UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Ponderação baseada em expertise para modelos de regressão com rótulos ruidosos

**Milene Regina dos Santos**

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

**Milene Regina dos Santos**

# Ponderação baseada em expertise para modelos de regressão com rótulos ruidosos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Izbicki

**USP – São Carlos**
**Março de 2025**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

dos Santos, Milene Regina
d722p Ponderação baseada em expertise para modelos de regressão com rótulos ruidosos / Milene Regina dos Santos; orientador Rafael Izbicki. -- São Carlos, 2025.
67 p.

Tese (Doutorado - Programa de Pós-Graduação em Matemática) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Modelo Ponderado. 2. Rótulos Ruidosos. 3. Regressão. I. Izbicki, Rafael , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**Milene Regina dos Santos**

# Expertise-based weighting for regression models with noisy labels

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Izbicki

**USP – São Carlos**
**March 2025**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

## Folha de Aprovação

Defesa de Tese de Doutorado da candidata Milene Regina dos Santos, realizada em 18/02/2025.

### Comissão Julgadora:

Prof. Dr. Rafael Izbicki (UFSCar)

Profa. Dra. Teresa Cristina Martins Dias (UFSCar)

Prof. Dr. Paulo Henrique Ferreira da Silva (UFBA)

Prof. Dr. Rafael Bassi Stern (IME-USP)

Prof. Dr. Victor Fossaluza (IME-USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

*Este trabalho é dedicado a todos que sabem que aprender é eterno,*
*que cada um tem seu tempo e*
*que todo rótulo é viesado.*

# AGRADECIMENTOS

Agradeço a Deus, por me iluminar a cada dia.

Ao meu orientador, Dr. Rafael Izbicki, pela compreensão, paciência, compartilhamento, aprendizado e, principalmente, por tornar este trabalho possível.

Aos secretários do programa, Monique da Conceição e Julio Cezar de Barros, que sempre estão dispostos a esclarecer as dúvidas e auxiliar os discentes.

Agradeço ao meu marido, Ivan Carlos Cagnin, por me apoiar a cada dia.

Ao meu pai, Mauro José dos Santos, por ser o meu espelho de honestidade e caráter.

À minha mãe, Maria Izabel Gonçalves dos Santos, por me ensinar a nunca ter vergonha de perguntar.

À minha irmã, Aline Cristina dos Santos, por me ensinar como ser resiliente e perdoar a quem nos tem ofendido.

À minha sobrinha, por ser como a filha que abdiquei momentaneamente para consolidar minha carreira.

Por fim, agradeço a todos os trabalhos CLTs que tive até o momento, pois me proporcionaram flexibilidade para a construção deste trabalho e me mostraram inúmeras aplicações com a estatística, o que me trouxe mais fascínio por esta área.

*“(...) o fato de que um time leve o troféu*
*não serve como indicação confiável*
*de que realmente é melhor time do compeonato."*
*Do livro "O andar do bêbado"*
*de Leonard Mlodinow.*

# RESUMO

SANTOS, M. R. **Ponderação baseada em expertise para modelos de regressão com rótulos ruidosos**. 2025. 67 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Os métodos de regressão tradicionais pressupõem a disponibilidade de rótulos precisos para o treinamento dos modelos. No entanto, em muitos contextos, obter rótulos totalmente precisos pode não ser factível, sendo necessário recorrer a múltiplos especialistas cujas opiniões podem divergir devido a ruídos humanos intrínsecos e difíceis de mensurar. Esses ruídos podem estar nas variáveis de entrada, já que diferentes especialistas podem interpretar certas observações de maneiras distintas, devido as *expertises*.

Neste trabalho a proposta é uma abordagem inovadora para o treinamento de modelos de regressão em cenários nos quais os rótulos apresentam ruído, resultante de múltiplas opiniões divergentes de especialistas. O método proposto consiste, primeiramente, em estimar a *expertise* de cada especialista de forma geral e a nível de instância, atribuindo pesos às suas opiniões. Em seguida, realiza-se uma média ponderada dessas opiniões, utilizando os pesos aprendidos para ajustar o modelo de regressão com base nas variáveis de entrada.

A abordagem proposta tem fundamentação teórica sólida e, por meio de experimentos com dados simulados e reais, demonstrou-se empiricamente superior a métodos tradicionais. Em suma, o método oferece uma solução simples, rápida e eficaz para o treinamento de modelos de regressão em cenários com rótulos ruidosos, gerados por diferentes opiniões de especialistas.

**Palavras-chave:** Modelo ponderado, Rótulos Ruidosos, Regressão.

# ABSTRACT

SANTOS, M. R. **Expertise-based weighting for regression models with noisy labels**. 2025. 67 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Traditional regression methods assume the availability of precise labels for training models. However, in many contexts, obtaining fully accurate labels may not be feasible, requiring reliance on multiple experts whose opinions may diverge due to intrinsic human noise, which is difficult to measure. This noise can be present in the input variables, as different experts may interpret certain observations in distinct ways due to their expertise.

In this work, we propose an innovative approach to training regression models in scenarios in which the labels contain noise, resulting from multiple divergent expert opinions. The proposed method first estimates each expert's expertise both generally and at the instance level, assigning weights to their opinions. Then, a weighted average of these opinions is computed, using the learned weights to adjust the regression model based on the input variables.

The proposed approach has a solid theoretical foundation and, through experiments with both simulated and real data, has been empirically demonstrated to outperform traditional methods. In summary, this method provides a simple, fast, and effective solution for training regression models in scenarios with noisy labels generated by differing expert opinions.

**Keywords:** Weighted Model, Noisy Labels, Regression.

# LISTA DE ILUSTRAÇÕES


Figura 1 – Boxplot das variabilidades dos especialistas - Dados 1 - Simulados . . . . . 43
Figura 2 – Boxplot das variabilidades dos especialistas - Dados 2 - Simulados . . . . . 44
Figura 3 – Boxplot das variabilidades dos especialistas - Dados 3 - Simulados . . . . . 46
Figura 4 – Boxplot das variabilidades dos especialistas - Dados 4 - Simulados . . . . . 48
Figura 5 – Skin Segmentation - Boxplot das variabilidades dos especialistas . . . . . 51
Figura 6 – Skin Segmentation - Gráfico de dispersão entre o y observado simulado e o y estimado . . . . . 52
Figura 7 – Internet Firewall Data Versão 1 - Boxplot das variabilidades dos especialistas 54
Figura 8 – Internet Firewall Data - Versão 1 - Boxplot das 4 variáveis selecionadas como variâncias por observações de cada especialista, respectivamente . . . . . 54
Figura 9 – Internet Firewall Data - Gráfico de dispersão entre o y observado simulado e o y estimado pelo método WEAR-INS . . . . . 55
Figura 10 – 3D Road Network (North Jutland, Denmark) - Boxplot das covariáveis . . . 56
Figura 11 – 3D Road Network (North Jutland, Denmark) - Boxplot das variabilidades dos especialistas . . . . . 56
Figura 12 – 3D Road Network (North Jutland, Denmark) - Gráfico de dispersão entre o y observado simulado e o y estimado . . . . . 57
Figura 13 – Firewall de Internet (Versão 2) - Boxplot das variabilidades dos especialistas 60
Figura 14 – Firewall de Internet (Versão 2) - Gráfico de dispersão entre o y observado simulado e o y estimado . . . . . 60
Figura 15 – Firewall de Internet (Versão 3) - Boxplot das variabilidades dos especialistas 61
Figura 16 – Firewall de Internet (Versão 3) - Gráfico de dispersão entre o y observado simulado e o y estimado . . . . . 61

# LISTA DE TABELAS


Tabela 1 – Dados 1 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 43
Tabela 2 – Dados 2 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 45
Tabela 3 – Dados 3 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 46
Tabela 4 – Dados - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 48
Tabela 5 – Resultados comparativos dos EQMs de diferentes modelos em quatro conjuntos de dados simulados. . . . . 49
Tabela 6 – Resultados comparativos dos EQMs de diferentes modelos em três conjuntos de dados reais. . . . . 50
Tabela 7 – Skin Segmentation - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 52
Tabela 8 – Internet Firewall Data - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 55
Tabela 9 – 3D Road Network (North Jutland, Denmark - Pesos estimados para cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 57
Tabela 10 – Firewall de Internet - Versão 2 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 59
Tabela 11 – Firewall de Internet - Versão 3 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar. . . . . 59
Tabela 12 – Firewall de Internet (Versão 2 e Versão 3) - Comparação entre diferentes métodos, a partir do erro preditivo. . . . . 59

# SUMÁRIO


1 INTRODUÇÃO . . . . . 23
1.1 Objetivos . . . . . 27
*1.1.1 Objetivos Específicos* . . . . . *27*

2 DESENVOLVIMENTO . . . . . 29
2.1 Motivação . . . . . 30
2.2 Modelo Ponderado Proposto . . . . . 36
*2.2.1 WEAR: Regressão Média Ponderada pela Expertise* . . . . . *36*
*2.2.2 WEAR-INS: Regressão Média Ponderada pela Expertise por Observação* . . . . . *37*

3 EXPERIMENTOS . . . . . 39
3.1 Dados Simulados . . . . . 42
*3.1.1 Dados 1* . . . . . *42*
*3.1.2 Dados 2* . . . . . *43*
*3.1.3 Dados 3* . . . . . *45*
*3.1.4 Dados 4* . . . . . *46*
3.2 Dados Reais . . . . . 49
*3.2.1 Skin Segmentation* . . . . . *50*
*3.2.2 Internet Firewall Data - Versão 1* . . . . . *53*
*3.2.3 3D Road Network (North Jutland, Denmark)* . . . . . *55*
*3.2.4 Firewall de Internet - Versão 2 e Versão 3* . . . . . *57*

4 CONCLUSÃO . . . . . 63
4.1 Trabalhos Futuros . . . . . 64

REFERÊNCIAS . . . . . 65

CAPÍTULO

# 1

# INTRODUÇÃO

Com o avanço da tecnologia, estatística e computação têm se tornado cada vez mais interdependentes, especialmente em áreas como inteligência artificial, aprendizado de máquina e ciência de dados. Esses campos estão transformando a forma como problemas complexos são abordados, oferecendo soluções que combinam poder preditivo com análise de grandes volumes de dados (FACELI *et al.*, 2021). No centro dessas inovações está o aprendizado de máquina supervisionado, que, ao usar dados rotulados para treinar modelos, enfrenta desafios como a presença de rótulos ruidosos. Neste trabalho, exploramos esses desafios e propomos novas abordagens para lidar com o impacto de rótulos imprecisos, que afetam diretamente a eficácia dos modelos preditivos.

Os métodos de regressão de aprendizado supervisionado têm ampla aplicação em cenários reais, auxiliando nas tomadas de decisões baseadas em dados. Na área médica, esses métodos são amplamente utilizados, empregando o histórico do paciente, variáveis antropométricas, imagens de ressonância, medidas de exames, dentre outros, como dados de entrada para auxiliar no diagnóstico e minimizar os efeitos de um prognóstico (PAIXÃO *et al.*, 2022). No setor financeiro, eles são usados para prevenir fraudes e estimar o risco do cliente, ou seja, estratégias que visam minimizar as perdas financeiras das instituições (ALBUQUERQUE; MEDINA; SILVA, 2017). Há também aplicações no setor jurídico, nas quais esses métodos podem estimar o valor de uma causa ganha utilizando dados de jurisprudência, entre outras aplicações (SOUSA, 2022).

Esses métodos buscam encontrar uma função $g(\mathbf{x})$ que preveja um rótulo real $\mathbf{Y} \in \Re$, com base nas covariáveis de entrada $\mathbf{X} = (X_1, X_2, \ldots, X_d)$. Para isso, parte-se frequentemente da suposição de que se dispõe de um conjunto de dados rotulados, $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. No entanto, em muitos cenários, a obtenção dos rótulos reais $Y$, também denominados como padrão ouro, pode ser dispendiosa, demorada ou até mesmo inviável. Nesses casos, é comum recorrer a especialistas para fornecer estimativas dos rótulos reais de cada observação, que seriam os

rótulos precisos. Contudo, esses especialistas podem ter diferentes níveis de *expertise*, levando a opiniões discrepantes. Esse cenário é frequentemente referido como dados com rótulos ruidosos (RAYKAR *et al.*, 2010).

A *expertise* é definida neste trabalho como a habilidade do especialista em fornecer previsões precisas. Exemplos em que essa *expertise* pode variar incluem, a avaliação da qualidade de um vinho, a interpretação de exames de ressonância magnética por diferentes especialistas, a divergência entre aparelhos que medem o mesmo fenômeno, julgamentos em concursos de beleza ou por júris populares. Esses cenários são tão comuns que, em muitos setores, como nas avaliações de ginástica artística e rítmica, opta-se por descartar a maior e a menor nota, com o objetivo de trazer mais estabilidade ao resultado final. Também há situações em que o ruído está no mesmo especialista, com rótulos observados em momentos distintos (KAHNEMAN; SIBONY; SUNSTEIN, 2021). Outro exemplo e ainda muito conhecido no Brasil é o cenário das correções das redações do ENEM. Atualmente dois especialistas corrigem a mesma redação. Caso as notas sejam muito distantes, sendo o critério adotado pela regra do concurso, um terceiro corretor é acionado e a nota final do candidato é a média aritmética das duas notas mais próximas. Esta informação pode ser vista na "Cartilha do Participante" disponibilizada pela INEP.

Tradicionalmente, em situações como essas, que usam diversos rotuladores, os rótulos precisos são estimados por uma métrica que agrege a opiniões individuais dos especialistas. Comumente, a adaptação da variável resposta é realizada pela média aritmética para problemas de regressão e pela frequência ou moda para problemas de classificação (XU; FRANK, 2004; GULSHAN *et al.*, 2016; CIOBOTARU *et al.*, 2022). Contudo, essa abordagem pode gerar viéses, especialmente quando os especialistas possuem diferentes níveis de *expertise*, visto que a média aritmética e a frequência atribuem o mesmo peso a todos os especialistas (YAN *et al.*, 2010; NARIMANZADEH *et al.*, 2023; CZEKAJ *et al.*, 2019).

Além disso, a *expertise* é apenas um dos fatores que podem influenciar as notas dos avaliadores. Existem também variáveis intrínsecas ou abstratas, muitas vezes impossíveis de capturar diretamente, que geram ruídos e justificam o nome dado a esses cenários. Exemplos incluem o cansaço do avaliador, seu estado de saúde mental e física, a ambiência física e cultural ou até mesmo uma relação inconsciente com o objeto de estudo. Esses fatores subjetivos podem afetar o julgamento, introduzindo variações nas notas que vão além do conhecimento técnico ou experiência dos especialistas.

Embora diversos autores tenham abordado o tema de rótulos ruidosos, a maioria das pesquisas não tem como foco central a "justiça na rotulagem", que, sob uma perspectiva estatística, busca estabilidade nos resultados. Este aspecto, particularmente relevante quando se utilizam seres humanos para rotular dados de treinamento destinados a automatizações, é um dos maiores desafios e potenciais ganhos desse tipo de abordagem.

