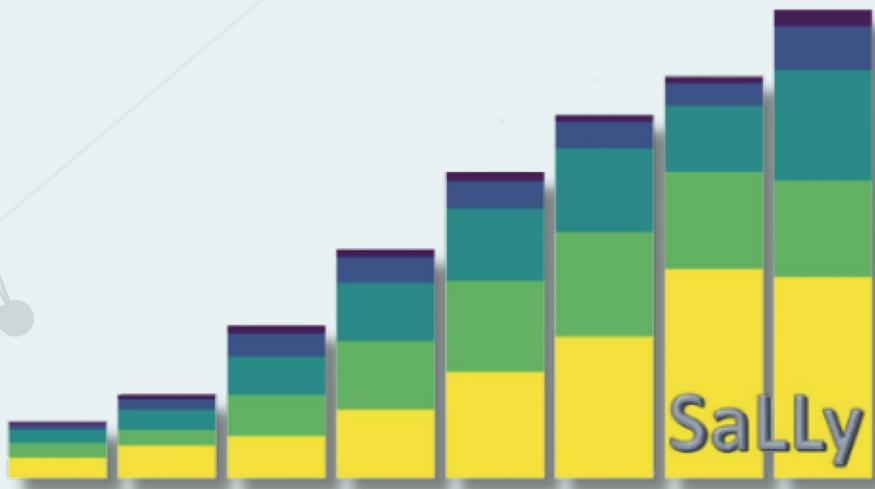


BOOK OF ABSTRACTS



1st SaLLy Day

IME,UFBA, 28 de Outubro de 2023



Livro de resumos do 1st SaLLy Day

28 de outubro de 2023
Salvador - BA, Brasil

Editado por

Paulo Canas Rodrigues

Beatriz Lopes

Universidade Federal da Bahia, Salvador, Brasil

Jonatha Sousa Pimentel

Universidade Federal de Pernambuco, Recife, Brasil

Web Design

João Vitor Rocha da Silva

SaLLy, UFBA, Salvador, Brasil

Design da Capa

Ana Caroline Pinheiro

SaLLy, UFBA, Salvador, Brasil

Citar como:

Rodrigues, P.C.; Pimentel, J.S.; Lopes, B.; Silva, J.V.R. Livro de resumos do 1st SaLLy Day, 2023.

Contents

Part I. Introdução

Bem vindo ao 1st SaLLy Day, UFBA, Salvador, Brasil	7
Comissões	9

Part II. Programação Científica

Programação Científica	13
------------------------------	----

Part III. Conferência de abertura

Aprendizado Estatístico: Transformando Dados em Conhecimento	17
<i>Paulo Canas Rodrigues</i>	

Part IV. Minicursos

MC1: Introdução ao Aprendizado de Máquina Não Supervisionado: Teoria e Prática ..	21
<i>João Vítor Rocha da Silva e Arthur Rios de Azevedo</i>	
MC2: Mineração de Texto com Orange Datamining	22
<i>Crysttian Paixão</i>	

Part V. Mesa Redonda

MR: Mercado de Trabalho em Estatística e Ciência de Dados	25
<i>Suzana de Lima, Gabriela Borges, Júnia Ortis, Carlos Senra e João Vitor da Silva</i>	

Part VI. Sessão de Pôsteres

SP1: Uso de Modelos de Machine Learning para Previsão da Saída de Energia em Planta de Ciclo Combinado	29
<i>Amilton Souza, Ana Oliveira e Karla Esquerre</i>	
SP2: Séries Temporais Hierárquicas para Previsão de Focos de Queimadas no Brasil ..	30
<i>Ana Caroline Pinheiro da Cruz e Paulo Canas Rodrigues</i>	
SP3: MDM: Desenvolvimento de Pacote e Aplicações	31
<i>Arthur Rios Azevedo, Mariana Almeida e Lilia Costa</i>	

SP4: Proposta de Novos Gráficos de Controle para o Monitoramento de Séries Temporais de Dados de Contagem com Sobredispersão	32
<i>David Regalado, Paulo Henrique Ferreira da Silva e Diego Carvalho do Nascimento</i>	
SP5: Mudança de Covariável em Modelos de Aprendizagem de Maquina Epidêmicos	33
<i>Diego Santos Souza e Paulo Canas Rodrigues</i>	
SP6: Combinação de modelos de transferência de aprendizado: uma nova abordagem para a detecção do câncer de pele	34
<i>Fernando Moraes, Adriano Suzuki, Francisco Louzada Neto e Ricardo Rocha</i>	
SP7: Uma Abordagem Bayesiana para Previsão de Resultados de Partidas do Campeonato Brasileiro de Futebol de 2022	35
<i>Gabriel Ribeiro, Lilia Costa e Paulo Ferreira</i>	
SP8: Análise dos Microdados do ENEM de 2009-2019 do Estado da Bahia com Foco em Gênero e Raça	36
<i>Jeilly Costa e Karla Esquerre</i>	
SP9: Avaliação da Eficiência na NBA Através de Aprendizado de Máquina e Análise de Séries Temporais	37
<i>João Vítor Rocha da Silva e Paulo Canas Rodrigues</i>	
SP10: Application of Machine Learning Techniques for Fake News Classification	38
<i>Kim Silva, Crysttian Paixão e Paulo Canas Rodrigues</i>	
SP11: Identifying Adult Asthma Subtypes Through Latent Class Analysis and Uncovering Genetic Panels Using Machine Learning	39
<i>Luciano Gomes, Álvaro Cruz, Maria Rabélo, Raísa Coelho, Gabriela Pinheiro, Cinthia Santana, Jamille Fernandes, Meher Boorgula, Monica Campbell, Kathleen Barnes, Adelmir Machado, Rafael Veiga, Ryan Costa e Camila Figueiredo</i>	
SP12: Analise dos Dados sobre a COVID-19 no Município de Salvador	40
<i>Luiza Moura e Karla Esquerre</i>	
SP13: Uma Abordagem Híbrida de Modelagem Robusta-ponderada AMMI com Esquemas de Pesos Generalizados	41
<i>Marcelo Fonsêca, Paulo Canas Rodrigues e Vanda Lourenço</i>	
SP14: Classification of Images of Fruits and Vegetables with Deep Learning	42
<i>Márcio Henrique Matos de Freitas</i>	
SP15: Escolas Técnicas: Uma Análise do Desempenho dos Estudantes no Exame Nacional do Ensino Médio	43
<i>Miguel Alves, Karla Esquerre e Rogério Filho</i>	
SP16: Ciência de Dados, Inteligência Artificial e Engenharia: Formação Científica com Atuação na Sociedade	44
<i>Vinícius Nascimento</i>	

