instituição Coparticipante

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: COORTE DAS UNIVERSIDADES MINEIRAS (CUME): IMPACTO DO PADRÃO ALIMENTAR BRASILEIRO E DA TRANSIÇÃO NUTRICIONAL SOBRE AS DOENÇAS E AGRAVOS NÃO TRANSMISSÍVEIS

Pesquisador: Adriano Marçal Pimenta

Área Temática:

Versão: 3

CAAE: 07223812.3.3001.5153

Instituição Proponente: UNIVERSIDADE FEDERAL DE MINAS GERAIS

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 596.741-0

Data da Relatoria: 18/01/2013

Apresentação do Projeto:

Trata-se de Protocolo de Pesquisa que analisa Coorte das Universidades Mineiras (CUME) Impacto do Padrão Alimentar Brasileiro e da Transição Nutricional sobre as doenças e Agravos não transmissíveis

Objetivo da Pesquisa:

Realizar análise comparativa com relação às Instituições Universidades Mineiras referentes ao padrão alimentar do Brasileiro

Avaliação dos Riscos e Benefícios:

Não há riscos para os indivíduos

Comentários e Considerações sobre a Pesquisa:

Trata-se de Protocolo de pesquisa relevante e que oferecere retorno para a sociedade

Considerações sobre os Termos de apresentação obrigatória:

Todos os documentos pertinentes ao Protocolo de Pesquisa foram apresentados.

Recomendações:

Recomenda-se a aprovação

Endereço: Universidade Federal de Viçosa, prédio Arthur Bernardes, piso inferior
Bairro: campus Viçosa **CEP:** 36.570-000
UF: MG **Município:** VICOSA
Telefone: (31)3899-2492 **Fax:** (31)3899-2492 **E-mail:** cep@ufv.br

ANEXO B – PARECER DA UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Plataforma Brasil

PARECER CONSUBSTANCIADO DO CEP

DADOS DA EMENDA

Título da Pesquisa: Coorte de Universidades MinEiras (CUME): impacto do padrão alimentar brasileiro e da transição nutricional sobre as doenças crônicas não transmissíveis - fase 2

Pesquisador: Adriano Marçal Pimenta
Área Temática:
Versão: 2
CAAE: 44483415.5.1001.5149
Instituição Proponente: UNIVERSIDADE FEDERAL DE MINAS GERAIS
Patrocinador Principal: FUNDACAO DE AMPARO A PESQUISA DO ESTADO DE MINAS GERAIS

DADOS DO PARECER

Número do Parecer: 2.491.366

Apresentação do Projeto:
Mesma apresentação que consta no parecer 1.137.860 de 03/07/2015.

Objetivo da Pesquisa:
Mesmos objetivos que constam no parecer 1.137.860 de 03/07/2015.

Avaliação dos Riscos e Benefícios:
Mesmos riscos e benefícios que constam no parecer 1.137.860 de 03/07/2015.

Comentários e Considerações sobre a Pesquisa:
Pesquisador solicita a inclusão de participantes da UFOP, UFLA e UFJF no projeto CUME.

Considerações sobre os Termos de apresentação obrigatória:
Foi acrescentado novo TCLE para a linha de base de inclusão e novo projeto adaptado à emenda.

Conclusões ou Pendências e Lista de Inadequações:
SMJ, sou favorável à aprovação da emenda.

Considerações Finais a critério do CEP:
Tendo em vista a legislação vigente (Resolução CNS 466/12), o COEP-UFMG recomenda aos Pesquisadores: comunicar toda e qualquer alteração do projeto e do termo de consentimento via emenda na Plataforma Brasil, informar imediatamente qualquer evento adverso ocorrido durante o

Endereço: Av. Presidente Antônio Carlos,6627 2º Ad SI 2005
Bairro: Unidade Administrativa II **CEP:** 31.270-901
UF: MG **Município:** BELO HORIZONTE
Telefone: (31)3409-4592 **E-mail:** coep@prpq.ufmg.br

ANEXO C – PARECER DA UFOP

UNIVERSIDADE FEDERAL DE OURO PRETO

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Coorte de Universidades MinEiras (CUME): impacto do padrão alimentar brasileiro e da transição nutricional sobre as doenças crônicas não transmissíveis - fase 2

Pesquisador: JULIA CRISTINA CARDOSO CARRARO
Área Temática:
Versão: 1
CAAE: 44483415.5.2003.5150
Instituição Proponente: Universidade Federal de Ouro Preto
Patrocinador Principal: FUNDACAO DE AMPARO A PESQUISA DO ESTADO DE MINAS GERAIS