Este trabalho traz avanços nestes cenários, pois apesar de já existir diversos estudos relacionados, a grande maioria desenvolveu apenas métodos de classificação, não exploraram a

diferença entre os especialistas de forma geral e por observação ou não consideraram ajustes com métodos não paramétricos, que com frequência, conseguem trabalhar de forma mais otimizada com covariáveis independentes que são funções das variáveis originais ou interações. Como exemplo de alguns métodos de classificação na literatura, temos Smyth *et al.* (1995), foi um dos primeiros trabalhos a utilizar *machine learning* para expor rótulos ruidosos. Para classificação de imagens, temos Izbicki e Stern (2013), que propôs, na presença de rótulos ruidosos, um método de trazer *sparsity* para o modelo provinda da função de perda desenvolvida; Frénay e Verleysen (2013) abordou diversos tipos de ruídos nos rótulos e seus efeitos; Jindal *et al.* (2019) e Algan e Ulusoy (2021), que usaram *deep learning* no método, ou ainda, Nguyen *et al.* (2021), que desenvolveu um método semelhante a este trabalho, pois considera a ponderação dos dados ruidosos, entretanto não se aprofundou em regressão, que é o foco desta dissertação, levando em consideração relações não lineares entre a variável dependente e as independentes. Temos também outros trabalhos, como Yan *et al.* (2010), Chittaranjan, Aran e Gatica-Perez (2011), Zheng *et al.* (2017), Tanno *et al.* (2019), Zhang (2022).

Alguns estudos tratam de regressão, como o trabalho de Raykar *et al.* (2010), que foi um motivador para o desenvolvimento deste estudo, pois incorporou a diferença entre os especialistas no modelo de regressão, desenvolvendo um método iterativo baseado no algoritmo *Expectation-Maximization* (EM), ou Rodrigues *et al.* (2017), que trata de maneira distinta os especialistas, considerando a presença de viés entre eles, temos Cooney e White (2021), que usa as divergências das opiniões dos especialistas sobre o tempo até o evento de interesse em modelos de análise de sobrevivência, utilizando essas discrepâncias para ponderar a função de verossimilhança, ou ainda, Wang, Sun e Fu (2022), que desenvolveu um método de penalizar o ruído detectado. Todavia, a maior parte, como estes exemplos, assumem um modelo linear nas covariáveis, que é um pressuposto muito forte perante a variedade de formas que as variáveis de entrada possam apresentar em relação à variável dependente.

Um método que abrange relações não lineares entre a variável dependente e as independentes foi apresentado por Xiao, Xiao e Eckert (2013), que introduziu o modelo utilizando um processo gaussiano. Contudo, este é um método muito específico e geralmente técnicas que utilizam *kernel* não possuem um bom comportamento quando há muitas covariáveis para serem incluídas no modelo e ainda não permite o uso de técnicas tradicionais como LASSO, Floresta Aleatória, Mínimos Quadrados, dentre outros.

Diante desses desafios, propomos os métodos WEAR *(Weighted Average Regression with Specialization)* e WEAR-INS *(Weighted Average Regression with Instance Specialization)*, uma abordagem versátil e eficaz para treinar a função $g$ em um contexto de regressão. Estes métodos consideram a diversidade entre rotuladores, sejam eles especialistas ou leigos, refletindo a realidade de plataformas modernas, como a *Amazon Mechanical Turk*, que coletam contribuições de pessoas com variados níveis de conhecimento. Nossa metodologia permite que o modelo se adapte tanto à *expertise* geral de cada rotulador quanto às especificidades de

cada observação, ponderando as diferentes contribuições dos avaliadores. Com isso, busca-se corrigir a instabilidade entre as opiniões, resultando em uma variável resposta mais robusta e próxima da realidade — um requisito essencial para alcançar a "justiça na rotulagem", conforme mencionado.

A metodologia WEAR-INS é composta por nove etapas principais: (1) Separar os dados em três partes: treino, validação e teste. (2) Ajusta-se uma função para cada especialista utilizando os dados de treinamento. Essa etapa identifica o modelo mais adequado às variáveis de entrada para aquele especialista. Intuitivamente, considerando que as variáveis selecionadas sejam apropriadas, e que o especialista rotule a instância de forma coerente, o ruído associado a essa observação tende a ser baixo. É com base nesse conceito intuitivo que o restante do algoritmo é desenvolvido. (3) Utilizamos esses modelos para prever a variável resposta nos dados de validação. Essa etapa tem como objetivo gerar estimativas da variável dependente, permitindo o cálculo do risco do modelo de cada especialista na etapa seguinte. Além disso, usar os dados de validação, busca mitigar o viés por otimismo, garantindo uma avaliação mais realista do desempenho dos modelos. (4) Utiliza-se o resultado do item (3) para calcular o risco do modelo, definido neste trabalho como o logaritmo natural do erro quadrático, representado pela notação $Z_{kj}$, visto em 2.2.2, sendo $k$ o número de observações no conjunto de validação e $j$ o número de especialistas. O uso do logaritmo natural do erro quadrado é uma alternativa vantajosa, pois ajuda a comprimir os efeitos de escala, tornando a avaliação mais robusta em relação a outliers. Outro benefício do logaritmo natural é de recuperar a escala original aplicando a exponencial, obtendo-se apenas dados positivos, que representam mais fielmente a variância que sempre será maior ou igual a zero. (5) Utiliza-se esse erro encontrado no item (4) como variável resposta para ajustar os dados de validação, ou seja, modelamos a função dos erros dos especialistas. O modelo resultante, aplicado aos dados de treinamento, gera a variância empírica da observação, conforme descrito por Wasserman (2006). Esta técnica é comumente empregada em modelos heterocedásticos. (6) Para retornar à escala original, aplicamos a exponencial da previsão, (7) Utilizar a informação do item (6) da forma inversa como o peso de cada observação. (8) Assim, calculamos uma média ponderada da variável resposta de cada especialista para cada instância, **resultando na estimativa do rótulo verdadeiro**. (9) Agora, finalizado todos os passos anteriores, é possível aplicar qualquer método de aprendizado de máquina supervisionado, o que confere a flexibilidade ao WEAR-INS, especialmente para a utilização de modelos não paramétricos, em cenários que a função verdadeira do modelo é desconhecida.

O artigo (SANTOS; IZBICKI, 2023), desenvolvido como parte da qualificação deste trabalho, no qual apresentamos o método WEAR, é uma versão simplificada do WEAR-INS. Essa simplificação desconsidera os pesos específicos por observação, adotando em seu lugar um peso geral aplicado a todas as instâncias. Com um número reduzido de etapas, portanto com tempo de processamento relativamente menor em comparação ao WEAR-INS, esse método foi descrito em detalhes ao longo desta dissertação, demonstrando-se que, em determinadas situações, ele é suficientemente eficaz para ajustar modelos de maneira adequada.

O presente trabalho está organizado em quatro capítulos. No primeiro, introduziu-se o tema central e o problema a ser abordado. O segundo capítulo descreve o método proposto, com ênfase no teorema que fundamenta o desenvolvimento da pesquisa. No terceiro capítulo, são apresentadas aplicações do método em dados simulados e reais, demonstrando empiricamente que seu desempenho supera o de abordagens tradicionais, como a média aritmética. Por fim, o quarto capítulo traz as conclusões do estudo, juntamente com sugestões para futuras linhas de pesquisa e possíveis aprimoramentos.

## 1.1 Objetivos

O objetivo deste trabalho é desenvolver um método de regressão capaz de prever o rótulo verdadeiro em novas observações, mesmo na ausência desse rótulo durante a fase de treinamento, utilizando as divergentes opiniões de especialistas na presença de ruídos.

### *1.1.1 Objetivos Específicos*

- Propor um método de estimar a *expertise* dos especialistas de forma geral;
- Propor uma técnica para quantificar a discordância entre as avaliações dos especialistas;
- Desenvolver um estimador de regressão que utilize a média ponderada das opiniões dos especialistas para cada observação, oferecendo uma alternativa aos métodos tradicionais, como a média aritmética.

CAPÍTULO

# 2

# DESENVOLVIMENTO

No aprendizado supervisionado, o principal objetivo é prever uma variável dependente com base em variáveis explicativas. Para isso, é essencial obter o rótulo real para o treinamento do modelo. No entanto, em algumas situações, torna-se inviável obter esse rótulo. Nesses casos, recorre-se às opiniões de diversos especialistas, que podem ter *expertises* distintas, resultando em opiniões divergentes sobre o que está sendo observado. Esse tipo de cenário é comumente chamado de dados com rótulos ruidosos. Tradicionalmente, usa-se a média aritmética dessas opiniões como estimador do rótulo real, entretanto, isto nem sempre é correto, dado que, na presença de um especialista muito divergente dos outros, considerar o mesmo peso para todos os especialistas pode resultar em uma estimativa incorreta.

Este estudo resume-se em desenvolver um método de estimar a *expertise* de cada especialista a fim de calcular à média ponderada do rótulo, o que é uma alternativa a média aritmética. A técnica que usa a média ponderada é superior à técnica que usa a média aritmética, pois pondera a opinião do especialista com o peso baseado na sua *expertise*.

Para encontrar o estimador para o rótulo real e posteriormente desenvolver o modelo que prevê este rótulo a partir de novas observações, precisa-se ajustar modelos que utilizam de técnicas paramétricas ou não paramétricas. Em cenários que o número de covariáveis é alto, corre-se o risco de obtermos modelos paramétricos inadequados, pois este tipo de modelo exige uma forma explícita da função $g(\mathbf{x})$. Dado que as covariáveis podem ser funções das variáveis originais e há o risco da forma explícita de $r(\mathbf{x})$ não capturar isto, o método não paramétrico torna-se mais flexível. Este trabalho permite comparações entre essas duas técnicas.

As vantagens deste trabalho estão no desenvolvimento de uma técnica, baseada no teorema que foi desenvolvido, para estimar a *expertise* geral e por observação do especialista e incorporar esta informação no modelo, podendo utilizar técnicas não paramétricas.

Incluir a *expertise* e ainda trabalhar com técnicas não paramétricas ocasiona em riscos preditivos menores em relação ao método que considera o mesmo peso para todos os especialistas.

## 2.1 Motivação

Nesta seção introduzimos o teorema que motivou o método deste trabalho.

Tem-se que uma das funções de risco mais conhecidas para verificar a eficiência do modelo preditivo é o erro quadrático médio (EQM). O EQM de um estimador $\hat{\theta}$ é definido como:

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2], \tag{2.1}$$

que pode ser reescrito em função da variância e um viés:

$$EQM(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \tag{2.2}$$

$$= Var(\hat{\theta}) + Viés^2. \tag{2.3}$$

Portanto, o EQM de um estimador não viesado é a própria variância.

Agora, suponha que exista a opinião de um especialista e usa-se o EQM para calcular o erro da opinião do especialista em relação ao rótulo real. Então, o EQM não viesado neste caso é a própria variância do especialista. O teorema 1 descreve melhor o problema.

**Teorema 1.** Seja $Y_j$ a opinião do $j$-ésimo especialista, $Y$ o rótulo real, com $Y_j$ e $Y$ independentes, $\mathbf{X}$ o vetor de covariáveis fixo, $w_j$ o peso do $j$-ésimo especialista, sendo $\sum_{j=1}^{J} w_j = 1$. Suponha que

$$E[Y_j|\mathbf{X}] = E[Y|\mathbf{X}].$$

A solução para

$$\min_{w_1,\ldots,w_{J-1}} E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right)^2 \middle| \mathbf{X} = \mathbf{x}\right], \tag{2.4}$$

é utilizar $w_j$ como,

$$\frac{\frac{1}{Var[Y_j|\mathbf{X}=\mathbf{x}]}}{\sum_{j=1}^{J} \frac{1}{Var[Y_j|\mathbf{X}=\mathbf{x}]}}.$$

***Prova:*** Considere a variância em função da esperança, de tal forma que:

$$Var[X] = E[X^2] - (E[X])^2.$$

Logo,

$$\begin{aligned} E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right)^2 \middle| \mathbf{X}=\mathbf{x}\right] = \\ Var\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right) \middle| \mathbf{X}=\mathbf{x}\right] + E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right) \middle| \mathbf{X}=\mathbf{x}\right]^2. \end{aligned} \tag{2.5}$$

Assumindo as hipóteses do teorema, tem-se que o termo:

$$E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right) \middle| \mathbf{X}=\mathbf{x}\right]^2$$

da equação (2.5) é zero. Então,

$$E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right)^2 \middle| \mathbf{X}=\mathbf{x}\right] = Var\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right) \middle| \mathbf{X}=\mathbf{x}\right]. \tag{2.6}$$

Note que o lado esquerdo da equação (2.6) é a função de risco.

Considerando a independência entre $Y_j$ e $Y$ tem-se,

$$Var\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right) \middle| \mathbf{X}=\mathbf{x}\right] = Var\left[\left(\sum_{j=1}^{J} w_j Y_j\right) \middle| \mathbf{X}=\mathbf{x}\right] + Var[Y \mid \mathbf{X}=\mathbf{x}]. \tag{2.7}$$

Dado que $w_j$ é constante, a expressão se simplifica para:

$$\begin{aligned} Var\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right) \middle| \mathbf{X}\right] = \\ \mathbf{x} = \sum_{j=1}^{J} \left(w_j^2 Var\left[Y_j \middle| \mathbf{X}=\mathbf{x}\right]\right) + Var[Y \mid \mathbf{X}=\mathbf{x}] = \\ = w_1^2 Var[Y_1 \mid \mathbf{X}=\mathbf{x}] + w_2^2 Var[Y_2 \mid \mathbf{X}=x] + w_3^2 Var[Y_3 \mid \mathbf{X}=\mathbf{x}] + \cdots + \\ \left(1 - \sum_{1}^{j-1} w_j\right)^2 Var[Y_j \mid \mathbf{X}=\mathbf{x}] + Var[Y \mid \mathbf{X}=\mathbf{x}]. \end{aligned} \tag{2.8}$$

Para atingir o objetivo de minimizar a função descrita, devemos derivar a expressão 2.8 em função de $w_1$ e igualar a zero. Então:

$$\begin{aligned} \frac{d}{dw_1}\left(\sum_{j=1}^{J} w_j^2 \operatorname{Var}\left[Y_j \mid \mathbf{X}=\mathbf{x}\right] + \operatorname{Var}[Y \mid \mathbf{X}=\mathbf{x}]\right) = \\ 2w_1 \operatorname{Var}[Y_1 \mid \mathbf{X}=\mathbf{x}] + \left(1 - \sum_{j=1}^{J-1} w_j\right)(-1) \operatorname{Var}[Y_j \mid \mathbf{X}=\mathbf{x}] = 0. \end{aligned} \tag{2.9}$$

Portanto,

$$\frac{w_j}{w_1} = \frac{Var[Y_1 \mid \mathbf{X} = \mathbf{x}]}{Var[Y_j \mid \mathbf{X} = \mathbf{x}]}. \tag{2.10}$$

Considerando $w_1 = K/Var[Y_1 \mid \mathbf{X} = \mathbf{x}]$ e substituindo em (2.10), tem-se que,

$$w_j = K/Var[Y_j \mid \mathbf{X} = \mathbf{x}]$$

para qualquer $j$. Conforme $\sum_{j=1}^{J} w_j = 1$, então

$$\sum_{j=1}^{J} w_j = \sum_{j=1}^{J} \frac{K}{Var[Y_j \mid \mathbf{X} = \mathbf{x}]} = 1, \tag{2.11}$$

sob $K$ uma constante, tem-se de (2.16) que,

$$K \sum_{j=1}^{J} \frac{1}{Var[Y_j \mid \mathbf{X} = \mathbf{x}]} = 1. \tag{2.12}$$

Portanto, dado $K = \frac{1}{\sum_{j=1}^{J} \frac{1}{Var[Y_j|\mathbf{X}=\mathbf{x}]}}$, o valor de $w_j$ que minimiza a variância dos erros é:

$$\frac{\frac{1}{\sum_{j=1}^{J} \frac{1}{Var[Y_j|\mathbf{X}=\mathbf{x}]}}}{Var[Y_j \mid \mathbf{X} = \mathbf{x}]}. \tag{2.13}$$

□

A equação (2.13) apresenta o valor do $w_j$ dependendo das observações, o que representa situações da vida real, pois, em alguns cenários, algumas observações são mais fáceis de serem rotuladas do que outras, entretanto, segue de forma direta do teorema que também é possível estimar $w_j$ não dependendo das observações. Com este propósito, segue o corolário.