SP17: Improved Process Capability Assessment Through Semiparametric Piecewise Modeling	45
<i>Pedro Luiz Ramos, Paulo Henrique Ferreira, Nixon Jerez-Lillo e Vinicius da Costa Soares</i>	
Index	47

Part I

Introdução

Bem vindo ao 1st SaLLy Day, um dia de Aprendizado Estatístico e Exploração!

Bem Vindo ao 1st SaLLy Day!

Em nome da Comissão Organizadora Local e da Comissão Científica, é com grande entusiasmo que damos as boas-vindas a todos os amantes da estatística, da aprendizagem de máquina e da ciência de dados para o primeiro SaLLy Day, um evento emocionante e enriquecedor organizado pelo Statistical Learning Laboratory (SaLLy) da Universidade Federal da Bahia.

Neste dia de imersão intelectual, convidamos você a se juntar a nós para uma jornada repleta de descobertas, insights e aprendizados profundos no mundo da estatística e da análise de dados. O 1st SaLLy Day não é apenas um evento, mas uma oportunidade única de se conectar com especialistas renomados, colegas entusiasmados e mentes criativas que estão moldando o cenário da aprendizagem estatística.

A organização deste encontro foi realizada pelo SaLLy - Statistical Learning Laboratórý e o seu objetivo é reunir investigadores e profissionais, da academia e da indústria, que desenvolvam e apliquem métodos estatísticos e computacionais para ciência de dados. Este evento proporcionará um fórum para compartilhar e discutir formas de melhorar o acesso ao conhecimento e promover colaborações interdisciplinares.

O programa científico inclui uma palestra de abertura, uma mesa redonda sobre o mercado de trabalho em Estatística e Ciência de Dados, dois minicursos e uma sessão de pôsters. Os apresentadores de pôster presentes neste 1st SaLLy Day tiveram a possibilidade de concorrer ao prêmio de melhor pôster.

Os organizadores gostariam de agradeceràs instituições que forneceram apoio para tornar esta organização possível. Muito obrigado ao Instituto de Matemática e Estatística por disponibilizar as instalações e ao programa de Mestrado em Matemática pelo apoio financeiro. Por último, mas não menos importante, agradecemos aos palestrantes, aos debatedores da mesa redonda, aos apresentadores de pôsteres, e a todos os participantes por sua contribuição para a confecção de um grande programa científico. Obrigado a todos pela vossa contribuição!

Desejamos-lhe uma estadia agradável e bons momentos em Salvador!

Em nome do Comité do Programa Científico e do Comité Organizador Local,

Paulo Canas Rodrigues
Coordenador da Comissão Científica do 1st SaLLy Day

João Vitor R. Silva
Coordenador da Comissão Organizadora Local do 1st SaLLy Day

Comissão Organizadora Local

- João Vitor Rocha Silva (Coordenador), SaLLy, UFBA
- Ana Caroline Pinheiro da Cruz, SaLLy, UFBA
- Arthur Rios de Azevedo, SaLLy, UFBA
- Beatriz Lopes, SaLLy, UFBA
- Jonatha Sousa Pimentel, SaLLy, UFPE
- Kim Leone, SaLLy, UFBA
- Marcelo Fonseca, SaLLy, UFBA
- Maria Andreina Moreira, SaLLy, UFBA
- Paulo Canas Rodrigues, SaLLy, UFBA

Comissão Científica

- Paulo Canas Rodrigues (Coordenador), SaLLy, UFBA
- Crysttian Paixão, SaLLy, UFBA
- João Vitor Rocha Silva, SaLLy, UFBA
- Nayguel Costa, SaLLy, Petrobras
- Rodrigo Bulhões, SaLLy, UFBA
- Valdério Anselmo Reisen, SaLLy, UFBA

Part II

Programação Científica

Programação Científica

09h00-09h30: Abertura do 1st SaLLy Day

09h30-10h30: Conferência de Abertura - Aprendizado Estatístico: Transformando Dados em Conhecimento, Prof. Dr. Paulo Canas Rodrigues

10h30-11h00: Coffee Break

11h00-12h30: Mesa Redonda - Mercado de Trabalho em Estatística e Ciência de Dados, Suzana de Lima, Gabriela Borges, Júnia Ortis, Carlos Senra e João Vitor da Silva

12h30-14h00: Intervalo para Almoço

14h00-17h00: Mini-Cursos

- Introdução ao Aprendizado de Máquina Não Supervisionado: Teoria e Prática, Arthur Rios e João Vítor Rocha da Silva
- Mineração de Dados com Orange Datamining, Crysttian Paixão

17h00-18h00: Sessão de Postères

Part III

Conferênciа de abertura

Aprendizado Estatístico: Transformando Dados em Conhecimento

Paulo Canas Rodrigues^{1,2}

¹ Universidade Federal da Bahia, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Email: paulocanas@gmail.com

Abstract

Nos últimos anos, foram geradas imensas quantidades de dados, provenientes de sensores, registros de transações de compras, sinais de GPS móveis, imagens digitais de satélite, mídias sociais, entre outros. A recente quarta revolução industrial trouxe quantidades ainda maiores de dados no contexto da internet das coisas. O aumento da coleta de dados trouxe a necessidade de profissionais com mentalidade quantitativa, capazes de transformar esses dados em informações e tomadas de decisão. Nesta palestra, darei uma visão geral sobre como a estatística e a ciência de dados são extremamente importantes em todas as disciplinas e na vida cotidiana. Também apresentarei algumas aplicações de big data e discutirei a utilidade das estatísticas na era da internet das coisas e da inteligência artificial.

