18/10/2023 (Quarta-Feira)

Palestra de Abertura

Palestrante: Mine Çetinkaya-Rundel (Duke University, USA)
Título: Data Science in a(n Ever-Evolving) Box
Resumo: What should a first course in data science for students who have limited to no experience with statistics and programming look like? How do we teach it in a way that lends itself to iteration as the landscape of data science evolves and that scales to more students and more instructors? In this talk I will aim to accomplish two goals to answer these questions: (1) Introduce a semester-long, modern introductory data science curriculum, along with its design philosophy, implementation details (particularly as class sizes increase), technical infrastructure, and real examples from course content as well as from student projects. (2) Discuss how I've open-sourced this curriculum at datasciencebox.org for sharing with and re-use / adaptation by other instructors and what it takes to maintain this open-source project as the landscape of data science, data science education curriculum guidelines, and data science tooling evolves.

Palestra 1

Palestrante: João Dorea (University of Wisconsin-Madison, USA)
Título: Machine Learning and Computer Vision to Optimize Farm Management Decisions
Resumo: The advance of AI systems in different fields of science has created incredible opportunities for the new generation of students and scientists to answer research questions that would not otherwise be possible before the recent progress towards more intelligent systems. AI technology such as computer vision, natural language processing, and robotics has become a real component of our lives through well-known applications as face recognition, speech-to-text, robotics, and virtual reality. The area of agriculture has leveraged the AI development by other scientific domains, and livestock systems have gradually experienced the implementation of modern solutions to solve critical problems related to animal monitoring systems for health and welfare, greenhouses gas emissions, animal traceability, and labor shortage. In this talk, we will discuss some examples of AI technologies with potential to revolutionize livestock systems in the next decades, such as computer vision systems, robotics, and mixed reality. We will discuss how these AI examples relate to real-world challenges currently faced by farmers, industry, and the scientific community.

Palestra 2

Palestrante: Guilherme J.M. Rosa (University of Wisconsin-Madison, USA)
Título: Regression and Classification Applied to Precision Agriculture
Resumo: Vast amounts of data are routinely collected in agriculture, encompassing traditional operational farm data and more recent data acquired through digital tools, such as remote and on-site sensing technologies. Furthermore, numerous additional sources of information can be integrated with farm data, including economic and weather variables. The integration and analysis of such data can yield essential insights and data-driven decision tools to optimize agricultural systems. Nevertheless, handling such large and intricate observational datasets, which involve multicollinearities, redundancies, nonlinear

relationships, as well as spatial and temporal dependencies, necessitates suitable statistical and data mining tools. In this presentation, we will delve into various regression and classification techniques tailored for prediction and causal inference, with a special focus on their applications in agriculture. Specific modeling approaches will encompass generalized additive models, structural equation models, and mixed-effect models. We will explore valuable algorithms and fitting strategies, including dimension-reduction techniques, regularization methods like penalized regression and Bayesian hierarchical approaches, as well as cross-validation strategies for variable selection and model comparison. To illustrate these methods and applications, we will showcase examples utilizing infrared spectroscopy data, satellite remote sensing, image analysis, and computer vision, among other relevant domains.

Palestra 3

Palestrante: Tiago A. Marques (University of St Andrews, UK; Universidade de Lisboa, PT)
Título: How to obtain ACCURATE estimates of wildlife abundance from passive acoustics data
Resumo: Traditionally estimation of wildlife abundance has been achieved via a wide variety of approaches which in general relied on visual detection of the animals of interest. Recent technological developments are changing the way we monitor wildlife. In this talk we introduce the topic with a focus on passive acoustic monitoring (PAM) density estimation (DE), and with a strong (conscious) bias towards marine mammals. For cetacean species, which are visually cryptic and spend long periods submerged, making them hard to survey by conventional visual methods, PAM has become routine. A possible approach is cue counting, when a sample of detected sounds produced by the animals are used to estimate density. Among other multipliers related to detector performance, including false negatives and false positives, a cue production rate is required to convert the number of sounds detected into animal density. The cue production rate corresponds to the average number of sounds produced per animal per unit time. Cue counting PAM based reliable estimates of animal density therefore depend upon using accurate and unbiased estimates of the cue rate, i.e., cue rates that correspond to the time and place the survey took place. The ACCURATE project, funded by the US Navy Living Marine Resources program, is a St Andrews-led large project involving several partners, looking at everything related to cue production by marine mammals to inform PAM DE exercises. We describe advances obtained within the project. Building on PAM DE, I will present the project objectives, some of the main results to date and discuss factors that affect cue rates and report on cue rate estimates for some of the species we have been working on, showcasing the methods we have used and some of the research threads we are pursuing. This will include thoughts about what is the best sampling unit to estimate cue rates, challenges with caller identification and conspecific interferences, getting cue rates from tags without acoustics and the potential for site-specific cue rates.
This research was conducted under the ACCURATE project, funded by the US Navy Living Marine Resources program (contract no. N3943019C2176). TAM thanks partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020).