**Corolário 1.1.** Suponha que as variâncias, $Var[Y_j \mid \mathbf{X} = \mathbf{x}]$, sejam constantes para todo $\mathbf{x}$. Então, o valor de $w_j$ ótimo, não dependerá das observações.

O teorema 1 está associado ao método WEAR-INS, enquanto o corolário 1.1 derivado do teorema se aplica ao método WEAR, conforme discutido no Capítulo 1.

É importante destacar que o pressuposto do teorema,

$$E[Y_j|\mathbf{X}] = E[Y|\mathbf{X}],$$

é essencial para sua validade. Por exemplo, considere um especialista com uma variância próxima de zero: se o viés for elevado, o peso estimado não refletirá adequadamente a verdadeira *expertise* do especialista. Essa condição ressalta a importância de garantir que o viés seja mínimo para que o método seja eficaz.

Para flexibilizar esta situação, segue o Teorema 2.

**Teorema 2.** Seja $Y_j$ a opinião do $j$-ésimo especialista, $Y$ o rótulo real, com $Y_j$ e $Y$ independentes, $\mathbf{X}$ o vetor de covariáveis fixo, $w_j$ o peso do $j$-ésimo especialista, sendo $\sum_{j=1}^{J} w_j = 1$. Suponha que

$$E[Y_j|\mathbf{X}] \neq E[Y|\mathbf{X}].$$

A solução para

$$\min_{w_1,\ldots,w_{J-1}} E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right)^2 \middle| \mathbf{X} = \mathbf{x}\right] \tag{2.14}$$

é utilizar $w_j$ como,

$$w_j = \frac{\frac{1}{\text{Var}[Y_j|\mathbf{X}=\mathbf{x}]+\text{Viés}_j^2}}{\sum_{k=1}^{J} \frac{1}{\text{Var}[Y_k|\mathbf{X}=\mathbf{x}]+\text{Viés}_k^2}}.$$

***Prova:*** Considere que: $Y$ é a variável resposta real, $Y_j$ é o rótulo fornecido pelo especialista $j$, $Y$ e $Y_j$ independentes, $E[\mathbf{Y} \mid \mathbf{X}=\mathbf{x}]$ é a esperança do valor real $Y$ dado $X = x$, $E[\mathbf{Y_j}|\mathbf{X} = \mathbf{x}]$ é a esperança do rótulo do especialista $j$ dado $X = x$ e $\text{Viés}_j = E[Y_j|X = x] - E[Y|\ \mathbf{X}=\mathbf{x}]$ representa o viés do especialista $j$ em relação ao valor real e sendo o objetivo minimizar o erro quadrático esperado:

$$E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right)^2 \mid \mathbf{X}=\mathbf{x}\right],$$

sujeito a condição de que $\sum_{j=1}^{J} w_j = 1$, temos:

$$E\left[\left(\sum_{j=1}^{J} w_j Y_j - Y\right)^2 \mid \mathbf{X}=\mathbf{x}\right] = \text{Var}\left[\sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X}=\mathbf{x}\right] + \left(E\left[\sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X}=\mathbf{x}\right]\right)^2. \tag{2.16}$$

O primeiro termo da expressão (2.16) após o sinal de igual é a variância, que pode ser escrita como:

$$\text{Var}\left[\sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X}=\mathbf{x}\right] = \text{Var}\left[\sum_{j=1}^{J} w_j Y_j \mid \mathbf{X}=\mathbf{x}\right] + \text{Var}[Y \mid \mathbf{X}=\mathbf{x}].$$

Dado que os especialistas são independentes entre si e de $Y$, temos:

$$\text{Var}\left[\sum_{j=1}^{J} w_j Y_j \mid \mathbf{X}=\mathbf{x}\right] = \sum_{j=1}^{J} w_j^2 \text{Var}[Y_j \mid \mathbf{X}=\mathbf{x}].$$

Portanto:

$$\text{Var}\left[\sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X}=\mathbf{x}\right] = \sum_{j=1}^{J} w_j^2 \text{Var}[Y_j \mid \mathbf{X}=\mathbf{x}] + \text{Var}[Y \mid \mathbf{X}=\mathbf{x}].$$

O segundo termo da expressão (2.16) após o sinal de igual é o viés ao quadrado, dado por:

$$\left( E\left[ \sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X} = \mathbf{x} \right] \right)^2.$$

Devido $E[Y_j|X=x] = E[Y|X=x]+$ o viés do $j$-ésimo especialista, temos:

$$E\left[ \sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X} = \mathbf{x} \right] = \sum_{j=1}^{J} w_j E[Y_j|X=x] - E[Y|X=x].$$

Substituindo $E[Y_j|X=x] = E[Y|X=x] + \text{Viés}_j$, obtemos:

$$E\left[ \sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X} = \mathbf{x} \right] = \sum_{j=1}^{J} w_j \left( E[Y|X=x] + \text{Viés}_j \right) - E[Y|X=x].$$

Cancelando os termos $E[Y|X=x]$:

$$E\left[ \sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X} = \mathbf{x} \right] = \sum_{j=1}^{J} w_j \text{Viés}_j.$$

Então, o viés quadrático é:

$$\left( E\left[ \sum_{j=1}^{J} w_j Y_j - Y \mid \mathbf{X} = \mathbf{x} \right] \right)^2 = \left( \sum_{j=1}^{J} w_j \text{Viés}_j \right)^2.$$

Portanto, combinando os dois, o erro quadrático esperado é:

$$E\left[ \left( \sum_{j=1}^{J} w_j Y_j - Y \right)^2 \mid \mathbf{X} = \mathbf{x} \right] = \sum_{j=1}^{J} w_j^2 \text{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] + \text{Var}[Y \mid \mathbf{X} = \mathbf{x}] + \left( \sum_{j=1}^{J} w_j \text{Viés}_j \right)^2.$$

Este erro precisa ser minimizado em relação aos pesos $w_j$, sob a restrição $\sum_{j=1}^{J} w_j = 1$. O problema de minimização é:

$$\min_{w_1,\ldots,w_J} \sum_{j=1}^{J} w_j^2 \text{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] + \left( \sum_{j=1}^{J} w_j \text{Viés}_j \right)^2,$$

sujeito a:

$$\sum_{j=1}^{J} w_j = 1.$$

Substituímos $w_J$ na função por $w_J = 1 - \sum_{j=1}^{J-1} w_j$. Então, o termo com a variância segue a forma:

$$\sum_{j=1}^{J} w_j^2 \text{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] = \sum_{j=1}^{J-1} w_j^2 \text{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] + \left( 1 - \sum_{j=1}^{J-1} w_j \right)^2 \text{Var}[Y_J \mid \mathbf{X} = \mathbf{x}].$$

Agora o termo de viés:

$$\left(\sum_{j=1}^{J} w_j \text{Viés}_j\right)^2 = \left(\sum_{j=1}^{J-1} w_j \text{Viés}_j + \left(1 - \sum_{j=1}^{J-1} w_j\right) \text{Viés}_J\right)^2.$$

Expandindo o quadrado:

$$\left(\sum_{j=1}^{J-1} w_j \text{Viés}_j\right)^2 + 2\left(\sum_{j=1}^{J-1} w_j \text{Viés}_j\right)\left(1 - \sum_{j=1}^{J-1} w_j\right) \text{Viés}_J + \left(1 - \sum_{j=1}^{J-1} w_j\right)^2 \text{Viés}_J^2.$$

Agora, basta minimizar a função, derivamos em relação a cada $w_j$.

A derivada parcial em relação a $w_j$ é:

$$\frac{\partial}{\partial w_j}\left[\sum_{j=1}^{J-1} w_j^2 \text{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] + \left(1 - \sum_{j=1}^{J-1} w_j\right)^2 \text{Var}[Y_J \mid \mathbf{X} = \mathbf{x}]\right]$$

+

$$\frac{\partial}{\partial w_j}\left[\left(\sum_{j=1}^{J-1} w_j \text{Viés}_j + \left(1 - \sum_{j=1}^{J-1} w_j\right) \text{Viés}_J\right)^2\right].$$

Após expandir e rearranjar conforme visto no Teorema 1, temos:

$$w_j \propto \frac{K}{\text{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] + \text{Viés}_j^2}.$$

Portanto, temos:

$$w_j = \frac{\frac{1}{\text{Var}[Y_j|\mathbf{X}=\mathbf{x}]+\text{Viés}_j^2}}{\sum_{k=1}^{J} \frac{1}{\text{Var}[Y_k|\mathbf{X}=\mathbf{x}]+\text{Viés}_k^2}}.$$

□

Essa solução, ao reconhecer que a esperança condicional de Y (o valor real) pode divergir da esperança dos rótulos fornecidos pelos especialistas, introduz uma flexibilidade importante. Contudo, o cenário real impõe um desafio crucial: a inexistência do valor Y verdadeiro. Para implementar essa abordagem em um algoritmo prático, seria imprescindível definir uma variável substituta para o padrão ouro. Uma possível solução seria utilizar a variância estimada de cada especialista como um indicador de sua precisão, presumindo que o especialista com a menor variância oferece a estimativa mais confiável de $Y$.

Contudo, o método atualmente utilizado para estimar o melhor especialista é baseado no Teorema 1, que pressupõe especialistas não viesados em relação ao $Y$ real. Assim, uma alternativa seria desenvolver um estimador para o viés ou considerar outro estimador para $Y$, não vinculado ao Teorema 1.

Assim, o algoritmo proposto e os experimentos apresentados baseiam-se unicamente no Teorema 1.

## 2.2 Modelo Ponderado Proposto

Baseado no Teorema 1, foi descrito o passo a passo do algoritmo para o método proposto, que visa estimar as *expertises* dos especialistas por observação, a fim de calcular a média ponderada de suas opiniões como estimador para o rótulo real. Em sequência, como obter um modelo para prever o rótulo real a partir de novas observações.

Considere que toda a amostra é i.i.d (independente e identicamente distribuída) tal que, $(X_1, Y_1), \ldots, (X_m, Y_m) \sim (X, Y)$, com $m$ o número de observações no conjunto treino, $h$ o número de observações no conjunto de validação e $J$ o número de especialistas.

### 2.2.1 WEAR: Regressão Média Ponderada pela Expertise

O método WEAR é uma variação do WEAR-INS, pois provém do corolário 1.1 do Teorema 1 e possui o objetivo de estimar um peso geral para cada especialista e posteriormente calcular uma média ponderada dos rótulos. O algoritmo pode ser reduzido em comparação ao WEAR-INS e segue da forma:

1. Separar os dados em treinamento, validação e teste.
2. Estimar $r_j(\mathbf{x}) E[Y_j|\mathbf{X} = \mathbf{x}]$, a regressão do rótulo fornecido pelo $j$-ésimo especialista em $\mathbf{X}$, usando o conjunto de treinamento, sendo $\widehat{r}_j$ essa estimativa. Neste item qualquer modelo supervisionado pode ser explorado.
3. Aplicar a predição nos dados de validação.
4. Calcular o risco preditivo de cada especialista:

$$\widehat{R}_j := \frac{\sum_{j=1}^{J} \left(Y_{h,j} - \widehat{r}_j(\mathbf{x}_h)\right)^2}{h},$$

   com $h$ variando da primeira até a última observação dos dados de validação e $h$ sendo a quantidade de observações nos dados de validação.
5. Calcular o inverso do valor encontrado no item anterior.
6. Usar o valor encontrado no item anterior como o peso da informação fornecida por cada especialista, conforme o corolário:

$$\widehat{w}'_j := \frac{\widehat{R}_j^{-1}}{\sum_{j=1}^{J} \widehat{R}_j^{-1}}.$$

   Calcular $\overline{Y}^w$, a média ponderada dos especialistas nos dados de treinamento:

$$\overline{Y}^w = \sum_{j=1}^{J} \widehat{w}_j Y_{m,j} \tag{2.17}$$

7. Estimar a função de regressão verdadeira, $r(\mathbf{x}) := E[Y|\mathbf{x}]$, fazendo a regressão de $\overline{Y}^w$ em $\mathbf{x}_h$ usando a amostra de treinamento, podendo usufruir de qualquer método supervisionado.

### 2.2.2 WEAR-INS: Regressão Média Ponderada pela Expertise por Observação

Nosso método utiliza o Teorema 1 combinado com a forma de estimar a variância de forma não paramétrica, consistindo nos seguintes passos:

1. Separar o conjunto de dados em conjuntos de treinamento, validação e teste.

2. Estimar $r_j(\mathbf{x})E[Y_j|\mathbf{X} = \mathbf{x}]$, a regressão do rótulo fornecido pelo $j$-ésimo especialista em $\mathbf{X}$, usando o conjunto de treinamento, sendo $\widehat{r}_j$ essa estimativa. Neste item qualquer método supervisionado pode ser explorado.

3. Aplicar a predição nos dados de validação.

4. Definir a estimativa $R_j$ do Teorema 1, porém, em forma logarítmica da base *euler* e por observação:

$$\widehat{Z}_{hj} := ln\left(Y_{h,j} - \widehat{r}_j(\mathbf{x}_h)\right)^2,$$

   com $h$ variando da primeira até a última observação dos dados de validação.

5. Considerar $\widehat{Z}_{hj}$, encontrado no item anterior, como a variável resposta e ajustar, para cada especialista, usando as variáveis de entrada dos dados de validação. Esse modelo é a função que estima os erros empiricamente:

$$\hat{e}_j(x_h) := E[\widehat{\mathbf{Z}}_{hj}|\mathbf{x_h}].$$

6. Usar o modelo encontrado no item acima para aplicar a predição nos dados de treinamento. Esse vetor precisa retornar à escala original, e por isso, é aplicado o exponencial: $exp(\hat{e}_j(x_m))$, com $m$ variando da primeira até a última observação dos dados de treinamento. Os vetores encontrados são os erros estimados por cada especialista e observação.