Part IV

Minicursos

MC1: Introdução ao Aprendizado de Máquina Não Supervisionado: Teoria e Prática

João Vítor Rocha da Silva^{1,2} e Arthur Rios de Azevedo^{1,2}

¹ Universidade Federal da Bahia, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

E-mail: rochajoaovitor@yahoo.com; arthur.rios.az@hotmail.com

Resumo: O aprendizado de máquina é um ramo da inteligência artificial que permite que os computadores aprendam sem serem explicitamente programados. Isso significa que os computadores podem aprender a realizar tarefas por meio da análise de dados e da identificação de padrões. O aprendizado de máquina está na moda por vários motivos. Primeiro, os avanços tecnológicos nos últimos anos tornaram possível coletar e armazenar grandes quantidades de dados. Esses dados podem ser usados para treinar modelos de aprendizado de máquina que podem ser usados para realizar uma variedade de tarefas como reconhecimento facial, reconhecimento de voz, classificação de imagens e muito mais. Segundo, o aprendizado de máquina é uma tecnologia que pode ser aplicada a uma ampla gama de problemas. Ele pode ser usado para melhorar a precisão de diagnósticos médicos, prever tendências de mercado, personalizar recomendações de produtos e serviços, e muito mais. A relação entre estatística e aprendizado de máquina é estreita. A estatística fornece a base teórica e os métodos para o aprendizado de máquina. Os algoritmos de aprendizado de máquina são geralmente baseados em técnicas estatísticas, como regressão, classificação e análise multivariada. Neste curso, iremos explorar a base teórica por trás das técnicas estatísticas de análise multivariada: Análise de Componentes Principais e Análise Fatorial. Algoritmos de aprendizado de máquina não supervisionado amplamente utilizados. Na segunda parte, iremos mostrar detalhadamente como aplicar as técnicas estudadas nas linguagens de programação R e Python, com estudos de caso para uma melhor compreensão e exemplificação. Na parte final, exploraremos a riqueza de informações que os resultados nos entregam, discutindo e interpretando as análises realizadas ao longo do mini-curso.

MC2: Mineração de Texto com Orange Datamining

Crysttian Paixão^{1,2}

¹ Universidade Federal da Bahia, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

E-mail: crysttian@gmail.com

Resumo: O desenvolvimento tecnológico é responsável pelo aumento no volume de informações geradas e compartilhadas. A extração do conhecimento de um volume crescente de dados é essencial para o estudo em diferentes áreas. Neste curso será apresentado uma das ferramentas que permite explorar diferentes bases de dados, em especial as textuais, utilizando diferentes abordagens, o Orange Data Mining.

Part V

Mesa Redonda

MR: Mercado de Trabalho em Estatística e Ciência de Dados

Suzana de Lima¹, Gabriela Borges², Júnia Ortiz³, Carlos Senra⁴ e João Vitor da Silva^{4,5}

¹ Zup Innvation, Brasil

² iFood, Brasil

³ Senai Cimatec, Brasil

⁴ VX CASE, Brazil

⁵ SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Resumo: Nesta mesa redonda alunos e ex-alunos da UFBA irão discutir suas experiências e vivências dentro do âmbito profissional em diversos estágios de suas carreiras, diversas áreas do amplo mercado de Estatística e Ciência de Dados atual, e perspectivas para o futuro das profissões.

Participantes:

- Suzana de Lima (Zup Innvation)
E-mail: Suzilima81@gmail.com
- Gabriela Borges (iFood, Brasil)
E-mail: gabrilimaborges@hotmail.com
- Júnia Ortiz (Senai Cimatec, Brasil)
E-mail: junia.ortiz@gmail.com
- Carlos Senra (VX CASE, Brasil)
E-mail: crsbsenra@gmail.com
- João Vitor da Silva (VX CASE, Brasil)
E-mail: rochajoaovitor@yahoo.com

Part VI

Sessão de Pôsteres

SP1: Uso de Modelos de Machine Learning para Previsão da Saída de Energia em Planta de Ciclo Combinado

Amilton Souza¹, Ana Oliveira¹ e Karla Esquerre¹

¹ Universidade Federal da Bahia, BA, Brasil
Email: amiltonsouza099@gmail.com

Abstract

Este estudo tem como objetivo realizar uma comparação entre diferentes modelos de Aprendizado de Máquina (Machine Learning - ML) com o propósito de prever a produção de energia em uma Central de Ciclo Combinado de Energia (CCPP). O conjunto de dados utilizado foi extraído do "UCI Machine Learning Repository: Combined Cycle Power Plant Data Set". A implementação prática deste estudo foi realizada no ambiente Google Collaboratory, utilizando as bibliotecas numpy, pandas, plotly e scikitlearn. Os dados provenientes da planta consistem em quatro variáveis independentes: temperatura, pressão ambiente, umidade relativa e vácuo de exaustão, juntamente com uma variável dependente que representa a energia total gerada pela usina, objeto de nossa predição. O banco de dados foi dividido em dois conjuntos: o conjunto de dados de treinamento, contendo as variáveis utilizadas para treinar os modelos, e o conjunto de dados de teste, utilizado para avaliar a eficácia dos modelos. Cinco modelos de Aprendizado de Máquina foram empregados para a previsão: Regressão Múltipla Linear, Regressão de Lasso, Regressão de Ridge, Elastic Net e Random Forest. A avaliação dos resultados foi realizada por meio de métricas como o Erro Quadrático Médio (MSE), o Coeficiente de Determinação (R^2) e o Erro Absoluto Médio (MAE). Os resultados indicam que o modelo Random Forest apresentou o melhor desempenho na previsão da produção de energia, com valores menores de MSE e MAE em comparação aos outros modelos (11, 90 e 2, 42, respectivamente) e um coeficiente de determinação R^2 próximo de 1 (0, 96).