Palestra 4

Palestrante: Pedro A. Morettin (Universidade de São Paulo, Brasil)
Título: Estatísticos e Cientistas de Dados: Almas Gêmeas?
Resumo: Nesta palestra iremos abordar temas como Inteligência Artificial, Aprendizado de Máquina, Aprendizado Estatístico e Ciência de Dados. Também discutiremos sugestões para um currículo de graduação e pós-graduação em Ciência de Dados e se existem diferenças, ou não, entre Cientista de Dados e Estatísticos.

Palestra 5

Palestrante: Gloria Icaza (SEREMI de Salud del Maule, Chile)
Título: La desigualdad mata: cómo los datos nos cambian la vida
Resumo: En esta charla mostraré cómo ha sido el proceso de pasar de ser una tranquila investigadora universitaria a una autoridad regional de salud del gobierno del Presidente Gabriel Boric.
Como bioestadísticos aprendemos a interpretar los datos, siguiendo el método científico, siendo cautos al momento de sacar conclusiones. Por décadas, me he dedicado al análisis epidemiológico de la mortalidad por enfermedades crónicas. Estos análisis muestran evidencia de desigualdades en Chile.

Durante el estallido social de octubre de 2019, sentimos la urgencia de mostrar estos resultados con mayor fuerza que nuestra acostumbrada cautela científica, ya que es evidente que “La desigualdad mata*”. Esto se evidencia en mapas comunales de mortalidad por cáncer gástrico y de vesícula y de mortalidad por cirrosis en Chile.

*La desigualdad mata, es el título, en español, del libro de Goran Therborn, Alianza Editorial, 2015

19/10/2023 (Quinta-Feira)

Tutorial 1

Ministrante: Antonio Mendes Magalhães Júnior (PPGEE/UFLA)
Título: Introdução às Redes Neurais Artificiais com R
Resumo: Neste tutorial, os participantes conhecerão a história das RNAs, explorando seus diferentes tipos e mecanismos de funcionamento, com ênfase nas redes do tipo Perceptron Multicamadas (MLP) de aprendizado supervisionado. Será discutida a importância e aplicação de comitês de máquinas, bem como as métricas essenciais para avaliar sistemas de classificação e regressão. Para consolidar o aprendizado, diferentes problemas práticos serão apresentados e os participantes terão a oportunidade de aplicar os conceitos aprendidos utilizando o software R.

Ao final do minicurso, os participantes terão adquirido uma compreensão dos fundamentos das RNAs e estarão preparados para resolver problemas práticos utilizando RNAs.

Tópicos que serão abordados:
- Contextualização histórica das RNAs
- Fundamentos e estruturação das RNAs (com foco em MLP)
- Introdução aos comitês de máquinas
- Métricas de avaliação para sistemas de classificação e regressão
- Estudo de caso prático em R.

Tutorial 2

Ministrante: Luiz Otávio de Oliveira Pala (DES/UFLA)
Título: Modelos para séries temporais de contagem
Resumo: A análise de séries temporais a partir da classe de modelos autorregressivos e de médias móveis, ARMA(p, q), é amplamente adotada em estudos aplicados. No entanto, séries temporais de contagem necessitam de certa atenção pois podem apresentar características estilizadas, como alta dispersão e excesso de zeros, que devem ser levadas em conta pelo pesquisador. Ampliações da classe ARMA(p, q) para a modelagem de séries de contagem têm sido propostas na literatura estatística e talvez a mais difundida seja a classe de modelos autorregressivos e de médias móveis generalizados, GARMA(p, q), que será apresentada neste tutorial. Adicionalmente, exemplos serão apresentados no programa R para o caso Poisson e Binomial negativo.