   Observação: O logaritmo não está definido no zero. Portanto, na aplicação real, é necessário ajustar os dados quando houver erros iguais a zero. Nestas situações sugerimos escolher um valor próximo de zero que se aproxime da média das opiniões do especialista quando aplicadas a exponencial para retornar a escala original. Por exemplo, ao retornar os valores a escala original a média, sem considerar o zero, se aproximava de -16, então escolheu-se o valor 0.0000001 como substituição do zero.

7. Calcular o inverso de cada vetor para cada instância.

8. Usar o valor encontrado no item anterior como o peso da informação fornecida por cada especialista, conforme o Teorema 1. Isso é necessário porque queremos considerar o menor peso para a observação com maior risco:

$$\frac{1}{exp(\hat{e}_j(x_m))} = \hat{w}_{m,j},$$

sendo $w'_{j,m}$ o peso do $j$-ésimo especialista na $m$-ésima instância. Aproximar os pesos ótimos do Teorema 1:

$$\widehat{w}'_{m,j} := \frac{\widehat{R}^{-1}_{m,j}}{\sum_{j=1}^{J} \widehat{R}^{-1}_{m,j}}.$$

Calcular $\overline{Y}^{w}m$, a média ponderada dos especialistas para a instância no $h$-ésimo ponto da amostra de treinamento, usando

$$\overline{Y}^{w}m = \sum_{m=1}^{m} \widehat{w}_{m,j} Y_{m,j} \tag{2.18}$$

9. Estimar a função de regressão verdadeira, $r(\mathbf{x}) := E[Y|\mathbf{x}]$, fazendo a regressão de $\overline{Y}^{w}_{m}$ em $\mathbf{x}_h$ usando a amostra de treinamento, podendo usufruir de qualquer método supervisionado.

Ambos os métodos oferecem grande flexibilidade, pois podemos utilizar qualquer algoritmo de aprendizado de máquina nos passos 2, 5 e 9 no método WEAR-INS e nos passos 2 e 7 no método WEAR. Por exemplo, se conhecermos a forma exata da equação, podemos usar ferramentas paramétricas; caso contrário, podemos confiar em métodos não paramétricos. Ou seja, dependendo dos dados disponíveis e de suas características, existem algoritmos mais adequados para um bom desempenho. Poderíamos, em qualquer caso, testar muitos deles, mas o processamento se tornaria inviável. O ideal é conhecer o objetivo e como a base de dados está formada para entender quais métodos seriam mais adequados para testar.

Na próxima seção, demonstraremos o método com dados simulados e reais, com especialistas fictícios em ambas as situações. Foi possível mostrar que os métodos propostos alcançam alto desempenho em comparação com outros métodos tradicionais, como a média aritmética entre os especialistas.

CAPÍTULO

# 3

# EXPERIMENTOS

Nesta seção, apresentamos a validação empírica dos métodos propostos, através de dois conjuntos de experimentos: simulações controladas e análise de dados reais. O objetivo é comparar o desempenho do WEAR e WEAR-INS com os métodos de referência, utilizando diferentes cenários e o EQM como métrica de avaliação.

Enquanto os cenários totalmente simulados permitem alto controle sobre os parâmetros, os dados reais trazem desafios práticos, como ruído e vieses inerentes, fornecendo uma visão complementar da eficácia do método proposto.

Para ajustar o modelo e validar o método proposto, foram utilizadas várias técnicas, incluindo Regressão Linear, Árvores de Decisão, Florestas Aleatórias e LASSO. Optamos por comparar algoritmos paramétricos e não paramétricos, demonstrando que é possível aplicar diferentes técnicas dentro do método proposto.

Além da comparação entre algoritmos, diferentes métodos de estimativa para a variável resposta foram avaliados em relação às metodologias propostas (WEAR-INS e WEAR), como o método desenvolvido por Raykar *et al.* (2010), a média aritmética das opiniões dos especialistas e, finalmente, o padrão ouro (rótulo preciso), que compara o método desenvolvido neste trabalho para verificar sua capacidade de representar adequadamente a resposta verdadeira.

Na seção 3.1 ou primeira aplicação, simulamos quatro cenários distintos, com detalhes descritos em cada subseção. O resultado final das comparações entre os métodos é mostrado na Tabela 5, que apresenta o EQM e o erro padrão (entre parênteses), fornecendo uma base sólida para avaliar a significância dos métodos ao verificar a sobreposição dos intervalos de confiança.

Para aumentar a precisão dos resultados, utilizou-se o Método de Monte Carlo, fundamentado na Lei dos Grandes Números, simulando cada cenário 50 vezes e calculando as métricas com base nas simulações, com erro padrão incluído para fornecer intervalos de confiança mais robustos.

A segunda aplicação utilizou dados reais, com especialistas simulados com base na variável resposta, conforme descrito na subseção 3.2. Essa abordagem segue o teorema proposto, onde $Y_j$ denota a opinião do $j$-ésimo especialista ($j = 1, \ldots, J$), $Y$ representa o rótulo verdadeiro, e $\mathbf{x}$ é o vetor de covariáveis. Para todos os especialistas, assume-se que $E[Y_j|\mathbf{x}] = E[Y|\mathbf{x}]$.

Assim, na maior parte dos experimentos, decidiu-se que a opinião de cada especialista seria simulada com base na variável resposta real, adicionando-se uma distribuição normal para cada observação, com média zero e variância igual ao valor da covariável escolhida para representar o ruído do especialista, conforme a equação:

$$Y_j = Y + N(0, |x_{ki}|), \tag{3.1}$$

onde $Y$ é a resposta real, $Y_j$ é a opinião do $j$-ésimo especialista, $k$ é o índice da covariável no modelo, e $i$ é o índice da instância da covariável.

Esse cenário foi aplicado a 3 diferentes conjuntos de dados.

Os resultados estão na Tabela 6. Ela apresenta o EQM de cada método para cada conjunto de dados, com o erro padrão mostrado entre parênteses ao lado dos resultados.

Utilizou-se o software R <https://www.r-project.org/> para a análise, e, em todos os exemplos, os hiperparâmetros foram configurados com valores *default*. Portanto, não foram feitos estudos para a otimização dos hiperparâmetros. Esta escolha vem de que o principal objetivo é a comparação entre os métodos. Por isso, é necessário manter tudo o mais constante.

Os resultados desses experimentos estão organizados em blocos nas Tabelas 5 e 6. O primeiro e o segundo bloco inclui todos os algoritmos testados com as variáveis respostas desenvolvidas neste trabalho, o método WEAR-INS (Método de Regressão Ponderada por Instância) e o WEAR (Método de Regressão Ponderada). O terceiro utiliza o algoritmo de Raykar *et al.* (2010). O quarto bloco utiliza a média aritmética como linha de base tradicional, e o quinto utiliza o rótulo verdadeiro ($Y$ real) como referência para avaliar o desempenho do método proposto.

Para avaliar se o algoritmo pode estimar a variância por observação, será apresentado um boxplot da variação de cada especialista. O boxplot é uma ferramenta visual útil para capturar a variabilidade entre as opiniões dos especialistas em cada observação. Nos cenários simulados, ele permitirá uma análise clara da variabilidade 1 iteração das reamostragens, enquanto nos dados reais, o boxplot representará as variâncias empíricas estimadas, proporcionando *insights* valiosos sobre a dispersão das opiniões.

Além disso, serão apresentadas Tabelas contendo os pesos gerais estimados tanto pelo método WEAR quanto pelo algoritmo proposto por Raykar *et al.* (2010).

A divisão dos dados seguiu a proporção de 60% para treinamento, 10% para validação e 30% para teste. Essa separação foi definida de forma parcialmente arbitrária, considerando

que o conjunto de treinamento demandava maior volume de dados para o aprendizado e que o conjunto de teste necessitava de um volume superior ao de validação, com o objetivo de garantir intervalos de confiança suficientemente distintos na maioria dos experimentos. Não foram realizados estudos para avaliar se os métodos propostos apresentariam melhor desempenho caso o conjunto de validação tivesse um tamanho equivalente ao de treinamento, especialmente considerando que também é ajustado o modelo de risco dos especialistas.

A métrica de desempenho utilizada foi o EQM, acompanhado do erro padrão para a análise dos intervalos de confiança. Os menores valores de EQM de cada conjunto estão destacados em negrito.

Os resultados dos experimentos confirmam que os métodos WEAR-INS e WEAR demonstram desempenho superior em cenários com maior variância entre as opiniões dos especialistas, com o WEAR-INS alcançando resultados ainda melhores. Portanto, entende-se que o método WEAR-INS conseguiu estimar a resposta verdadeira de maneira mais precisa.

## 3.1 Dados Simulados

### 3.1.1 Dados 1

O conjunto de dados 1 consiste em 4 especialistas e 5 variáveis independentes, geradas a partir de distribuições normais. Uma das variáveis ($x_0$) foi gerada com média zero e variância 1, enquanto as demais apresentam média 5 e variâncias distintas que refletem a função do peso de cada especialista: 1, 25, 125 e 15625, respectivamente. Dessa forma, o especialista 1 tem menor variância em relação aos demais, com a variância aumentando progressivamente do especialista 1 para o especialista 4. O verdadeiro valor de $y$ foi construído somando-se os quadrados das covariáveis mais um termo de erro, proveniente de uma distribuição normal, conforme a equação:

$$y = -0,25x_0^2 + 0,06x_1^2 + 0,01x_2^2 + 0,003x_3^2 + 0,0003x_4^2 + \varepsilon, \tag{3.2}$$

onde $\varepsilon$ segue uma distribuição normal com média 0 e variância 9.

Essa configuração foi escolhida para simular um cenário no qual a relação entre $y$ e as variáveis preditoras é não linear, permitindo avaliar o desempenho comparativo entre métodos paramétricos e não paramétricos.

Os coeficientes foram definidos de forma arbitrária, de modo que, quanto maior a variância do especialista, menor o coeficiente associado.

Na coluna Dados 1 da Tabela 5, os resultados mostram que, ao considerar o verdadeiro $y$, o algoritmo de Floresta Aleatória apresentou o menor EQM, com um valor em torno de 10. Esse padrão de desempenho foi também observado nos métodos WEAR-INS e WEAR, sugerindo que ambos são excelentes alternativas na ausência do padrão-ouro. Contudo, o método WEAR-INS, que incorpora a variância por observação, demonstrou um desempenho superior ao WEAR, evidenciado pelos intervalos de confiança não sobrepostos.

O método proposto por Raykar *et al.* (2010) obteve um desempenho similar ao dos métodos paramétricos. Isso ocorre porque ele utiliza uma forma fechada da função no primeiro passo (passo E) do algoritmo EM, embora ele tenha estimado de forma coerente os pesos para cada especialista, sendo o maior peso para o especialista com menor variância, como visto na Tabela 1.

Como esperado, o algoritmo de Floresta Aleatória apresentou o melhor desempenho, uma vez que o valor de $y$ foi gerado para refletir uma relação não linear com as variáveis independentes. Para avaliar a consistência desse desempenho na estimativa das variâncias por observação, foi realizado um análise do boxplot das médias das variâncias estimadas para cada especialista, com base nas reamostragens. A Figura 1 revela que todos os algoritmos capturaram adequadamente as diferenças entre os especialistas. Além disso, a Tabela 1 mostra que os pesos gerais também foram estimados com base na variância fornecida, sendo o modelo de Floresta Aleatória o que

obteve os melhores resultados. Dessa forma, ambos os métodos propostos demonstraram eficácia na estimativa da variância dos especialistas.

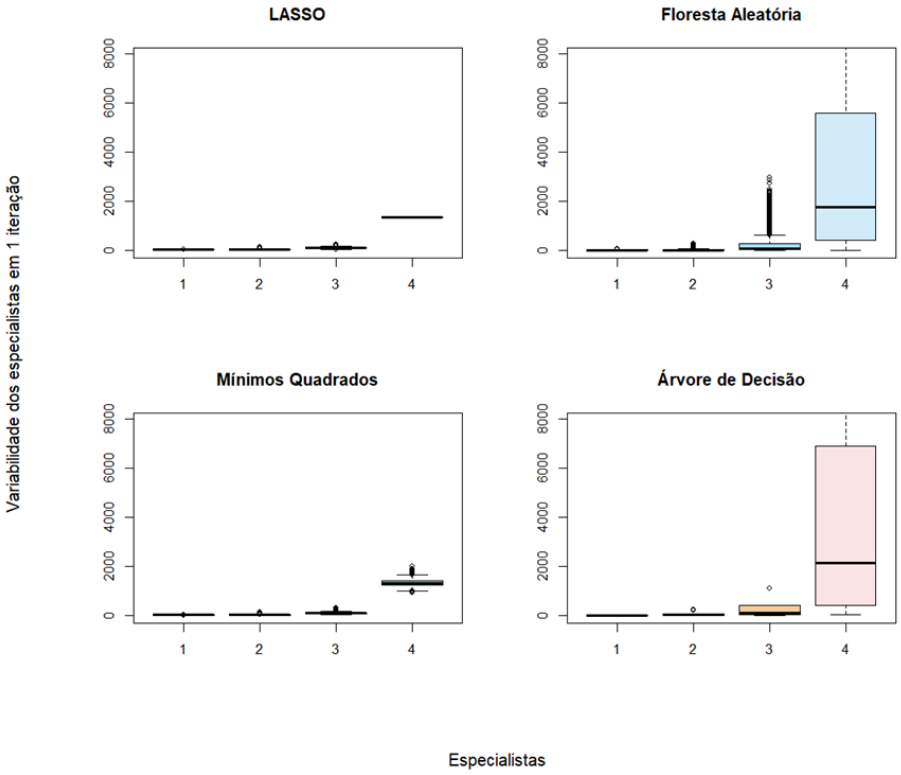


Figura 1 – Boxplot das variabilidades dos especialistas - Dados 1 - Simulados

| Abordagem | Modelo | Especialista 1 | Especialista 2 | Especialista 3 | Especialista 4 |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,6449 | 0,6322 | 0,4917 | 0,3989 |
| WEAR | Floresta | 38,7638 | 64,3699 | 685,1851 | 16611,0061 |
| | Árvore | 45,0529 | 71,4088 | 667,5519 | 15939,3711 |
| | LASSO | 88,5801 | 113,3042 | 688,1223 | 15947,4274 |
| | Mínimos Quadrados | 88,5908 | 113,3098 | 687,9490 | 15965,2647 |

Tabela 1 – Dados 1 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

### 3.1.2 Dados 2

O conjunto de dados 2 foi gerado utilizando cinco variáveis independentes, sendo quatro provenientes da distribuição normal e uma da distribuição exponencial.