SP2: Séries Temporais Hierárquicas para Previsão de Focos de Queimadas no Brasil

Ana Caroline Pinheiro da Cruz^{1,2} e Paulo Canas Rodrigues^{1,2}

¹ Universidade Federal da Bahia, BA, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Email: anacarolinep.cruz@gmail.com

Abstract

Incêndios florestais são um problema recorrente no Brasil, e a previsão de incêndios florestais desempenha um papel fundamental na concepção de políticas eficazes de prevenção e controle desses fenômenos. Dessa forma, este trabalho tem como objetivo a comparação dos diferentes métodos de previsão de séries temporais hierárquicas utilizando dados dos focos de queimadas ocorridos no Brasil no período de 2011 a 2022. Foi considerada uma série temporal hierárquica de três níveis, os focos de incêndio do Brasil, que são desagregados por bioma, e os focos dos biomas desagregados por município. Foram testados o modelo ARIMA e ETS para as previsões de base, e nove abordagens de reconciliação para garantir que essas previsões sejam coerentes. Essas abordagens foram o bottom-up (BU), três abordagens top-down (TD), middle-out (MO) e quatro métodos de combinação ótima (OLS, WLS(v), WLS(s), Mint(s)). Para avaliar a precisão das previsões, os dados foram divididos em dois, dados de treinamento para estimar os parâmetros do modelo e dados de teste para avaliar sua precisão. As métricas utilizadas foram a Média Absoluta dos Erros (MAE) e a Média Relativa MAE (AvgRelMAE), que determina a melhoria do MAE das previsões reconciliadas em comparação com o MAE das previsões de base. Os resultados mostram que o modelo ARIMA e as abordagens BU, MINT(s), MO e WLS(v) apresentaram os menores MAEs. Dessas quatro, o MINT(s) e MO apresentaram valores baixos de AvgRelMAE.

SP3: MDM: Desenvolvimento de Pacote e Aplicações

Arthur Rios Azevedo^{1,2}, Mariana Almeida¹ e Lilia Costa¹

¹ Universidade Federal da Bahia, BA, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Email: arthur.rios@ufba.br

Abstract

O Modelo Dinâmico de Multiregressão (MDM) é um modelo grafo para séries temporais multivariadas que permite que a estimação da conexão entre variáveis varie no tempo. O MDM também pode ser definido com um modelo de espaço de estados o qual impõe restrições acíclicas, que apesar de dificultar a escolha para a melhor estrutura de rede, pode implicar em uma interpretação causal, de acordo com os modelos de Pearl (2000), para as conexões em um contexto dinâmico. Isto é, a dinamicidade do MDM permite modelar interações entre variáveis em diferentes pontos no tempo, e consequentemente, estimar as conexões da rede de maneira a distinguir os gráficos direcionais acíclicos Markov equivalentes. O mdmr é uma implementação do método desenvolvido por Costa et. al. em linguagem R e código aberto. O pacote está em constante desenvolvimento e a versão estável da livraria pode ser acessada em <https://github.com/arzevedo/mdmr>. Neste pôster apresentaremos duas aplicações que mostram a versatilidade do método. Na primeira iremos investigar a relação entre os novos casos de COVID-19 entre as macrorregiões brasileiras, e na segunda utilizaremos uma aplicação relacionada a neurociência.

SP4: Proposta de Novos Gráficos de Controle para o Monitoramento de Séries Temporais de Dados de Contagem com Sobredispersão

David Regalado¹, Paulo Henrique Ferreira da Silva¹ e Diego Carvalho do Nascimento²

¹ Universidade Federal da Bahia, BA, Brasil

² Universidad de Atacama, Copiapó, Chile

Email: davidregalado@ufba.br

Abstract

O projeto tem como finalidade a obtenção de um sistema onde os usuários possam carregar seu banco de dados, eleger suas variáveis independentes e dependentes e o sistema realize uma análise descritiva delas, depois processar os dados através de modelos estatísticos para séries temporais e com ferramentas de Controle Estatístico de Processos (CEP), com vistas a interpretar melhor os dados e projetar o comportamento deles no futuro. O sistema é responsivo, ou seja, quando algum dos parâmetros é modificado, afeta os gráficos e resumos estatísticos. Alguns aspectos do projeto foram revisados, mas o objetivo original – trabalhar com modelagem de séries temporais (inclusive, para dados de contagem) e desenvolver a parte de CEP correspondente – se mantém. A mudança permitiu integrar o Prof. Dr. Diego Carvalho do Nascimento (do Departamento de Matemática da Faculdade de Engenharia da Universidade de Atacama, Chile) como coorientador, inclusive disponibilizando bancos de dados reais e sua vasta experiência na área de Estatística e no desenvolvimento de sistemas estatísticos responsivos. Os objetivos e metas são estudar, conhecer, construir e interpretar relatórios responsivos e interativos. A partir destes, o usuário poderá obter, ao proporcionar seus próprios dados, a informação estatística (com indicadores básicos e a continuação a uma análise mais aprofundada) referente aos mesmos. Com isso, é possível desenvolver maiores habilidades sobre dados, estatística descritiva, desenvolver programas em softwares estatísticos, estudar sobre os principais métodos utilizados para relatórios responsivos e modelagem dos dados referentes a cada caso particular em cada um dos bancos de dados.