Palestra 6

Palestrante: Denilson Alves Pereira (DCC/UFLA)
Título: Modelos de Linguagem para Processamento de Linguagem Natural
Resumo: Processamento de Linguagem Natural (PLN) é um campo da Linguística e da Aprendizagem de Máquina focado no entendimento da linguagem humana. Tem como objetivo entender o contexto das palavras. O estado-da-arte em PLN está focado no desenvolvimento de modelos de linguagem gerados por Redes Neurais Artificiais. Esta palestra apresenta os conceitos e fundamentos sobre essa tecnologia, bem como exemplos de aplicações. Exemplos de modelos de linguagem são o GPT, o BERT e o BARD e suas

ferramentas, como o ChatGPT e o Copilot. Exemplos de aplicações são: busca semântica em máquinas de busca, extração de respostas para perguntas, chatbots, classificação de documentos, reconhecimento de entidades nomeadas, tradução de textos, sumarização de documentos e sistemas de recomendação.

Palestra 7

Palestrante: Ana Claudia Festucci de Herval (ICTIN/UFLA)
Título: Uma Introdução à Análise de Dados do Mercado Acionário Brasileiro
Resumo: Para se trabalhar com dados do mercado acionário, é fundamental compreender que sua natureza influencia a abordagem necessária para sua interpretação. Nesta palestra abordaremos as peculiaridades das séries temporais no mercado financeiro, incluindo horizontes de investimento de baixas frequências (diário, mensal, anual) e os de alta frequência, através de dados intradiários (coletados em intervalos de minutos ou horas). Além disso, vamos explorar uma das características mais marcantes: os saltos de volatilidade, muitas vezes desencadeados por anúncios de notícias macroeconômicas. Por fim, apresentaremos um método para detecção de saltos em dados reais para uma série de preços das ações da Petrobrás.

Mesa Redonda

Moderadora: Izabela R. Cardoso de Oliveira - DES/UFLA
Membros:

- Éric Fernandes de Mello Araújo - DCC/UFLA
- Helena Maria Ferreira - DEL/UFLA
- Viviane Costa Silva - PPGEE/UFLA

Tema: A inteligência artificial generativa na educação

20/10/2023 (Sexta-feira)

Minicurso 1

Ministrantes: Marluce Rodrigues Pereira (DAC/UFLA) e Carlos Pereira da Silva (Pós-Doutorando, DES/UFLA)
Título: Execução de aplicações em cluster de computadores
Resumo: A definir

Minicurso 2

Ministrante: Geraldo Magela da Cruz Pereira (DES/UFLA)
Título: Aprendizado de máquina para análise de sobrevivência
Resumo: A análise de sobrevivência é uma área da estatística utilizada para a investigação de dados nos quais a variável resposta é o tempo até a ocorrência de um evento de interesse, como, por exemplo, o falecimento ou a recuperação de um paciente, a falha de um equipamento eletrônico, o divórcio, a evasão de um aluno, etc. Em muitos casos, durante a realização de um estudo, algumas observações podem não experimentar o evento de interesse e são, portanto, classificadas como censuradas. Os conceitos de análise de sobrevivência são amplamente aplicados em áreas como Medicina e Engenharia, com a finalidade de estimar, respectivamente, o tempo necessário para a recuperação de pacientes e o período até que um componente apresente falhas. Além disso, a análise possibilita a avaliação dos fatores que influenciam no tempo de sobrevivência. Este minicurso tem como objetivo introduzir a análise de sobrevivência, abordando seus conceitos fundamentas, bem como explorar a utilização de métodos não paramétricos (Estimador de Kaplan-Meier), semiparamétricos (regressão de Cox e regressão penalizada de Cox) e técnicas de aprendizado de máquina (Random Survival Forest, Gradient Boosted Models e Survival Support Vector Machine) para a análise de sobrevida de pacientes e de componentes eletrônicos, fazendo uso das bibliotecas survival, fastcox, gbm e randomForestSRC do software R.