A variável resposta $y$ foi modelada com uma relação linear em função das variáveis independentes, conforme a equação:

$$y = -0,5x_1 + 0,5x_2 + 0,5x_3 + 0,1x_4 + 0,1x_5 + \varepsilon, \tag{3.3}$$

na qual $\varepsilon$ é uma variável aleatória proveniente de uma distribuição normal com média 0 e variância 1. As variáveis $x_1, x_2, x_3$ e $x_5$ seguem distribuições normais com média 0 e variâncias de 0,01; 0,0625; 1 e 25, respectivamente. Por sua vez, $x_4$ é proveniente de uma distribuição exponencial com parâmetro $\lambda = 0,7$, resultando em uma variância de aproximadamente 2,0408.

Neste cenário, as variâncias associadas aos especialistas 1, 2, 3 e 4 foram definidas com base nas variáveis $x_2$, $x_3$, $x_4$ e $x_5$, respectivamente.

Os resultados deste experimento demonstram que os métodos propostos neste trabalho apresentam desempenho superior aos métodos concorrentes. Em particular, o método WEAR mostrou-se tão eficiente quanto o algoritmo de Raykar *et al.* (2010), enquanto o método WEAR-INS destacou-se como o melhor entre todos os avaliados. Já esperava-se que o método de Raykar *et al.* (2010) tivesse esse resultado devido às covariáveis não serem funções na formação do $y$.

Além disso, esperava-se que os modelos paramétricos fornecessem melhores estimativas dos pesos por observação. No entanto, os modelos não paramétricos surpreenderam ao fornecer as ordens de grandeza esperadas, como visto na Figura 2. Isto deve-se a origem da geração do especialista 3 que provém de uma distribuição exponencial, sendo que os modelos paramétricos assumem relações lineares e distribuições mais simétricas, como a normal. Porém, conforme apresentado na Tabela 5, os métodos LASSO e Mínimos Quadrados foram os que apresentaram os menores riscos preditivos em cada abordagem empregada.

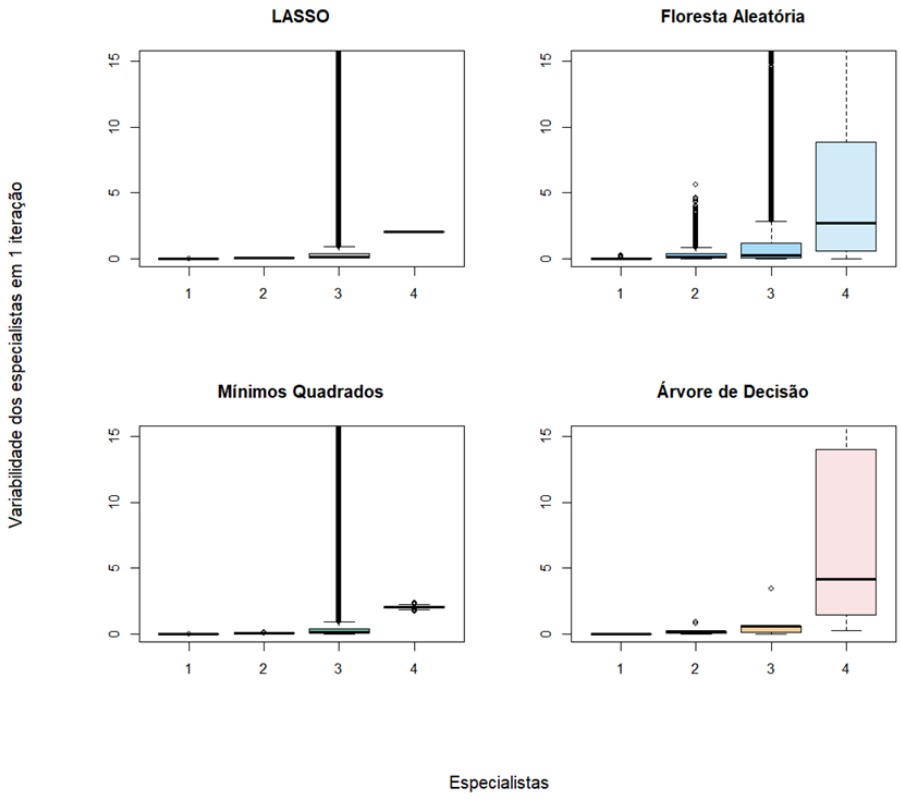


Figura 2 – Boxplot das variabilidades dos especialistas - Dados 2 - Simulados

Para ilustrar o modelo WEAR, a Tabela 2 apresenta a média das variâncias estimadas pelos quatro especialistas nos quatro algoritmos avaliados ao longo de 50 iterações, bem como os pesos estimados utilizando o modelo proposto por Raykar *et al.* (2010).

| **Abordagem** | **Modelo** | **Especialista 1** | **Especialista 2** | **Especialista 3** | **Especialista 3** |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,0799 | 0,0481 | 0,1080 | 0,2431 |
| WEAR | Floresta | 0,0608 | 1,1017 | 4,7154 | 24,7614 |
| | Árvore | 0,0605 | 1,0576 | 4,4851 | 24,1413 |
| | LASSO | 0,0594 | 1,0537 | 4,4832 | 24,1413 |
| | Mínimos Quadrados | 0,0594 | 1,0534 | 4,4824 | 24,1486 |

Tabela 2 – Dados 2 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

### 3.1.3 Dados 3

A simulação 3 tem como objetivo demonstrar que o método baseado na média ponderada por observação nem sempre apresenta o único melhor desempenho. Em certos cenários, o método proposto por Raykar *et al.* (2010) também se mostra uma excelente alternativa.

Nesta simulação, todas as covariáveis foram geradas a partir de uma distribuição normal com média zero e variâncias de 1, 4, 16 e 64, representando a variabilidade de cada especialista. A variável resposta verdadeira, $y$, foi modelada a partir de uma distribuição normal com variância também modelada por uma normal de média zero e variância 1. A relação entre $y$ e as variáveis independentes é linear, conforme a equação:

$$y = 0,5x_0 - 0,4x_1 + 0,3x_2 + 0,2x_3 + \varepsilon, \tag{3.4}$$

considerando

$$y_j = y + N(0, |x_{ki}|), \tag{3.5}$$

sendo $y$ a resposta real, $y_j$ a opinião do $j$-ésimo especialista, $k$ o índice da covariável no modelo e $i$ o índice da instância da covariável.

Como a soma de variáveis normais resulta em uma distribuição normal e todas as variáveis do modelo têm média zero, as diferenças entre elas refletem apenas a escala de suas variâncias. A Figura 3 demonstra que os métodos propostos capturaram com precisão as diferenças de variabilidade entre os especialistas. No entanto, o método de Raykar *et al.* (2010) mostrou-se igualmente adequado para modelar esses dados, indicando que o método proposto continua sendo uma excelente alternativa, embora não a única. Além disso a 3 apresenta que todos os modelos supervisionados tiveram um ótimo comportamento na estimação dos pesos dos especialistas, diferente do algoritmo de Raykar *et al.* (2010).

Por fim, em comparação com a abordagem tradicional de média aritmética, os métodos

que utilizam ponderação demonstraram desempenho superior. Os resultados completos dessa simulação estão apresentados na coluna Dados 3 da Tabela 5, refletindo as características específicas deste cenário.

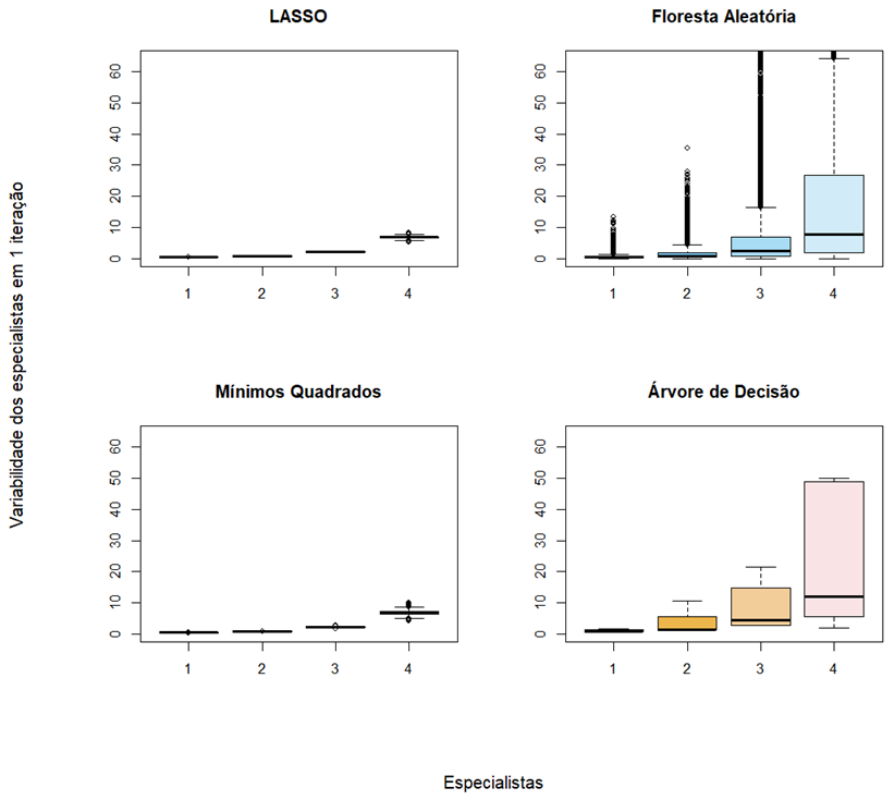


Figura 3 – Boxplot das variabilidades dos especialistas - Dados 3 - Simulados

| Abordagem | Modelo | Especialista 1 | Especialista 2 | Especialista 3 | Especialista 4 |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,8348 | 0,7807 | 0,7911 | 0,8323 |
| WEAR | Floresta | 2,1536 | 5,2998 | 17,8826 | 68,2928 |
| | Árvore | 3,5587 | 6,6851 | 19,2093 | 68,0908 |
| | LASSO | 2,0096 | 4,9779 | 16,9075 | 64,7809 |
| | Mínimos Quadrados | 2,0094 | 4,9779 | 16,9071 | 64,7802 |

Tabela 3 – Dados 3 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

### 3.1.4 Dados 4

As variáveis e o $y$ da simulação neste cenário são semelhantes aos dos Dados 3. O que foi diferenciado foram os especialistas, que foram gerados da seguinte forma:

$$y_{ji} = y_i + x_{ki}, \tag{3.6}$$

considerando que $y_{ji}$ é a i-ésima observação do j-ésimo especialista, $y_i$ é o rótulo preciso e $x_{ki}$ é a i-ésima observação da k-ésima covariável, com $k = j$. Em outras palavras, os especialistas são formados pelo padrão-ouro mais um ruído dado pela covariável. Essa forma de simular os especialistas é permitida pelo teorema 1, pois, de acordo com a subseção 3.1.3, todas as covariáveis são distribuídas normalmente com média zero. Dado que $y$ também vem de uma distribuição normal com média zero e que, se $X_1, X_2, \ldots, X_n$ são variáveis independentes normalmente distribuídas com a mesma média $\mu$, mas variâncias $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ , ou seja,

$$X_i \sim N(\mu, \sigma_i^2), \quad i = 1, 2, \ldots, n.$$

Então a soma dessas variáveis $S_n = \sum_{i=1}^{n} X_i$ será normalmente distribuída com média $n\mu$ e variância $\sum_{i=1}^{n} \sigma_i^2$ , ou seja,

$$S_n \sim \mathcal{N}\left(n\mu, \sum_{i=1}^{n} \sigma_i^2\right).$$

Essa propriedade pode ser vista em Hastie (2009).

Como $\mu = 0$, os especialistas também têm uma média de 0, o que está de acordo com a suposição do teorema 1.

Neste contexto específico, o método de Raykar *et al.* (2010) é o que apresenta desempenho mais próximo do método que utiliza o verdadeiro $y$. Isto porque foi o método que estimou melhor os pesos dos especialistas.

Além disso, o uso da Árvore de Decisão neste experimento traz um resultado inesperado: enquanto esperava-se que um algoritmo paramétrico fosse mais eficaz na estimativa da variância, é a Árvore de Decisão que se destaca. Como ilustrado na Figura 4, o algoritmo de Árvore de Decisão estimou com melhor precisão a variabilidade por observação dos especialistas.

Isto é devido à forma como os especialistas foram gerados, pois suas variações estão sistemáticas e este método segmenta melhor dados com padrões consistentes e previsíveis. Além disto existe uma alta colinearidade entre o ruído e as variáveis independentes.

A Tabela 4 mostra que nenhum estimador chegou em um resultado coerente com os pesos fornecidos.

O resultado desta e todas as outras simulações apresentadas acima estão na Tabela 5.

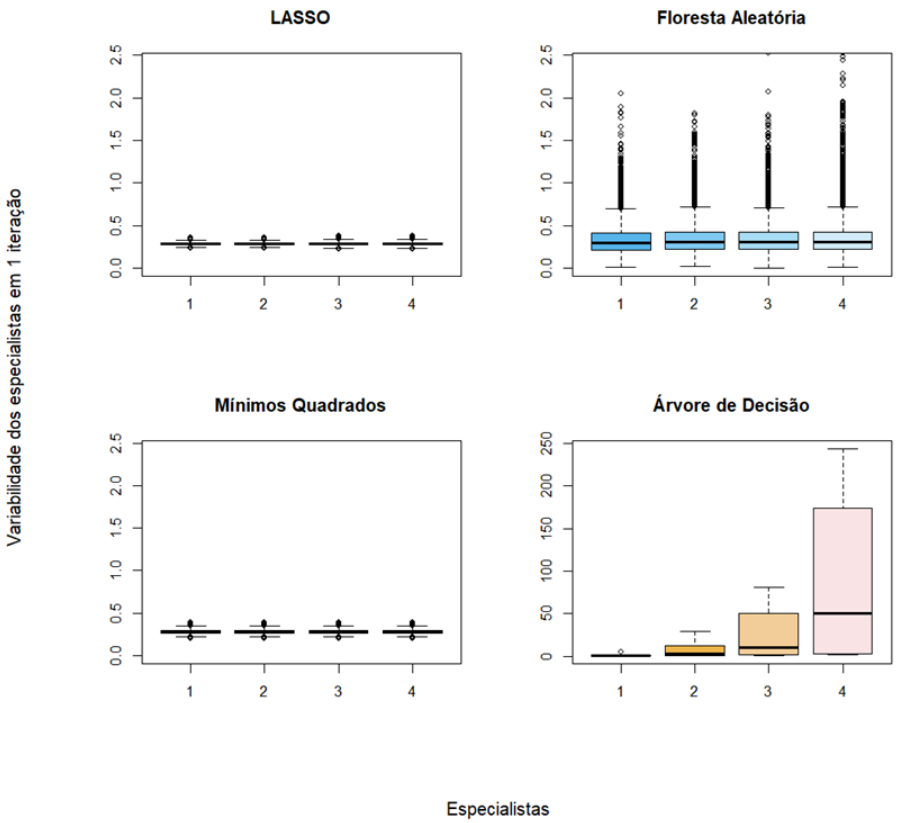


Figura 4 – Boxplot das variabilidades dos especialistas - Dados 4 - Simulados

| **Abordagem** | **Modelo** | **Especialista 1** | **Especialista 2** | **Especialista 3** | **Especialista 4** |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,7341 | 0,7866 | 0,8548 | 0,5970 |
| WEAR | Floresta | 1,1124 | 1,0956 | 1,1266 | 1,1705 |
| | Árvore | 3,5483 | 2,6903 | 5,2710 | 6,5557 |
| | LASSO | 0,9982 | 0,9981 | 0,9992 | 1,0023 |
| | Mínimos Quadrados | 0,9978 | 0,9978 | 0,9978 | 0,9978 |