SP5: Mudança de Covariável em Modelos de Aprendizagem de Maquina Epidêmicos

Diego Santos Souza^{1,2} e Paulo Canas Rodrigues^{1,2}

¹ Universidade Federal da Bahia, BA, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Email:seg.diego1355@gmail.com

Abstract

A pandemia de Covid-19 repercutiu vários anos na sociedade, seja em aspectos econômicos ou no aspecto mais grave, a mortalidade. Em estimativas iniciais, 15 milhões de pessoas morreram por covid, mas existem fortes evidências de que essas mortes estão subestimadas, pois a pandemia revelou-se um dos grandes problemas da modelagem empírica, que é a falta de qualidade dos dados, que se manifestam principalmente com subestimação e ausência de dados. Algumas metodologias foram propostas para contornar esses problemas como: uso de excesso de mortes, interpolação por modelo bayesiano e transferência de aprendizagem, sendo esta última tendo grande aplicabilidade para casos de subestimação e ausência de dados. Nesse trabalho demonstraremos a aplicabilidade do uso de transferência de aprendizagem para epidemia de covid19 utilizando, Covariate Shift, que permite utilizar modelos estimados com dados de um determinado contexto (covariáveis: sexo, idade, renda etc diferentes) e aplicá-los em outro contexto sem perda significativa de generalização de previsão. Para isso utilizamos dois países: Argentina e Equador com uma das maiores diferenças entre seus contextos, e construímos uma rede neural recorrente, para 60 dias iniciais da pandemia, predizer a mortalidade media usando dados da Argentina com correção de Covariate Shift para mortalidade do Equador. Usando o RMSE como métrica de avaliação, temos o modelo corrigido teve RMSE = 19,80 que é 0,38% menor que o modelo não corrigido, RMS 32,20, que usa dados da Argentina sem correção.

SP6: Combinação de modelos de transferência de aprendizado: uma nova abordagem para a detecção do câncer de pele

Fernando Moraes¹, Adriano Suzuki¹, Francisco Louzada Neto¹ e Ricardo Rocha²

¹ Universidade de São Paulo, SP, Brasil

² Universidade Federal da Bahia, BA, Brasil

Email: fernandomoraes@usp.br

Abstract

Recentemente os modelos de Deep Learning têm ganhado muito destaque na área de análise de imagem, porém necessitam de muitos dados para terem bons desempenhos. Isto dificulta a aplicação em bases de dados médicas visto que existe uma dificuldade em se obter muitas observações equiparadamente, entre os diferentes casos como, por exemplo, patológicos e normal. Para pequenas bases de dados pode-se utilizar modelos de transferência de aprendizado, ensemble e aumento de dados para se ter melhores desempenhos na tarefa de classificar imagens, já para base de dados desbalanceadas pode-se utilizar técnicas de reamostragem como undersampling e oversampling. Neste trabalho propomos uma nova abordagem baseada no ensemble de modelos de transferência de aprendizado com diferentes pesos para melhorar a predição de observações da classe minoritária em pequenas bases de dados desbalanceadas considerando reamostragem e aumento de dados. Ao final um experimento é realizado com a base de dados de imagens de câncer de pele, tem-se como objetivo classificar imagens de câncer de pele como malignas ou benignas. A partir dos resultados obtidos nota-se que as combinações com melhores métricas são obtidas com a utilização de reamostragem e as piores sem a utilização desta técnica. O que comprova que a utilização da metodologia de reamostragem melhora o desempenho da combinação em prever a classe minoritária.

SP7: Uma Abordagem Bayesiana para Previsão de Resultados de Partidas do Campeonato Brasileiro de Futebol de 2022

Gabriel Ribeiro¹, Lilia Costa¹ e Paulo Ferreira¹

¹ Universidade Federal da Bahia, BA, Brasil
Email: gabrielrbr76@gmail.com

Abstract

O futebol é o esporte mais praticado no mundo e é considerado um dos mais imprevisíveis, em que algumas vezes uma equipe considerada como mais fraca consegue superar uma outra mais forte. Dentro deste contexto, o campeonato brasileiro em especial é um dos maiores exemplos dessa imprevisibilidade, apresentando uma grande quantidade de resultados não esperados. Com o objetivo de fazer previsões de resultados de partidas do Campeonato Brasileiro da Série A de 2022, foram comparadas as seguintes estruturas de modelos de Poisson através da abordagem Bayesiana: não hierárquica (efeito fixo) e hierárquica (efeito aleatório com dois níveis), como propostas por Blanco et al (2021), e como proposta de implementação os modelos com inflação de zeros e com uma estrutura dinâmica para o modelo Poisson com inflação de zeros. Além disso, foram utilizadas distribuições a priori não informativas e informativas, incorporando informações do campeonato brasileiro do ano anterior. Para avaliação da qualidade das previsões foi usada a medida de DeFinetti, além da comparação da qualidade de ajuste utilizando a métrica Leave-One-Out Cross-Validation, em que os modelos apresentaram resultados satisfatórios. De acordo com a maioria das métricas utilizadas na comparação dos métodos, o modelo dinâmico de Poisson com inflação de zeros obteve o melhor resultado.