Comunicações orais

Comunicação Oral 1

Ministrante: Dulcídia Carlos G. Ernesto
Título: Transformada de wavelet discreta não decimada aplicada na análise de similaridades de genomas de vírus das famílias Coronaviridae e Paramyxoviridae
Resumo: A implementação de métodos estatísticos que dão suporte para tornar o processo de estudo de similaridades mais eficiente, tem sido muito recorrente nas pesquisas atuais, pois alguns métodos como as wavelets tornam o processamento de manipulação de dados mais rápido, e com maior capacidade para processar milhares de nucleotídeos de um sequenciamento genômico. Portanto, melhores estimativas podem ser obtidas. Fez-se inicialmente a extração das sequências do genoma completo das famílias pelo site do Centro Nacional de Biotecnologia e Informação (NCBI) e, em seguida analisou-se os padrões usando um método de processamento de sinais, com base na proporção de Guanina e Citosina (conteúdo GC). Para cada sequência foi feita a decomposição usando a transformada de wavelet não decimada discreta de Daubechies em seis níveis. A seguir calculou-se o expoente de Hurst para cada nível de decomposição e verificou-se os

sequenciamentos com padrões similares. Após a análise de agrupamento pôde-se ver que o par Betacoronavírus Inglaterra e Coronavírus do Médio Oriente mostrou-se similar em quase todos os métodos. O outro par de sequências que se mostrou similar pelos métodos de momentos absolutos e pelo método de variância agregada foi o par Pato coronavírus e Deltacoronavírus. O grupo composto pelos vírus Parainfluenza 1, 3, 4 e o Hendrahenipavírus apresentou-se similar pelos métodos de momentos absolutos e variância agregada. Já no método de análise R/S, ocorreu uma substituição do Parainfluenza4 pelo Parainfluenza5. Ao fim do estudo pôde-se concluir que a transformada discreta não decimada de wavelets nos permitiu decompor cada uma das sequências e por consequência, pôde-se fazer um estudo mais detalhado de similaridade entre os sequenciamentos genéticos de vírus dessas duas famílias, Coronaviridae e Paramyxoviridae. Trabalho em parceria com: Leila Maria Ferreira e Thelma Sáfadi.

Comunicação Oral 2

Palestrante: Momate Emate Ossifo

Título: Análise da produtividade do milho utilizando uma abordagem GAMLSS

Resumo: O milho (Zea mays L.) é um dos principais cereais cultivados em todo o mundo para alimentação humana e animal, apresentando um grande potencial de produção, cuja produção é limitada por vários fatores incluindo a utilização de variedades de baixo rendimento e o sistema de produção em sequeiro. Assim, neste estudo, objetivou-se analisar a produtividade do milho, relacionando-a com as variáveis específicas, por meio dos GAMLSS. Foram utilizadas informações referentes a 102 plantas de milho, que são parte da coleção permanente do Centro de Desenvolvimento Científico e Tecnológico para a Agricultura da Universidade Federal de Lavras. Para tal, foram consideradas as seguintes variáveis explicativas candidatas: altura da planta (AP, cm), altura da espiga (AE, cm), dias para a maturação (DPM), dias para o florescimento masculino (DFM) e dias para o florescimento feminino (DFF). Para a modelagem, foram utilizados os modelos aditivos generalizados para locação, escala e forma (GAMLSS) por conta de sua flexibilidade para explicar a variável resposta (PROD: Produtividade, kg ha -1). A distribuição escolhida para representar a resposta foi a Box-Cox exponencial potência, por ser capaz de modelar variáveis que assumem valores positivos e apresentam diferentes graus de curtose. Para o processo de seleção de covariáveis em cada um dos parâmetros da distribuição, foi utilizada a chamada Estratégia A, baseada no stepwise. As funções de suavização foram necessárias para modelar a mediana para AP e DFF e a curtose para AP. Além dessas variáveis utilizadas para explicar a variável resposta, foram detectadas ainda efeitos de DFM para mediana, AE para coeficiente de variação e DFF para assimetria, estatisticamente significativos (valor-p<0,05). Com base nos resíduos obtidos a partir do modelo final verificou-se que ele é adequado para explicar o conjunto de dados. Trabalho em parceria com: Joaquina da Márcia Jaime Muchico, Rafaela de Carvalho Salvador, Cesar Pedro, Luiz Ricardo Nakamura, João Cândido de Souza, Daniel Furtado Ferreira, Alex de Oliveira Ribeiro.