Tabela 4 – Dados - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

| Métodos | Modelos | Dados 1 | Dados 2 | Dados 3 | Dados 4 |
|---|---|---|---|---|---|
| WEAR-INS (Nossa abordagem por instância) | Mínimos Quadrados | 60,1533 (0,2112) | 0,0000142 (0,0000015) | **1,0002 (0,0016)** | 6,3163 (0,0115) |
| | LASSO | 60,1535 (0,2112) | **0,0000099 (0,0000012)** | **1,0004 (0,0016)** | 6,189 (0,0141) |
| | Floresta Aleatória | **10,5202 (0,0227)** | 0,0012236 (0,0000081) | 1,1172 (0,0016) | 6,0301 (0,0316) |
| | Árvore | 18,6110 (0,0672) | 0,0007798 (0,0000255) | 2,5269 (0,0082) | **3,1672 (0,0061)** |
| WEAR (Nossa abordagem Versão corolário) | Mínimos Quadrados | 60,1524 (0,2111) | **0,0000154 (0,0000017)** | **1,0001 (0,0016)** | 6,3163 (0,0115) |
| | LASSO | 61,2661 (0,2313) | 0,0015753 (0,0001184) | **1,0004 (0,0016)** | 6,0518 (0,0115) |
| | Floresta Aleatória | **10,7371 (0,0238)** | 0,0024141 (0,0000166) | 1,1351 (0,0016) | 5,7927 (0,0158) |
| | Árvore | 18,7136 (0,0601) | 0,0009357 (0,0000051) | 2,5543 (0,0097) | **4,8331 (0,0147)** |
| Método de Raykar | - | 60,18348 (0,2076) | 0,0000165 (0,0000014) | 0,9994 (0,0017) | 1,7846 (0,0028) |
| Média Aritmética | Mínimos Quadrados | 60,3769 (0,2087) | 0,0004476 (0,0000354) | **1,0013 (0,0016)** | 6,3163 (0,0115) |
| | LASSO | 60,4070 (0,2096) | **0,0003189 (0,0000320)** | **1,0015 (0,0016)** | 6,2053 (0,0113) |
| | Floresta Aleatória | 51,3337 (0,3583) | 0,0724483 (0,0006628) | 1,4008 (0,0025) | **6,0196 (0,0108)** |
| | Árvore | **35,1966 (0,4074)** | 0,0025561 (0,0000126) | 2,7474 (0,0062) | 6,9006 (0,0152) |
| Y real (Gold standard) | Mínimos Quadrados | 60,1513 (0,2110) | **0,0000000 (0,0000000)** | **0,9999 (0,0016)** | **1,0015 (0,0014)** |
| | LASSO | 60,1502 (0,2111) | 0,0000021 (0,0000000) | **1,0002 (0,0016)** | **1,0017 (0,0014)** |
| | Floresta Aleatória | **10,0731 (0,0222)** | 0,0000019 (0,0000001) | 1,0907 (0,0017) | 1,0924 (0,0016) |
| | Árvore | 18,2708 (0,0760) | 0,0000872 (0,0000003) | 2,4832 (0,0074) | 2,4788 (0,0059) |

Tabela 5 – Resultados comparativos dos EQMs de diferentes modelos em quatro conjuntos de dados simulados.

## 3.2 Dados Reais

Nesta subseção, para avaliar o desempenho do método desenvolvido, será apresentada a aplicação com dados reais. Para esse fim, foram considerados três bancos de dados:

- Conjunto de Dados 1: Skin Segmentation [1]
- Conjunto de Dados 2: Firewall de Internet [2]
- Conjunto de Dados 3: 3D Road Network (North Jutland, Denmark) 3D[3]

Todos os dados foram obtidos no *site* <https://archive.ics.uci.edu/>.

Os especialistas e seus respectivos desvios (por instância) foram simulados, em que cada observação foi gerada a partir da instância do padrão-ouro mais o desvio. O desvio foi construído a partir de uma distribuição normal com média zero e variância, representando uma função do peso simulado proveniente das observações. Para evitar números negativos na variância, usamos o valor absoluto da variável.

O padrão-ouro de cada conjunto de dados corresponde a uma de suas variáveis, a qual é quantitativa contínua, garantindo sua adequação para aplicação em métodos de regressão

Os métodos de aprendizado de máquina supervisionados utilizados foram os de Mínimos Quadrados Ordinários e LASSO como métodos paramétricos, e Árvore de Decisão e Floresta Aleatória como métodos não paramétricos. Eles foram aplicados nos dois métodos desenvolvidos neste trabalho (WEAR-INS e WEAR), no método que usa a média aritmética como variável

[1] <https://archive.ics.uci.edu/dataset/229/skin+segmentation>
[2] <https://archive.ics.uci.edu/dataset/542/internet+firewall+data>
[3] <https://archive.ics.uci.edu/dataset/246/3d+road+network+north+jutland+denmark>

resposta, bem como utilizando o padrão-ouro. Essa abordagem permitiu a comparação entre os métodos que usufruem destas técnicas tradicionais com o modelo proposto por Raykar *et al.* (2010). Os resultados podem ser vistos na Tabela 6.

| **Métodos** | **Modelos** | **Skin Segmentation** | **Internet Firewall Data** | **3D Road Network** |
|---|---|---|---|---|
| WEAR-INS (Nossa abordagem por instância) | Mínimos Quadrados | 0,1689 (0,0019) | 0,8624 (0,0149) | **0,056386 (0,000177)** |
| | LASSO | 0,1689 (0,0019) | 0,8538 (0,0127) | **0,056390 (0,000177)** |
| | Floresta Aleatória | **0,1252 (0,0012)** | **0,5306 (0,0080)** | 7,921954 (0,041696) |
| | Árvore | 0,1500 (0,0017) | 0,7151 (0,0107) | 0,084355 (0,000252) |
| WEAR (Nossa abordagem visão corolário) | Mínimos Quadrados | 0,1689 (0,0019) | 0,8626 (0,0126) | **0,056758 (0,000177)** |
| | LASSO | 0,1689 (0,0019) | 0,8675 (0,0148) | **0,056799 (0,000178)** |
| | Floresta Aleatória | **0,1286 (0,0012)** | **0,5679 (0,0157)** | 9,860449 (0,042159) |
| | Árvore | 0,1498 (0,0017) | 0,8245 (0,0435) | 0,084125 (0,000248) |
| Raykar | - | 0,1656943 (0,0018) | 0,9304 (0,0624) | 0,057082 (0,000155) |
| Média Aritmética | Mínimos Quadrados | 0,1689 (0,0019) | 0,8939 (0,0298) | **0,057104 (0,000177)** |
| | LASSO | 0,1690 (0,0019) | 0,9195 (0,0549) | 0,059966 (0,000191) |
| | Floresta Aleatória | **0,1289 (0,0012)** | **0,5559 (0,0127)** | 28,093236 (0,240713) |
| | Árvore | 0,2353 (0,0017) | 0,7599 (0,0206) | 0,084923 (0,000259) |
| Y real | Mínimos Quadrados | 0,1689 (0,0019) | 0,8498 (0,0121) | 0,055789 (0,000177) |
| | LASSO | 0,1689 (0,0019) | 0,8498 (0,0121) | 0,055790 (0,000177) |
| | Floresta Aleatória | **0,1257 (0,0012)** | **0,5309 (0,0080)** | **0,022780 (0,000159)** |
| | Árvore | 0,1496 (0,0017) | 0,7150 (0,0107) | 0,040034 (0,000168) |

Tabela 6 – Resultados comparativos dos EQMs de diferentes modelos em três conjuntos de dados reais.

### 3.2.1 Skin Segmentation

Na primeira simulação, foi utilizado o Conjunto de Dados 1, composto por 245,057 instâncias e três variáveis explicativas. A variável originalmente indicada como resposta pelos criadores do conjunto de dados é uma variável *dummy*, que identifica se a instância observada corresponde a um tecido cutâneo ou a uma parte não cutânea do corpo humano. Contudo, para fins de ilustração e aplicação do método de regressão desenvolvido neste trabalho, a variável *dummy* foi tratada como explicativa, enquanto a terceira variável do banco de dados original, que é uma variável quantitativa, foi utilizada como variável resposta.

Todas as variáveis quantitativas foram escaladas dividindo seus valores por 100, exclusivamente para alterar a escala dos dados, sem impacto em suas propriedades estatísticas. A variável *dummy*, por sua vez, permaneceu inalterada.

O desvio associado a cada especialista foi gerado multiplicando uma das variáveis explicativas quantitativas por constantes definidas no vetor [0,1,1,2,5]. Essa abordagem assegura que o primeiro especialista seja o mais preciso, enquanto o quarto apresenta o maior desvio, sendo o menos preciso.

Os resultados da aplicação no Conjunto de Dados 1 estão apresentados na Tabela 6. Com base nesses resultados, pode-se concluir que o modelo desenvolvido neste estudo (WEAR-INS) é o mais eficaz para este conjunto de dados. Além disso, observa-se que a técnica não paramétrica geralmente supera os métodos paramétricos.

Para avaliar se o método proposto por instância identificou efetivamente o pior especialista, foi construído um boxplot das variâncias estimadas pelo modelo. É evidente a partir da Figura 5 que, em todos os modelos, a variabilidade estimada para cada especialista está de

acordo com o desvio gerado.

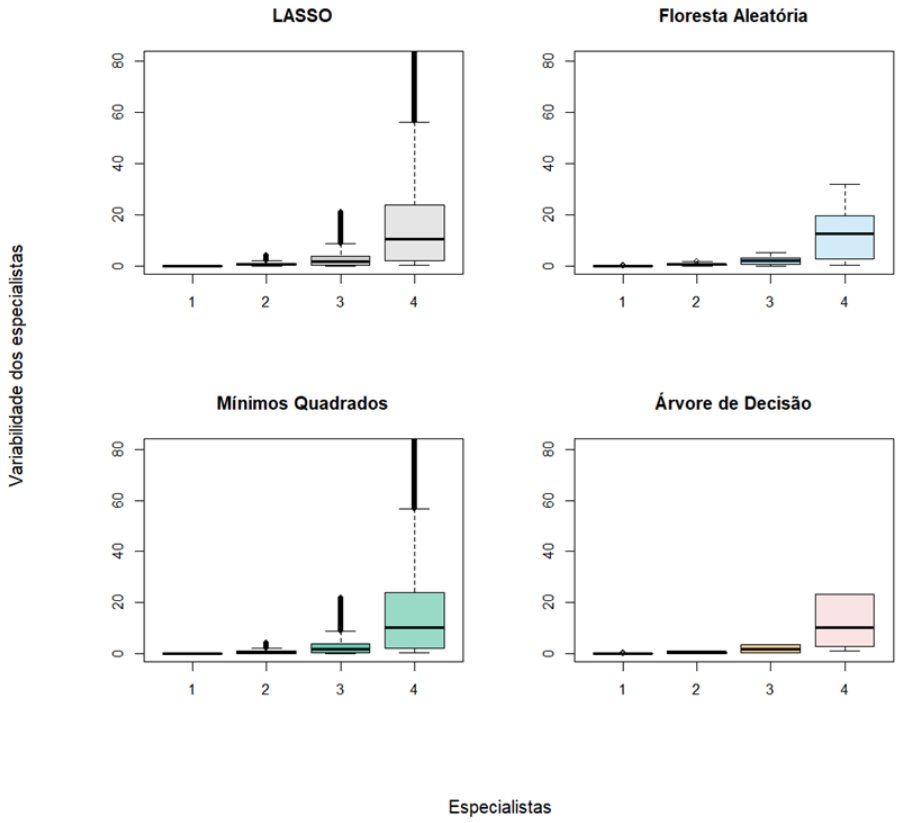


Figura 5 – Skin Segmentation - Boxplot das variabilidades dos especialistas

Para avaliar se os valores observados estão próximos dos estimados pelo modelo proposto por instância, foi construído um gráfico de dispersão para cada técnica do modelo WEAR-INS. Verificou-se que todas as técnicas apresentam uma relação linear positiva entre os valores observados e estimados. Veja na Figura 6.

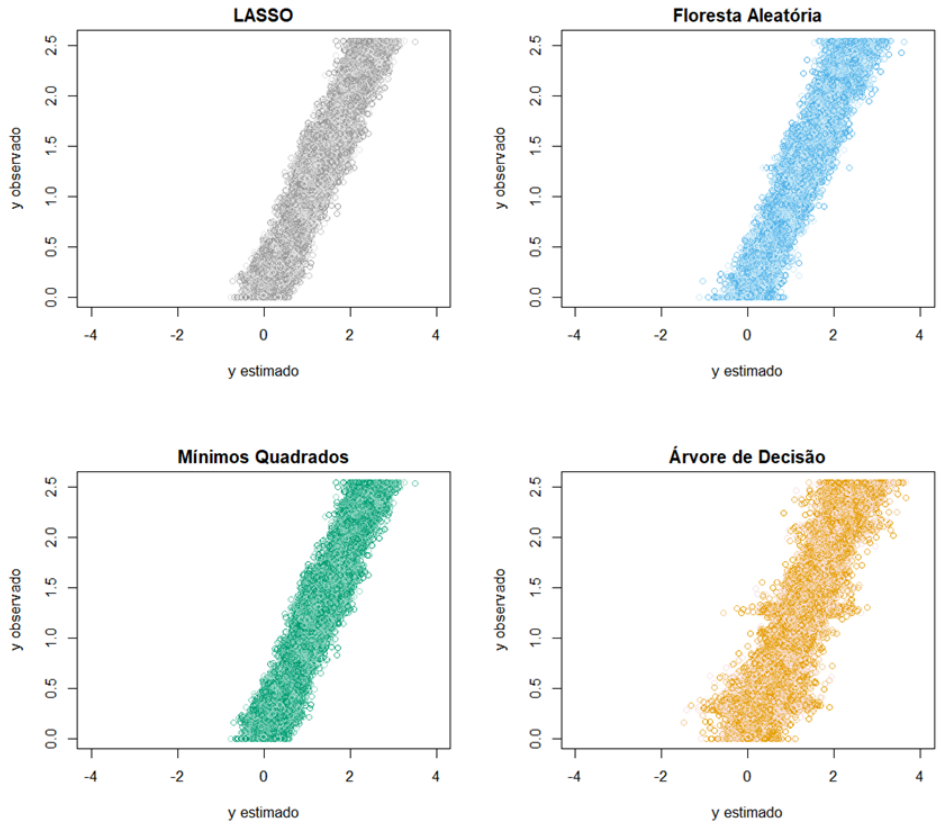


Figura 6 – Skin Segmentation - Gráfico de dispersão entre o y observado simulado e o y estimado

Ademais, na Tabela 7 mostramos que todos os estimadores dos pesos gerais tiveram ótimo resultado. Na Tabela 6 apresentamos que desempenho do modelo proposto por Raykar *et al.* (2010) aproxima-se dos métodos paramétricos, o que era esperado, uma vez que ele também emprega uma forma fechada dentro do algoritmo EM. Contudo, nesse cenário, embora a ponderação favoreça o resultado, o fato de o algoritmo ser limitado em relação à funcionalidade dos dados torna os dois métodos desenvolvidos neste trabalho significativamente superiores ao de Raykar *et al.* (2010). Entre eles, o algoritmo WEAR-INS destaca-se como o melhor, especialmente por não apresentar interpolação nos intervalos de confiança.

| **Abordagem** | **Modelo** | **Especialista 1** | **Especialista 2** | **Especialista 3** | **Especialista 4** |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 5,4164 | 0,4719 | 0,1260 | 0,0205 |
| WEAR | Floresta | 0,1391 | 2,1031 | 8,0274 | 48,7904 |
| | Árvore | 0,1649 | 2,1788 | 8,1553 | 49,1196 |
| | LASSO | 0,1822 | 2,1431 | 8,0778 | 48,7719 |
| | Mínimos Quadrados | 0,1822 | 2,1432 | 8,0777 | 48,7719 |

Tabela 7 – Skin Segmentation - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

### 3.2.2 Internet Firewall Data - Versão 1

O segundo conjunto de dados testado contém 65,532 linhas e 12 variáveis. Neste cenário, os 4 especialistas foram gerados a partir de 4 covariáveis distintas. Isso serve para ilustrar que algumas variáveis podem ser mais desafiadoras para alguns especialistas do que para outros. Neste exemplo, a variância do especialista é o valor da própria variável, conforme a seguinte equação:

$$esp_{ji} = y_i + N(0, (|X_{qi}|)),$$

na qual $j$ é o j-ésimo especialista, $i$ é a instância observada, e $q$ é o índice da covariável.