SP8: Análise dos Microdados do ENEM de 2009-2019 do Estado da Bahia com Foco em Gênero e Raça

Jeilly Costa¹ e Karla Esquerre¹

¹ Universidade Federal da Bahia, BA, Brasil
Email: jeilly.almeida.costa@gmail.com

Abstract

O presente trabalho foi desenvolvido buscando avaliar o desempenho no Exame Nacional do Ensino Médio dos alunos das escolas públicas estaduais da Bahia. Para a análise, foi utilizado um banco de dados contendo microdados do ENEM disponibilizados pelo INEP sobre o rendimento dos estudantes no exame entre os anos de 2009 e 2019. Assim, utilizando a linguagem de programação Python, foi possível mensurar as diferenças de desempenho com relação à raça e ao gênero dos participantes. Constatou-se que embora estudantes do gênero feminino fossem a maioria dos inscritos no ENEM durante o período de tempo avaliado (9000 a 13000 inscrições a mais), elas tiveram desempenho pior nos domínios de ciências da natureza (entre 3% e 7% abaixo), ciências humanas (entre 3% e 5% abaixo), matemática (entre 4% e 14% abaixo), linguagens e códigos (até 3% abaixo) e na nota geral (entre 2% e 4% abaixo), tendo rendimento superior ou aproximadamente igual aos estudantes do gênero masculino apenas na nota da redação (até 4% acima). Com relação à raça, os estudantes autodeclarados brancos e amarelos tiveram desempenho melhor do que as outras raças (até 6% acima) na nota geral. Já os estudantes autodeclarados indígenas tiveram, com exceção de 2011 e 2014, o pior desempenho geral (até 6% abaixo). A pesquisa realizada possibilita um melhor entendimento das estatísticas de desempenho dos estudantes com recortes de gênero e raça, podendo apoiar a proposição de estratégias e de políticas públicas educacionais.

SP9: Avaliação da Eficiência na NBA Através de Aprendizado de Máquina e Análise de Séries Temporais

João Vítor Rocha da Silva^{1,2} e Paulo Canas Rodrigues^{1,2}

¹ Universidade Federal da Bahia, BA, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Email: rochajoaovitor1@yahoo.com

Abstract

O basquete é um esporte popular que vem se desenvolvendo constantemente ao longo dos anos, com mudanças de regras, equipamentos e tecnologias que impactam diretamente a dinâmica do jogo. A ciência de dados voltada para análise esportiva, por sua vez, tem sido cada vez mais utilizada para auxiliar tomada de decisões e aprimoramento do desempenho de atletas, times e organizações. O presente estudo teve como objetivo analisar as temporadas regulares da NBA de 2013-14 a 2021-22, utilizando a análise de séries temporais. Primeiramente, a partir da utilização de técnicas de aprendizado de máquina, foi possível criar uma variável latente referente ao desempenho dos jogadores, que após o seu agrupamento por equipes, nos possibilitou criar séries históricas, de estrutura hierárquica, de desempenho das equipes, o que consequentemente nos permitiu estudar o ranqueamento das equipes, divisões, conferências e da liga em geral. Ao combinar uma análise descritiva e exploratória dos dados históricos de ranqueamento com modelos de análise e previsão de séries temporais, os resultados obtidos fornecem uma abordagem completa do estudo histórico das temporadas regulares da NBA de 2013-14 a 2021-22, onde foi possível estudar a eficiência de utilização de jogadores das equipes e da liga, principalmente durante o período da pandemia de COVID-19, e o impacto de práticas não permitidas pela NBA, como o tanking, na tomada decisão das equipes durante a temporada regular.

SP10: Application of Machine Learning Techniques for Fake News Classification

Kim Silva^{1,2}, Crysttian Paixão^{1,2} e Paulo Canas Rodrigues^{1,2}

¹ Universidade Federal da Bahia, BA, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil

Email: k.leone_silva@hotmail.com

Abstract

Fake News consists of disseminating fake news in various social and digital media such as newspapers, television networks, and the internet. Fake news is not a new phenomenon in human behavior. However, the current dissemination is very different from what happened in the past. Social networks and the contemporary world have made it possible for the spread of lies to occur quickly and even intentionally. This causes serious problems, and its impacts can be felt in the real world. The identification of Fake News can be useful in several contexts and can be used, for example, as a news filter in the virtual space. Thus, the present work aims to propose and evaluate strategies for processing and applying machine learning models, to improve the performance of classifiers in the problem of identifying Fake News in Brazilian news.

SP11: Identifying Adult Asthma Subtypes Through Latent Class Analysis and Uncovering Genetic Panels Using Machine Learning

Luciano Gomes¹, Álvaro Cruz², Maria Rabêlo¹, Raísa Coelho¹, Gabriela Pinheiro², Cinthia Santana², Jamille Fernandes³, Meher Boorgula⁴, Monica Campbell⁴, Kathleen Barnes⁴, Adelmir Machado², Rafael Veiga⁵, Ryan Costa¹ e Camila Figueiredo¹

¹ Universidade Federal da Bahia, BA, Brasil

² PROAR, BA, Brasil

³ Universidade Federal do Oeste da Bahia

⁴ University of Colorado, Denver, USA

⁵ The Babraham Institute, Cambridge, United Kingdom

Email: gama.luciano@hotmail.com

Abstract

Asthma is a chronic respiratory condition. Its diverse nature poses challenges in standardizing treatment due to variations in symptoms and severity among individuals. We aim to discern adult asthma subtypes using Latent Class Analysis and examine genetic panels for each classification through supervised methods. The ProAR cohort included 577 individuals with asthma diagnoses. Latent Class Analysis was applied to categorize subjects into distinct asthma subtypes based on various environmental exposures, clinical characteristics, and laboratory features. Ten different algorithms (K-Nearest Neighbors, Naïve Bayes, Artificial Neural Network, Support Vector Machine, Classification and Regression Trees, C5.0, Bagging, AdaBoost, Random Forest, and XGBoost) were utilized to build predictive classifier models. The study focused on 1,009,762 Single Nucleotide Variants using the Boruta algorithm for feature selection. We identified four asthma subtypes. Cluster 1 comprises individuals with mild to moderate asthma, easily manageable with a low exacerbation rate. These individuals exhibit atopic hyperreactivity, elevated IgE levels, and positive skin prick test results for allergens. Cluster 2 consists of patients with severe asthma and heightened responsiveness to environmental exposures, experiencing frequent exacerbations and impaired lung function. Cluster 3 represents Asthma-COPD Overlap Syndrome without reactivity to worsening exposure, with better asthma control and lung function compared to cluster 2. Cluster 4 represents mild hypo-reactive asthmatics with low atopic reactivity and superior pulmonary function. Support Vector Machine demonstrated the best performance in predicting gene panels. Our study distinguished unique subtypes of asthma and confirmed established genetic links. In addition, it reveals new genetic associations.