Comunicação Oral 3

Palestrante: Jéssica Gracielle Silva

Título: Diagramas de Hasse e delineamentos experimentais

Resumo: Delineamentos experimentais são totalmente descritos utilizando a teoria dos conjuntos. Uma ordem parcial, em teoria dos conjuntos, é dada pela relação de inclusão, ou seja, quando tem-se subconjuntos de conjuntos. Tal ordem é totalmente representada por

um recurso denominado diagrama de Hasse. Em situações experimentais a ordem parcial aparece naturalmente, visto que, em um experimento é possível ter canteiros que por sua vez estão contidos em casas de vegetação e tem-se a inclusão. Em teoria dos conjuntos podemos também definir partições. Uma partição de um conjunto Ω é a união disjunta de subconjuntos não vazios. Dadas duas partições F e G de Ω, tem-se a partição ínfimo $F \wedge G$ e tem-se também a partição supremo $F \vee G$.
Um conjunto de partições, $\mathfrak{F}$, é um conjunto parcialmente ordenado pela relação de inclusão das classes que pode também ser descrito pelo diagrama de Hasse. Este trabalho tem por objetivo aplicar a teoria dos conjuntos bem como as partições naturais existentes em situações experimentais na obtenção do esquema de análise de variância de um delineamento. Nesse sentido, o diagrama de Hasse definido por $\mathfrak{F}$ nos permite obter cada um dos subespaços V_F como soma direta de subespaços W_G com $F \preceq G $. Para relacionar as partições nos conjuntos de parcelas e as partições nos conjuntos de tratamentos, utiliza-se à função que aloca os tratamentos as parcelas. $\psi: \Omega \rightarrow \Theta$. Utiliza-se também a notação $\psi^{-1}(A) = \lbrace \omega \in \Omega;~ \psi(\omega) \in A \rbrace$, isto é, a pré-imagem de A. Assim, o diagrama de Hasse combinado (para tratamentos e parcelas) nos permite obter o esquema de análise de variância do delineamento. Sendo assim, os diagramas de Hasse e o esquema de análise de variância para um exemplo prático será apresentado. Trabalho em parceria com: Lucas Monteiro Chaves e Renato Ribeiro Lima.

Comunicação Oral 4

Palestrante: Edilene Cristina Pedroso Azarias

Título: Avaliação de modelos não lineares no estudo do crescimento de plantas daninhas e do efeito de herbicidas

Resumo: Plantas daninhas, frequentemente chamadas de "voluntárias" ou "infestantes", são, em sua maioria, indesejáveis e crescem junto a outras culturas agrícolas. Elas interferem nas atividades produtivas reduzindo tanto a produção quanto a sua qualidade, hospedam pragas e doenças, dificultam a colheita aumentando o custo de produção. Competem por recursos do meio como água, luz, espaço, gás carbônico e nutrientes, sendo, em geral, mais eficientes na captação desses elementos em comparação aos cultivos. Os herbicidas desempenham um papel fundamental no controle de plantas daninhas das culturas, viabilizando os cultivos de grandes áreas, por demandarem menos mão de obra, apresentando maior rapidez e menor custo em comparação aos outros métodos de controle. No entanto, nos plantios comerciais é essencial usá-los adequadamente, pois seu uso incorreto pode resultar em resíduos persistentes no solo e contaminação do ambiente, tornando essencial o entendimento de como esses produtos afetam as plantas e a determinação da dosagem adequada a ser utilizada em diferentes condições. Modelos não lineares são utilizados em diversas áreas, tais como: Computação, Administração, Engenharias, Biologia, Agronomia, Medicina e Saúde, Farmacologia, Sociologia, entre outras. Há vários modelos não lineares propostos na literatura e são utilizados de acordo com o comportamento dos dados. Em estudos de crescimento os mais usados são o Logístico, Gompertz, von Bertalanffy, Brody, Richards e Weibull. Em Agronomia, na descrição da curva dose-resposta de herbicidas, geralmente são utilizados os de Michaelis Menten, Brody, Weibull e derivações do modelo Logístico, propostas por Streibig, da qual foram obtidas diversas outras parametrizações. Os parâmetros, em geral,