Para garantir a aleatoriedade, dada a ordem arbitrária na qual os dados foram distribuídos, a primeira variável, "quantidade", tornou-se a variável resposta, e as últimas 4 tornaram-se os ruídos dos especialistas.

De acordo com a Tabela 6, os métodos discutidos neste trabalho alcançaram os menores riscos, sendo o melhor resultado de acordo com a métrica EQM, o método WEAR-INS, tendo um desempenho tão bom quanto o verdadeiro $y$ (padrão-ouro). Isso demonstra que, mesmo não havendo um padrão-ouro em cenários reais e as opiniões dos especialistas podendo ser ruidosas por instâncias, nosso método pode estimar o padrão-ouro e, assim, modelar os dados com mais precisão.

A Tabela 8 mostra que nenhum dos métodos deram menor peso para o pior especialista e deve ser por isso que o resultado do WEAR ficou pior do que a média.

Em relação ao modelo de Raykar *et al.* (2010), novamente o desempenho foi inferior. Já as outras alternativas de modelagem, há um empate técnico, conforme indicado pela sobreposição dos intervalos de confiança, entre o método que usa a média aritmética dos especialistas como variável resposta e o método WEAR, o que prova que há cenários em que a ponderação geral não é suficiente para diferir entre os especialistas. Além disso, entre todos os métodos apresentados, o algoritmo com o melhor desempenho para esses dados foi a Floresta Aleatória, o que não é surpreendente, já que a literatura sugere que esse algoritmo é um dos melhores para predição sem exigir muito pré-processamento e ajuste de hiperparâmetros, como visto em (ABDULKAREEM; ABDULAZEEZ, 2021) e (DONGES; URWIN; PIERRE, 2023).

Espera-se que o método WEAR-INS tenha o melhor desempenho devido à estimativa precisa da variância por observação. De acordo com a Figura 7 e os resultados da Tabela 6, os especialistas do modelo não paramétrico exibem maior variabilidade entre si, com o algoritmo de Floresta Aleatória, particularmente o especialista 2, mostrando o maior "desvio".

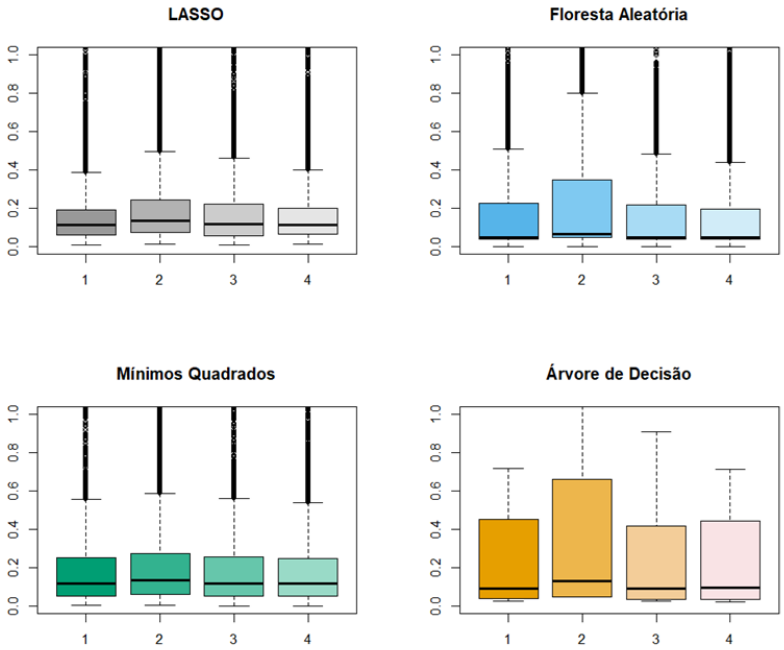


Figura 7 – Internet Firewall Data Versão 1 - Boxplot das variabilidades dos especialistas

Para validar isso, foi criado o mesmo estilo de gráfico (Figura 8) usando os dados de validação e selecionando as variáveis usadas para a construção das variâncias por observações dos especialistas, como:

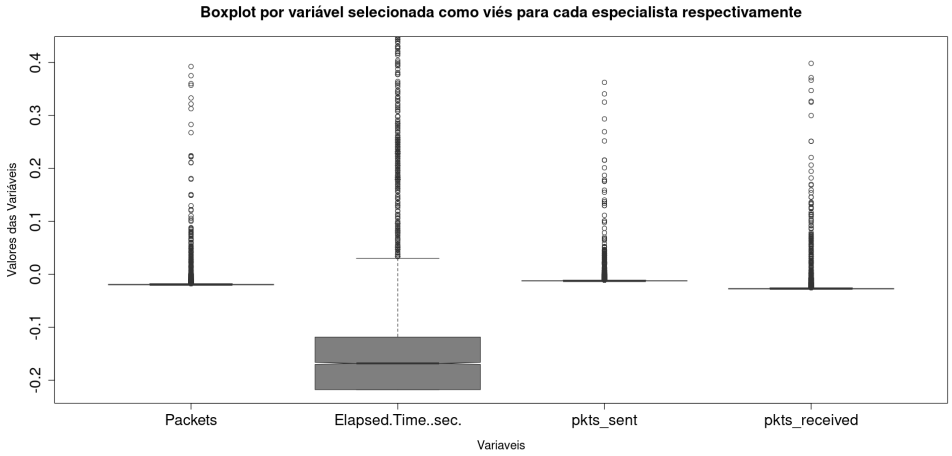


Figura 8 – Internet Firewall Data - Versão 1 - Boxplot das 4 variáveis selecionadas como variâncias por observações de cada especialista, respectivamente

A Figura 8 mostra que a variabilidade estimada pelo modelo de Floresta foi a mais próxima da variabilidade real dos dados.

Também foi criado um gráfico de correlação para verificar qual *y* estimado pelo método WEAR-INS está mais próximo do *y* real, conforme mostrado na Figura 9. O algoritmo de Floresta foi o que mais se aproximou do *y* real.

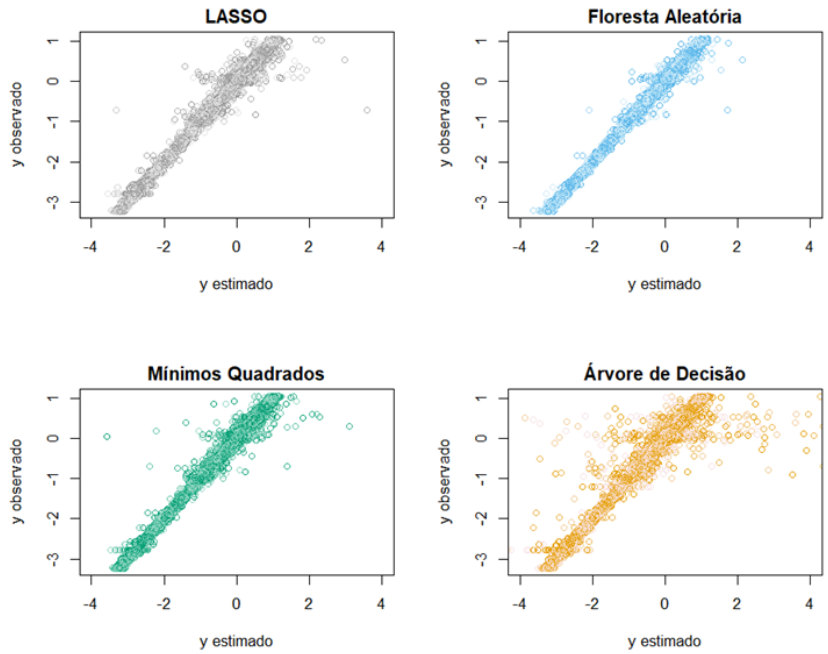


Figura 9 – Internet Firewall Data - Gráfico de dispersão entre o y observado simulado e o y estimado pelo método WEAR-INS

| Abordagem | Modelo | Especialista 1 | Especialista 2 | Especialista 3 | Especialista 4 |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 1,0980 | 0,5121 | 0,8216 | 0,8490 |
| WEAR | Floresta | 33,1463 | 2,0524 | 1,6374 | 3,6720 |
| | Árvore | 40,2527 | 2,1740 | 1,0058 | 4,2032 |
| | LASSO | 14,2118 | 2,5270 | 41,2620 | 4,2905 |
| | Mínimos Quadrados | 36,1455 | 3,9123 | 109,0288 | 214,4679 |

Tabela 8 – Internet Firewall Data - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

### 3.2.3 3D Road Network (North Jutland, Denmark)

Os dados neste cenário consistem em 434,874 observações e 4 variáveis, mas a primeira variável foi desconsiderada, restando apenas as variáveis 'latitude', 'altitude' e'longitude', sendo 'latitude' a variável resposta.

A figura 10 mostra a variabilidade de cada uma das covariáveis consideradas para serem os ruídos dos especialistas, sendo a altitude do especialista 1 e a longitude do especialista 2.

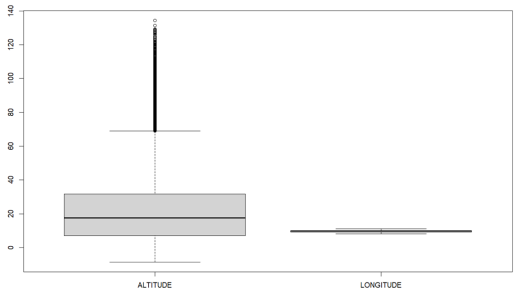


Figura 10 – 3D Road Network (North Jutland, Denmark) - Boxplot das covariáveis

De acordo com a figura 11, nos modelos paramétricos, o pior especialista é conforme o esperado. Isso também é observado no método de Floresta Aleatória, mas não no de Árvore de Decisão.

Considerando a diferença entre os especialistas, espera-se que os modelos ponderados apresentem desempenho superior. Isto apenas acontece no modelo WEAR-INS.

Conforme vemos na Tabela 9 todos os modelos que usam a técnica WEAR e o método de (RAYKAR *et al.*, 2010) recuperaram bem os pesos de cada especialista.

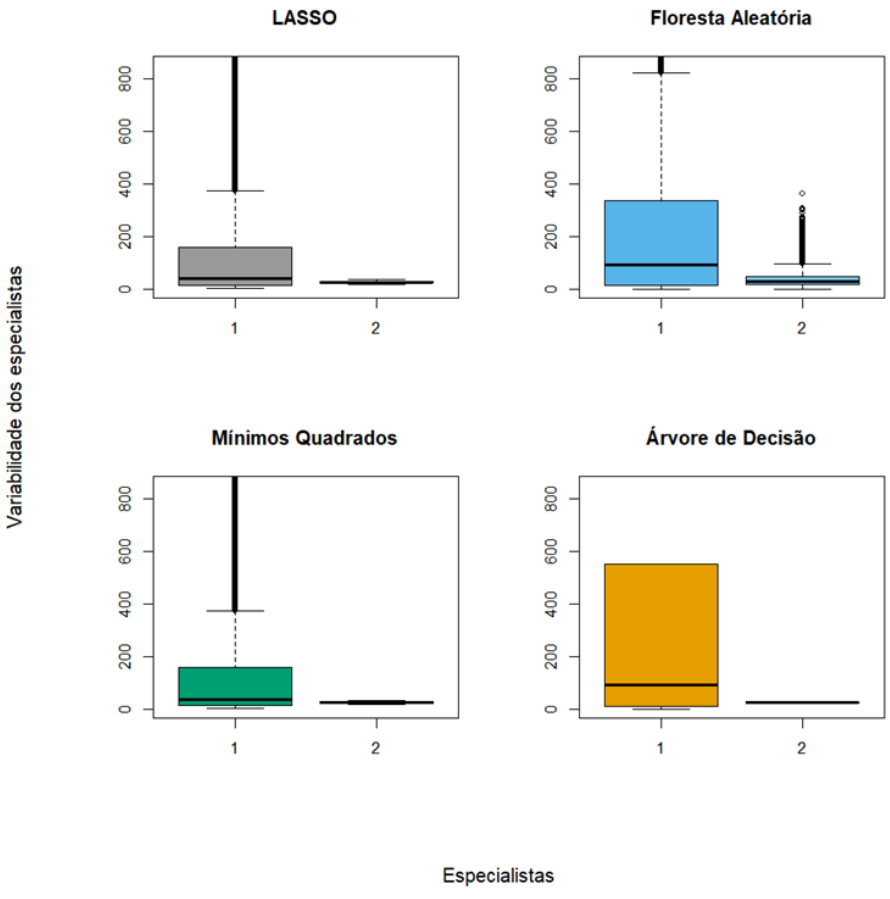


Figura 11 – 3D Road Network (North Jutland, Denmark) - Boxplot das variabilidades dos especialistas

A Figura 12, apresenta gráfico de correlação entre $\hat{y}$ ponderado por observação e o verdadeiro $y$, utilizando os dados de treinamento. Vemos que o mais próximo do valor considerado real é a variável resposta estimada pelo método floresta.