SP12: Analise dos Dados sobre a COVID-19 no Município de Salvador

Luiza Moura¹ e Karla Esquerre¹

¹ Universidade Federal da Bahia, BA, Brasil
Email: luiza.paula@ufba.br

Abstract

Este trabalho busca explorar os dados acerca dos casos de COVID-19 utilizando ferramentas de visualização de dados. Aqui foram consideradas informações em função do tempo e da localização geográfica da notificação do caso na cidade de Salvador a fim de verificar as nuances do número de notificações e do número de óbitos na cidade apoiada em função destas variáveis e se existem diferenças significativas nos dados em conformidade com elas. Para verificar isso foram utilizados dados indicadores da transparência da Secretaria Municipal de Saúde que abrangem o período entre os anos de 2020 e 2021 em todo o município de Salvador que foram expostos por meio de figuras produzidas usando as bibliotecas ggplot2 e plotly em linguagem R, além da biblioteca folium em linguagem Python. A partir disso foi possível verificar os meses de junho de 2020 e março de 2021 como meses marcantes em casos de COVID-19. Além disso, utilizando a localização geográfica, apesar dos bairros com mais casos serem Pituba, Brotas e Pernambués, eles tiveram coeficientes de incidência pequenos quando comparados aos maiores: Jardim Armação, Patamares e Retiro. Com estas ferramentas foi possível verificar as transformações em termos de casos, durante os anos de 2020 e 2021, de acordo com os meses de notificação e quanto a sua localização geográfica.

SP13: Uma Abordagem Híbrida de Modelagem Robusta-ponderada AMMI com Esquemas de Pesos Generalizados

Marcelo Fonsêca^{1,2}, Paulo Canas Rodrigues^{1,2} e Vanda Lourenço^{2,3}

¹ Universidade Federal da Bahia, BA, Brasil

² SaLLy - Statistical Learning Laboratory, UFBA, Brasil ³ Universidade NOVA Lisboa, Caparica, Portugal

Email: fonscabmarcelo@gmail.com

Abstract

O modelo AMMI e suas generalizações têm demonstrado ser altamente eficazes na identificação de genótipos adaptáveis e estáveis em condições ambientais específicas, desempenhando um papel crucial em programas de melhoramento de plantas. No entanto, a presença de dados atípicos em conjuntos de dados de culturas pode prejudicar o desempenho do AMMI. Portanto, é fundamental aprimorar o modelo AMMI com técnicas estatísticas para garantir resultados confiáveis, preservando sua eficácia nas tomadas de decisão no melhoramento de culturas. Neste contexto, apresentamos uma inovadora estrutura de modelagem AMMI denominada RW-AMMI, que combina métodos robustos e ponderados para capturar as interações entre genótipos e ambientes. Além disso, introduzimos um conjunto abrangente de nove esquemas de pesos para os modelos ponderados (W-AMMI), modelos robustos (R-AMMI) e modelos robustos-ponderados (RW-AMMI). Para avaliar a eficácia da nossa abordagem, conduzimos simulações de Monte Carlo em conjuntos de dados contaminados e não contaminados. Em seguida, comparamos os resultados da nossa abordagem com os modelos AMMI, W-AMMI e R-AMMI, utilizando os nove esquemas de pesos mencionados. Além disso, validamos a nossa abordagem em aplicações práticas, utilizando dados reais de culturas.

SP14: Classification of Images of Fruits and Vegetables with Deep Learning

Márcio Henrique Matos de Freitas^{1,2}

¹ Universidade Federal de Sergipe, SE, Brasil

² Instituto Federal de Sergipe, SE, Brasil

Email: marcio.freitas@dcomp.ufs.br

Abstract

The work addresses techniques of Deep Learning for image classification of fruits and vegetables in the agriculture environment, considering that the branch has relevance not only in the market aspect but also in the scientific scope. Agricultural production has a major impact market and increasingly has been optimizing its processes efficiently and quickly. In Brazil, activities in the field of agriculture represent about 5% of GDP national market, with revenues of 100 billion reais, being responsible for the commercial values cials that revolve around the country. The northeast region is a region that has a certain relevance in its plurality of cultivation of fruits and vegetables. Although, there is still little technological application in your region, this reflects in the little acceleration of production, quality of product. A loss rate of fruits and vegetables was identified and, consequently, a drop in the production of the branch in the state of Sergipe, thus emerged the need to understand the that could be causing, hence the fact of investigating such an event, where is the failure? in the process of separating the planting area ?. The lack of classification contributes enough for an occurrence of loss and, consequently, damage to the harvest. With the results, it was possible verify that the hit percentage of our network was above 90%, but precisely of 95.3% in the classification of fruits and vegetables. Showing that the techniques of deep learning used has been a great solver of problems in recent years.