são estimados pelo método de mínimos quadrados, utilizando o algoritmo de Gauss-Newton, pelo software R. Na estimação dos parâmetros para que a inferência seja adequada, é importante realizar a análise de resíduos. Esta análise permite verificar as pressuposições sobre o vetor de erros, ou seja, se são independentes, identicamente distribuídos com distribuição normal com média zero e variância constante, por meio dos testes de Shapiro-Wilk, Durbin-Watson e Breusch-Pagan, respectivamente. Como avaliadores de qualidade de ajuste podem ser utilizados o coeficiente de determinação R2, desvio padrão residual (DPR) e critério de informação de Akaike (AIC) e curvaturas de Bates e Watts. Também podem ser analisados por meio de derivações do modelo ajustado os pontos críticos: ponto de aceleração máxima (PAM), ponto de inflexão (PI), ponto de desaceleração máxima (PDM) e ponto de desaceleração assintótico (PDA). Nesse sentido, o objetivo desse trabalho foi apresentar o uso de modelos não lineares para o estudo do crescimento de plantas daninhas e a resposta a herbicidas. Os dados utilizados no estudo referem-se à eficácia dos herbicidas trifloxysulfuron-sodium e chlorimuron-ethyl e o crescimento de plantas do gênero Amaranthus (caruru). Os modelos Logístico e Gompertz foram aplicados aos dados de crescimento de caruru, apresentando bons ajustes aos dados. Para analisar a porcentagem de controle das plantas daninhas pelos herbicidas, o modelo mais adequado foi o de Michaelis Menten. Os resultados indicaram que a espécie A. deflexus acumulou menos massa seca total (1,4295) e nas raízes (25,5280) g planta−1, em comparação com a A. hybridus que acumulou mais massa seca total (10,0151) e nas raízes (57,1645) g planta−1. Além disso, a dose que proporcionou metade de controle ou redução do crescimento da A. hybridus e A. deflexus para o herbicida trifloxysulfuron-sodium foi de 0,41 e 1,55 g ha−1 e para o chlorimuron-ethyl de 2,50 e 8,97 g ha−1, evidenciando diferenças na sensibilidade das espécies. Trabalho em parceria com: Edilson Marcelino Silva, Joel Augusto Muniz.

Comunicação Oral 5

Palestrante: Daiane de Oliveira Gonçalves

Título: Proposição do uso de probabilidades de cobertura cruzadas como instrumento de diagnóstico de ajuste de modelos conjuntos para dados de sobrevivência e longitudinais

Resumo: Estudos relacionados a características de fenômenos/experimentos no tempo, como estudos longitudinais ou do tempo até a ocorrência de um evento de interesse, se fazem cada vez mais necessários em diversas áreas. Estes dois fenômenos podem ser tratados, respectivamente, por meio dos modelos lineares mistos e modelos de sobrevivência. No entanto, podem existir situações em que se objetiva investigar a relação entre uma ou mais respostas longitudinais e um evento de interesse, que pode ser realizada com o auxílio da modelagem conjunta de dados longitudinais e de sobrevivência. Entretanto, esses modelos podem apresentar problemas de convergência e serem computacionalmente exigentes, tornando inviável a utilização dos mesmos em muitos casos. Neste sentido, este estudo objetiva realizar, por meio de um estudo de simulação monte carlo, uma comparação entre modelos longitudinais e de sobrevivência, em função de diferentes porcentagens de censura e estrutura de covariância das medidas repetidas. Por meio dos resultados, verificou-se a existência de similaridade em termos de inferência entre os modelos. Em outras palavras, pode-se dizer que a escolha entre o modelo longitudinal e de sobrevivência para análise dos dados com respostas longitudinais e evento de interesse se torna arbitrária, principalmente quando o percentual de censura for de 15%. Trabalho em parceria com: Natália da Silva Martins Fonseca e Marcelo Angelo Cirillo.