DADOS DO PARECER

Número do Parecer: 2.565.240

Apresentação do Projeto:
"Tratar-se de um estudo epidemiológico, observacional, de delineamento transversal na sua linha de base e longitudinal no seu seguimento, que está sendo realizado, desde março de 2016, com indivíduos graduados na UFMG e na UFV a partir de janeiro de 1994. A principal característica deste estudo é o recrutamento permanentemente aberto, permitindo um contínuo crescimento da amostra a cada onda de seguimento, uma vez que ao mesmo tempo em que se aplica um novo questionário (Q_2, Q_4, ..., Q_n) a cada dois anos aos participantes recrutados previamente, envia-se o questionário da linha de base (Q_0) para o recrutamento de novos participantes. Neste sentido, o Q_0 será enviado a novos participantes da UFMG e da UFV e, a presente emenda, prevê o envio deste instrumento também a egressos da Universidade Federal de Ouro Preto (UFOP), Universidade Federal de Lavras (UFLA) e Universidade Federal de Juiz de Fora (UFJF), ampliando a amostra de participantes do projeto CUME."

Objetivo da Pesquisa:
"Objetivo Primário:
Avaliar o impacto do padrão alimentar brasileiro no desenvolvimento de DCNT em indivíduos graduados na Universidade Federal de Minas Gerais (UFMG), na Universidade Federal de Viçosa (UFV), na Universidade Federal de Ouro Preto (UFOP), na Universidade Federal de Lavras (UFLA) e na Universidade Federal de Juiz de Fora (UFJF).

Endereço: Morro do Cruzeiro-ICEB II, Sala 29 -PROPP/UFOP
Bairro: Campus Universitário **CEP:** 35.400-000
UF: MG **Município:** OURO PRETO
Telefone: (31)3559-1368 **Fax:** (31)3559-1370 **E-mail:** cep@propp.ufop.br

ANEXO D – PARECER DA UFJF

UFJF - HOSPITAL UNIVERSITÁRIO DA UNIVERSIDADE FEDERAL DE

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Coorte de Universidades MinEiras (CUME): impacto do padrão alimentar brasileiro e da transição nutricional sobre as doenças crônicas não transmissíveis - fase 2

Pesquisador: Ana Paula Boroni Moreira
Área Temática:
Versão: 1
CAAE: 44483415.5.2004.5133
Instituição Proponente: UNIVERSIDADE FEDERAL DE JUIZ DE FORA UFJF
Patrocinador Principal: FUNDACAO DE AMPARO A PESQUISA DO ESTADO DE MINAS GERAIS

DADOS DO PARECER

Número do Parecer: 2.015.738

Apresentação do Projeto:
Estudo epidemiológico de delineamento transversal na sua linha de base e longitudinal no seu seguimento, que está sendo realizado março de 2016 com egressos da UFMG e da UFV, cujo objetivo é avaliar o impacto do padrão alimentar brasileiro, de grupos de alimentos e fatores dietéticos específicos no desenvolvimento de Doenças Crônicas Não Transmissíveis (DCNT). A coleta de dados para a linha de base está sendo realizada por meio de questionário auto-respondido (Q_0), criado em ambiente virtual, procedimento que também será empregado para os questionários de seguimento que serão aplicados a cada dois anos, começando por aquele previsto na presente proposta (Q_2). Entre outubro de 2016 e fevereiro de 2017, foi realizado etapa de validação das variáveis que compõe os diagnósticos das DCNT, com a seleção de subamostra de participantes recrutados na linha de base que autodeclararam os valores das ditas variáveis e as mesmas foram, também, aferidas diretamente por entrevistadores treinados. Na linha de base, modelos de regressão de Poisson com variâncias robustas ou de regressão logística serão construídos para avaliar a associação entre o padrão alimentar brasileiro, de grupos de alimentos e fatores dietéticos específicos com as DCNT, ajustado por fatores de confusão. Na fase longitudinal, optar-se-á pela construção de modelos de regressão de Cox. Para a validação, foram utilizados testes
estatísticos próprios de comparação entre os valores autodeclarados e aqueles aferidos

Endereço: Rua Catulo Breviglieri, s/n
Bairro: Santa Catarina **CEP:** 36.036-110
UF: MG **Município:** JUIZ DE FORA
Telefone: (32)4009-5217 **E-mail:** cep.hu@ufjf.edu.br