SP15: Escolas Técnicas: Uma Análise do Desempenho dos Estudantes no Exame Nacional do Ensino Médio

Miguel Alves¹, Karla Esquerre¹ e Rogério Filho²

¹ Universidade Federal da Bahia, BA, Brasil

² Universidade Federal de Pernambuco, PE, Brasil

Email:miguelalves@ufba.br

Abstract

O ensino técnico no Brasil é frequentemente apontado como um vetor potencial para a melhoria dos indicadores educacionais e o desenvolvimento econômico nacional. Este estudo emprega uma abordagem quantitativa para avaliar o impacto dessa modalidade de ensino no desempenho acadêmico dos estudantes brasileiros. Utilizando dados provenientes do Exame Nacional do Ensino Médio (ENEM) de 2019, modelos de regressão linear foram implementados para analisar e comparar o desempenho acadêmico médio entre diferentes categorias de escolas. O escopo do estudo engloba a análise de desempenho dos estudantes matriculados no último ano do ensino médio em escolas públicas de ensino regular, escolas públicas de ensino técnico, escolas técnicas federais e escolas privadas de ensino regular. A média aritmética das 5 áreas de avaliação do ENEM foi utilizada como variável dependente enquanto fatores socioeconômicos dos estudantes e das escolas como variáveis independentes. Os resultados indicaram que a inclusão da modalidade técnica na rede pública não traz, em média, melhorias significativas, exceto nos estados do Ceará, São Paulo e Pernambuco, onde os estudantes das escolas públicas técnicas obtiveram desempenho significativamente superior ao das escolas de ensino regular. O Ceará, por exemplo, mostrou um aumento de performance de aproximadamente 5% quando fixado os mesmos parâmetros. As escolas federais, por sua vez, manifestam um desempenho superior em relação tanto às escolas técnicas quanto às escolas públicas regulares em âmbito nacional de um modo geral. Pesquisas futuras visam explorar análises pareadas de estudantes e escolas com perfis socioeconômicos semelhantes em diferentes tipos de escolas.

SP16: Ciência de Dados, Inteligência Artificial e Engenharia: Formação Científica com Atuação na Sociedade

Vinícius Nascimento¹

¹ Universidade Federal da Bahia, BA, Brasil
Email: vinicius.nascimento@ufba.br

Abstract

O objetivo deste trabalho é apresentar um relato de experiência da utilização da ciência de dados e a inteligência artificial (IA) como ferramenta de apoio aos projetos científicos desenvolvidos em um Clube de Ciências de uma escola pública de Salvador-BA. As ações foram promovidas pelo Projeto Ciência de Dados na Educação Pública em 2023. Foram realizados 12 encontros, com duração de aproximadamente uma hora e quarenta minutos, onde conteúdos foram expostos e práticas foram desenvolvidas nas temáticas de estatística, ciência de dados, metodologia científica, programação e inteligência artificial. Como resultado dessas ações, estudantes do ensino médio público puderam: conhecer os conceitos fundamentais de estatística, ciência de dados e IA; experimentar técnicas de programação em Python, explorar dados e interpretá-los a partir de gráficos; desenvolver competências técnicas e habilidades matemáticas presentes na Base Nacional Comum Curricular (BNCC). Portanto, tem-se a formação de jovens de ensino médio público mais capacitados em conceitos relacionados a dados e novas tecnologias, como IA. Além disso, esses estudantes puderam reconhecer o potencial desses assuntos como uma ferramenta que potencializa a formação científica, o crescimento intelectual e avanço da ciência. Desse modo, nota-se o impacto dessas abordagens que alinharam ciência de dados, inteligência artificial e formação científica com atuação na sociedade. A interseção entre tecnologia, educação, ciência e sociedade evidencia o potencial transformador dessas abordagens na formação de uma geração informada, capacitada e engajada.

SP17: Improved Process Capability Assessment Through Semiparametric Piecewise Modeling

Pedro Luiz Ramos¹, Paulo Henrique Ferreira², Nixon Jerez-Lillo¹ e Vinicius da Costa Soares²

¹ Pontificia Universidad Católica de Chile, Santiago, Chile

² Universidade Federal da Bahia, BA, Brasil

Email: vinicius4burame@gmail.com

Abstract

Piecewise models have gained popularity as a useful tool in reliability and quality control/monitoring, particularly when the process data deviates from a normal distribution. In this study, we develop maximum likelihood estimators (MLEs) for the process capability indices, denoted as C_{pk} , C_{pm} , C^*_{pm} and C_{mk} , using a semiparametric model. To account for bias in the MLEs for small sample sizes, we adopt a bias-correction approach to obtain improved estimates. Furthermore, we extend the proposed method to situations where the change-points in the density function are unknown. To estimate the model parameters efficiently, we employ the profiled maximum likelihood approach. Our simulation study reveals that the suggested method yields accurate estimates with low mean relative and squared errors. Finally, we provide real-world data applications to demonstrate the superiority of the proposed procedure over existing ones.

Index

- Almeida, M., 31
Alves, M., 43
Azevedo, A.R., 31
- Barnes, K., 39
Boorgula, M., 39
Borges, G., 25
- Campbell, M., 39
Coelho, R., 39
Costa, J., 36
Costa, L., 31, 35
Costa, R., 39
Cruz, A., 39
Cruz, A.C.P., 30
- Esquerre, K., 29, 36, 40, 43
- Fernandes, J., 39
Ferreira, P.H., 35, 45
Figueiredo, C., 39
Filho, R., 43
Fonsêca, M., 41
Freitas, M.H.M., 42
- Gomes, L., 39
- Jerez-Lillo, N., 45
- Lima, S., 25
Lourenço, V., 41
- Machado, A., 39
- Moraes, F., 34
Moura, L., 40
- Nascimento, D.C., 32
Nascimento, V., 44
Neto, F.L., 34
- Oliveira, A., 29
Ortis, J., 25
- Paixão, C., 22, 38
Pinheiro, G., 39
- Rabêlo, R., 39
Ramos, P.L., 45
Regalado, D., 32
Ribeiro, G., 35
Rios, A., 21
Rocha, J.V., 21, 25
Rocha, R., 34
Rodrigues, P.C., 17, 30, 33, 37, 38, 41
- Santana, C., 39
Senra, C., 25
Silva, J.V.R., 37
Silva, K., 38
Silva, P.H.F., 32
Soares, V.C., 45
Souza, A., 29
Souza, D.S., 33
Suzuki, A., 34
- Veiga, R., 39