Comunicação Oral 6

Palestrante: Iuri dos Santos Manoel

Título: Método robusto Jackknife no ajuste do modelo não linear Logístico

Resumo: Em ajustes de curvas de crescimento, a utilização de modelos não lineares é difundida entre os pesquisadores, destacando-se o modelo Logístico por se ajustar bem aos dados e apresentar parâmetros interpretáveis. No entanto, é necessário considerar sua sensibilidade quanto a valores discrepantes (outliers), uma vez que estes são oriundos de fatores aleatórios não controláveis. Isso pode implicar em amostras que não contêm informações que representam a população em estudo, prejudicando o ajuste de modelos. Frente a este problema, os métodos robustos são uma possível solução para melhorar o ajuste de modelos ao lidar com outliers no conjunto de dados. Neste trabalho, é apresentado o método robusto de reamostragem Jackknife como uma abordagem robusta para melhorar ajustes do modelo não linear Logístico na descrição de dados com valores discrepantes. O procedimento foi implementado no software estatístico R ao utilizar o método de mínimos quadrados para as convergências das estimativas do parâmetro do modelo Logístico. O conjunto de dados foi simulado via Simulação de Monte Carlo. A implementação do método robusto de reamostragem Jackknife ao modelo não linear Logístico apresentou melhoria no ajuste, de acordo com o desvio médio absoluto (DMA).

Comunicação Oral 7

Palestrante: Mateus Santos Peixoto

Título: Utilização de bibliometria e revisão sistemática com técnicas de aprendizado de máquina para avaliar a relevância da literatura científica

Resumo: A bibliometria é uma área que objetiva analisar informações e impactos da produção científica utilizando a combinação de métodos estatísticos e matemáticos. Ela pode ser aplicada em diversas áreas para estudar uma temática específica e compreender suas particularidades e estado da arte. Essa abordagem se baseia em conhecimentos estabelecidos para descrever padrões de distribuição de literatura científica, como a Lei de Bradford. Por outro lado, a revisão sistemática tem como objetivo encontrar, selecionar, sintetizar e avaliar um tema específico usando protocolos rigorosos e bem definidos. O aprendizado de máquina, por sua vez, é uma área da ciência de dados que confere às máquinas a capacidade de aprender com dados fornecidos, sendo dividido em duas categorias principais: Aprendizado supervisionado e Aprendizado não supervisionado. Quando a bibliometria se combina com a revisão sistemática e técnicas de aprendizado de máquina, resulta em um processo rigoroso de síntese e avaliação de material científico. Isso permite a avaliação eficiente de um grande número de literatura técnica em menos tempo, em comparação com métodos tradicionais. Este trabalho se baseou na aplicação da bibliometria, revisão sistemática e aprendizado de máquina para avaliar a bibliografia científica relacionada ao desenvolvimento regional, mudanças climáticas e gestão de água, com o objetivo de classificar e sintetizar a produção científica nessa área. Para isso, foram buscados bancos de dados contendo literatura técnica em fontes respeitáveis, como a Web of Science e Scopus, para análise. O software R 4.3.1 e sua IDE, o Rstudio, foram utilizados para limpar e tratar os dados. Em seguida, foi realizada uma revisão sistemática desse material por meio do software Asreview para selecionar o material mais relevante, com base

nos critérios estabelecidos. Posteriormente, os materiais selecionados foram submetidos a uma revisão bibliográfica usando a biblioteca Bibliometrix do software R. Foram coletadas 12.489 observações que foram submetidas ao processo de revisão sistemática. Dessas, 101 observações mais relevantes, classificadas pelo software com base nas preferências estabelecidas, foram retiradas para serem avaliadas pelo processo de revisão bibliográfica. Os resultados revelaram informações significativas, como um crescimento anual de 14,5% na produção de material. Foram observadas 66 fontes de publicação (jornais, livros etc.), com uma média de 28,33 citações por documento e um total de 3.690 referências. Foi evidenciado que a partir do ano de 2012 houve um salto na produção científica que perdura até os dias atuais. Foram encontradas também as fontes mais relevantes da área e uma representação espacial da distribuição da produção mundial do material científico. Esse trabalho forneceu informações precisas e relevantes, com base na revisão sistemática, sobre os materiais mais relevantes na área, contribuindo assim para o avanço do conhecimento e a promoção da produção científica de qualidade na temática estudada. Trabalho em parceria com: Gabriel Messias Santana Peixoto, Mateus Silva Rocha, Yuri Batista Oliveira Gomes, Tiago Almeida de Oliveira e Ricardo Alves de Olinda.

Comunicação Oral 8

Palestrante: Roger Almeida Pereira Melo

Título: Utilização de redes bayesianas para investigar fatores associados à exposição acidental de médicos veterinários de Minas Gerais às vacinas de Brucelose

Resumo: A Rede bayesiana (RB) é um método apresentado por Judea Pearl em 1985 que descreve um modelo probabilístico gráfico, representando um conjunto de variáveis e suas dependências condicionais por meio de um grafo acíclico direcionado (directed acyclic graph (DAG)). Os vértices (ou nós) representam proposições (ou variáveis), as arestas (ou arcos), quando são direcionadas, significam as dependências probabilísticas entre essas variáveis. Será apresentado uma aplicação em que foram utilizados os dados de uma pesquisa realizada entre 2018 e 2019, com médicos veterinários credenciados pelo Programa Nacional de Controle e Erradicação da Brucelose e Tuberculose Animal (PNCEBT) em Minas Gerais, com o objetivo de identificar os fatores de risco mais importantes associados à exposição acidental às vacinas anti-Brucella abortus (Brucelose). Uma das respostas de interesse no trabalho é a prevalência de exposição acidental às cepas vacinais S19 e RB51 entre esses profissionais, que foi estimada a partir de um modelo de regressão logístico. Ao utilizar RB, as covariáveis detectadas como mais importantes associadas à exposição acidental às vacinas foram a área principal de trabalho do profissional, a região em que o profissional reside e se o profissional está habilitado para realizar o diagnóstico de brucelose. Todas as análises foram realizadas no software R utilizando o pacote bnlearn. Trabalho em parceria com: Izabela Regina Cardoso de Oliveira e Júlio Sílvio de Sousa Bueno Filho.

Comunicação Oral 9

Palestrante: Lucas Ferreira Rosa

Título: Aplicação do supervisionado Support Vector Machine (SVM) na previsão de séries temporais financeiras

Resumo: A Aprendizagem de Máquina (Machine Learning), uma subárea da Inteligência Artificial (IA), tem como foco o desenvolvimento de algoritmos e modelos capazes de aprender com dados, permitindo que tomem decisões e façam previsões. Antecipar o comportamento de séries temporais, como as do mercado de ações, representa um desafio

notavelmente complexo, devido à influência de uma série de fatores. A literatura científica oferece diversos algoritmos de Machine Learning, tais como Random Forest (RF), Redes Neurais Artificiais (RNA) e Support Vector Machines (SVM), voltados para a previsão desses comportamentos. O algoritmo SVM, desenvolvido por Cortes e Vapnik em 1995, é amplamente reconhecido por sua eficácia e emergiu como um método de aprendizado em diversas áreas. Isso se deve aos seus mecanismos internos que asseguram uma generalização sólida, resultando em previsões precisas. Além disso, o SVM é conhecido por sua capacidade de lidar com distribuições não lineares, graças ao uso de funções específicas, e por sua eficiência no treinamento com grandes conjuntos de dados. Embora seja mais comumente empregado para classificação, o SVM também pode ser aplicado em tarefas de regressão, ou seja, na previsão de valores contínuos com base nos dados, em vez de prever as categorias às quais os dados pertencem. Este estudo apresenta uma aplicação do algoritmo SVM na previsão de séries temporais financeiras, particularmente nas ações do mercado financeiro. A avaliação da performance do SVM em tarefas de regressão pode ser conduzida por meio da utilização de métricas como Raiz do Erro Médio Quadrático (Root Mean Squared Error - RMSE), Erro Médio Absoluto (Mean Absolute Error - MAE) e Coeficiente de Determinação (R2). Essas métricas possibilitam a comparação do desempenho do SVM com outros modelos de aprendizado de máquina. Assim, este trabalho destaca a sólida aplicabilidade do algoritmo de aprendizado de máquina supervisionado SVM na previsão de valores futuros em séries temporais financeiras. Trabalho em parceria com: Paulo Henrique Sales Guimarães e Júlio Sílvio de Sousa Bueno Filho.