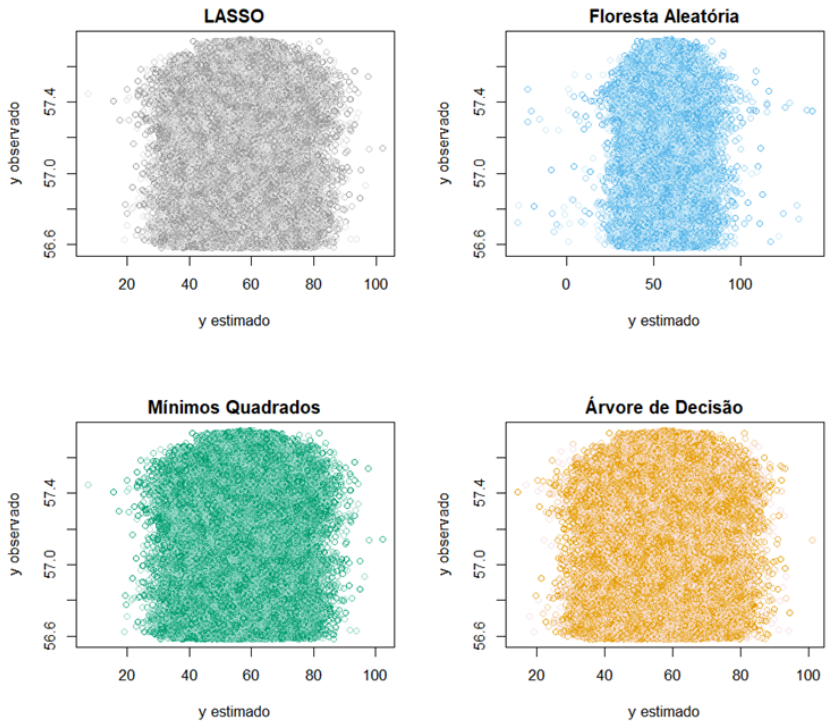


Figura 12 – 3D Road Network (North Jutland, Denmark) - Gráfico de dispersão entre o y observado simulado e o y estimado

| Abordagem | Modelo | Especialista 1 | Especialista 2 |
|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,0012 | 0,0105 |
| WEAR | Floresta | 935,0037 | 104,9015 |
| | Árvore | 844,6718 | 94,1905 |
| | LASSO | 844,7187 | 94,1773 |
| | Mínimos Quadrados | 844,7859 | 94,1774 |

Tabela 9 – 3D Road Network (North Jutland, Denmark - Pesos estimados para cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

### 3.2.4 *Firewall de Internet - Versão 2 e Versão 3*

Nesta subseção queremos apresentar cenários diferenciados usando o banco de dados Firewall de Internet, com todas as características do 3.2.2, exceto o modelo como os 4 especialistas são gerados.

Na Versão 2 simulamos os especialistas provindos todos da mesma covariável, veja:

$$especialista_j = y + N(0, x_1),$$

onde $x_1$ é a covariável 1 e $y$ é o $y$ real. O objetivo desta simulação é demonstrar que, em cenários que os especialistas apresentam opiniões similares, o método tradicional baseado na média aritmética é uma alternativa válida, com desempenho comparável aos métodos propostos, conforme visto na Tabela 12. Isso reforça os benefícios das abordagens desenvolvidas, pois, mesmo em situações de alta similaridade entre especialistas, elas mantêm um desempenho robusto. Conforme mencionado anteriormente, há interferências ou ruídos impossíveis de mensurar. Assim, na ausência de um conhecimento mais aprofundado sobre os especialistas e quando apenas um único método pode ser aplicado, o WEAR-INS se destaca por capturar a presença de ruídos associados tanto aos especialistas quanto às observações, tornando-se uma opção particularmente interessante.

Na Versão 3, os especialistas foram criados sem o vínculo com o y real, como:

$$especialista_j = N(0, x_j),$$

no qual o especialista $j$ vem de uma normal de média 0 e variância sendo a covariável $j$.

Este cenário foi elaborado para demonstrar que, mesmo em situações nas quais os especialistas não possuem vínculo direto com $y$, o método WEAR-INS, que utiliza ponderação tanto por especialista quanto por observação, apresenta o menor erro preditivo, desconsiderando o erro padrão. Conforme observado na Tabela 12, apesar de nenhum método ter se aproximado do menor erro preditivo obtido pelo $y$ real no algoritmo LASSO, nosso método mostrou-se tão eficiente quanto a média aritmética, ainda neste cenário específico.

As Tabelas 10 e 11 apresentam as estimativas dos pesos atribuídos a cada especialista, calculados pelo algoritmo de Raykar *et al.* (2010) e pelo método WEAR. Conforme o esperado, os resultados da Tabela 10 deveriam exibir pesos mais equilibrados entre os quatro especialistas. No entanto, isso não ocorreu devido à aleatoriedade presente no processo. Por outro lado, na Tabela 11, destaca-se que o único método que atribuiu um peso menor ao especialista 2 foi o LASSO, como era previsto com base na Figura 8. Esse comportamento está alinhado com os resultados observados na Tabela 12, reforçando a coerência do modelo com as características dos dados.

Os gráficos 13 e 15 estão consistentes com os resultados esperados a partir da simulação. Na Versão 2, observa-se uma variação uniforme entre os especialistas, enquanto na Versão 3 destaca-se uma variação mais elevada no Especialista 2.

| **Abordagem** | **Modelo** | **Especialista 1** | **Especialista 2** | **Especialista 3** | **Especialista 4** |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,1324 | 1,0148 | 0,9375 | 0,4289 |
| WEAR | Floresta | 0,7617 | 6,9719 | 5,2396 | 1,3230 |
| | Árvore | 3,4061 | 6,5135 | 6,4027 | 1,8216 |
| | LASSO | 1,5814 | 6,6361 | 9,7486 | 2,0655 |
| | Mínimos Quadrados | 9,6971 | 4,1846 | 9,9036 | 1,3424 |

Tabela 10 – Firewall de Internet - Versão 2 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

| **Abordagem** | **Modelo** | **Especialista 1** | **Especialista 2** | **Especialista 3** | **Especialista 4** |
|---|---|---|---|---|---|
| Algoritmo Raykar | Raykar | 0,1571 | 0,8707 | 39,8562 | 0,9826 |
| WEAR | Floresta | 0,0742 | 1,1292 | 3,2107 | 0,8738 |
| | Árvore | 2,5131 | 1,2908 | 3,6152 | 0,8298 |
| | LASSO | 1,8725 | 0,8700 | 3,3482 | 1,1534 |
| | Mínimos Quadrados | 8,4209 | 0,8809 | 5,1649 | 0,5285 |

Tabela 11 – Firewall de Internet - Versão 3 - Média das variâncias gerais estimadas por cada especialista em cada modelo tradicional da literatura e o peso estimado usando o algoritmo de Raykar.

| **Métodos** | **Models** | **Versão 2** | **Versão 3** |
|---|---|---|---|
| WEAR-INS (Nossa abordagem por instância) | Mínimos Quadrados | 7,8704 (6,7427) | 0,9942 (0,0221) |
| | LASSO | **0,5519 (0,0219)** | 0,9864 (0,0193) |
| | Floresta Aleatória | 0,6935 (0,0104) | 0,9799 (0,0151) |
| | Árvore | 0,9121 (0,0335) | 1,0276 (0,0564) |
| WEAR (Nossa abordagem visão corolário) | Mínimos Quadrados | 0,8667 (0,0259) | 1,0517 (0,0573) |
| | LASSO | **0,5965 (0,0588)** | 1,3033 (0,2954) |
| | Floresta Aleatória | 0,6984 (0,0105) | 0,9852 (0,0149) |
| | Árvore | 0,8555 (0,0251) | 0,9738 (0,0148) |
| Raykar | - | 0,8667 (0,0206) | 1,0084 (0,0324) |
| Média Aritmética | Mínimos Quadrados | 0,8959 (0,0559) | 1,0532 (0,0741) |
| | LASSO | **0,5443 (0,0125)** | 0,996 (0,0192) |
| | Floresta Aleatória | 0,7319 (0,0185) | 1,021 (0,0332) |
| | Árvore | 0,8662 (0,0131) | 0,9804 (0,0161) |
| Y real | Mínimos Quadrados | 0,8326 (0,0117) | 0,8326 (0,0117) |
| | LASSO | **0,5294 (0,0079)** | **0,5294 (0,0079)** |
| | Floresta Aleatória | 0,6936 (0,0104) | 0,6936 (0,0104) |
| | Árvore | 0,8320 (0,0118) | 0,832 (0,0118) |

Tabela 12 – Firewall de Internet (Versão 2 e Versão 3) - Comparação entre diferentes métodos, a partir do erro preditivo.

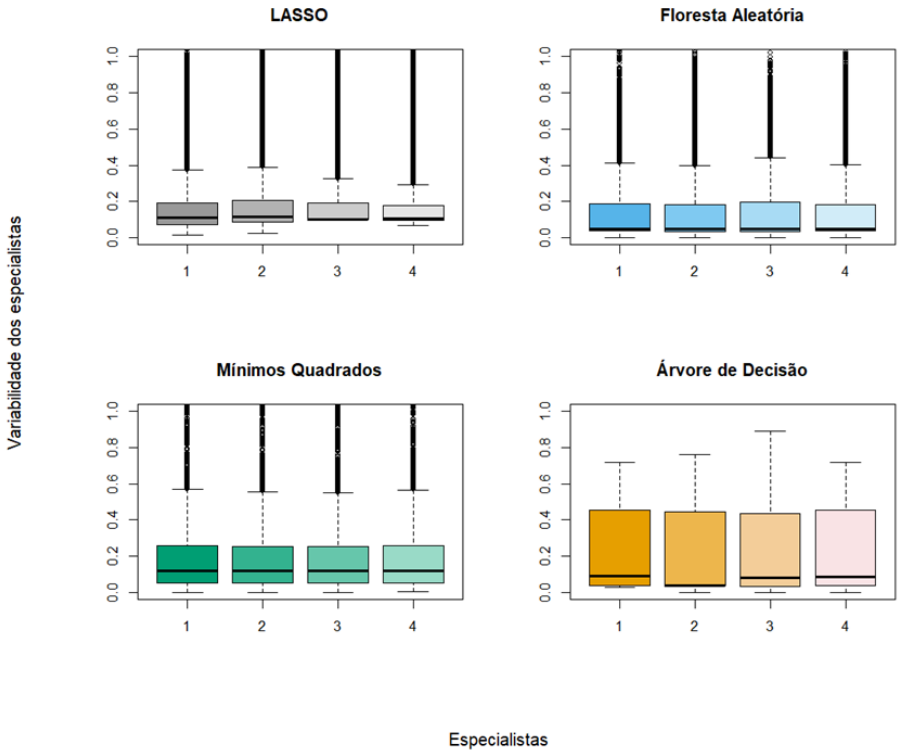


Figura 13 – Firewall de Internet (Versão 2) - Boxplot das variabilidades dos especialistas

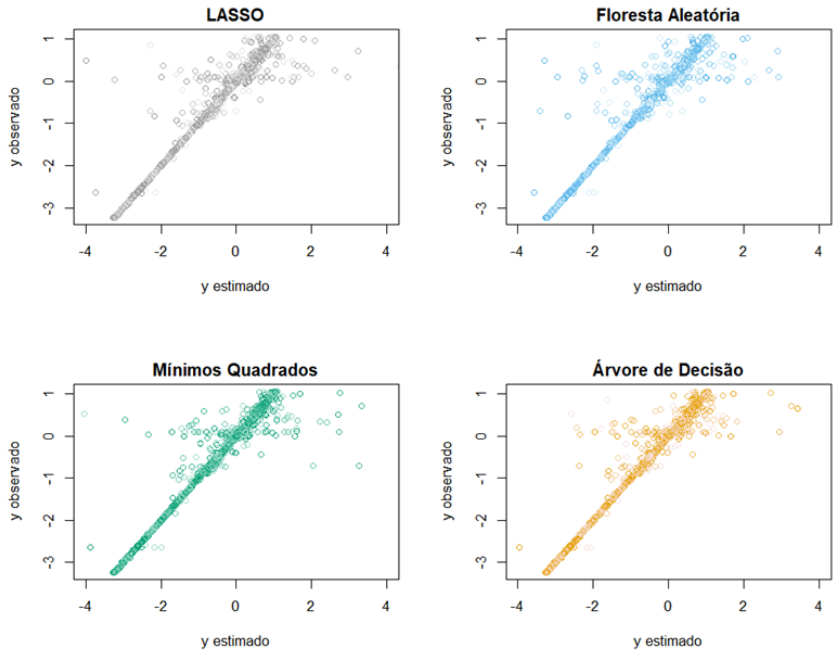


Figura 14 – Firewall de Internet (Versão 2) - Gráfico de dispersão entre o y observado simulado e o y estimado

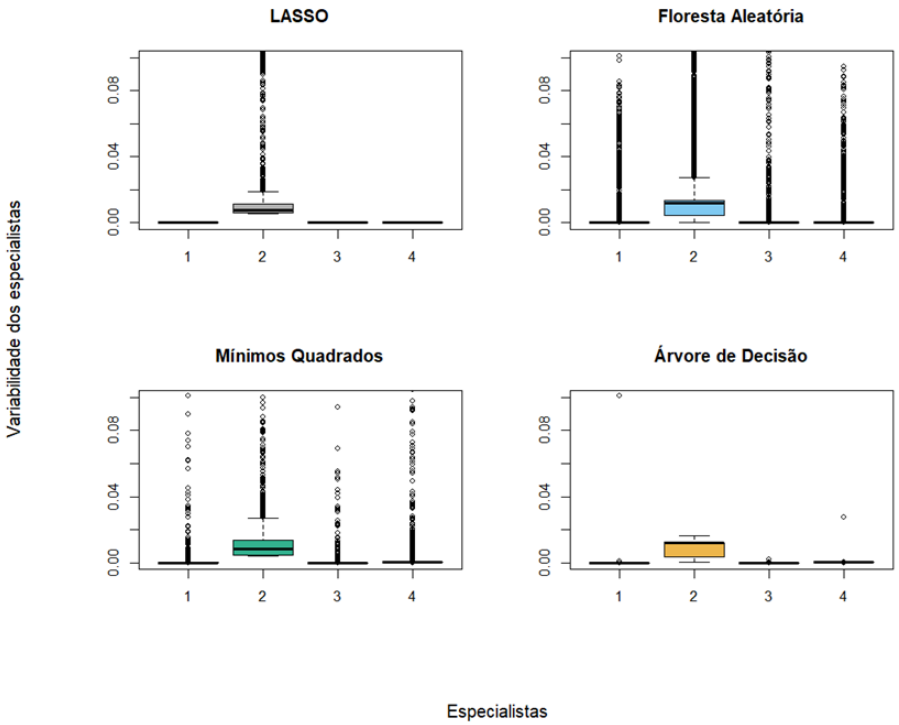


Figura 15 – Firewall de Internet (Versão 3) - Boxplot das variabilidades dos especialistas

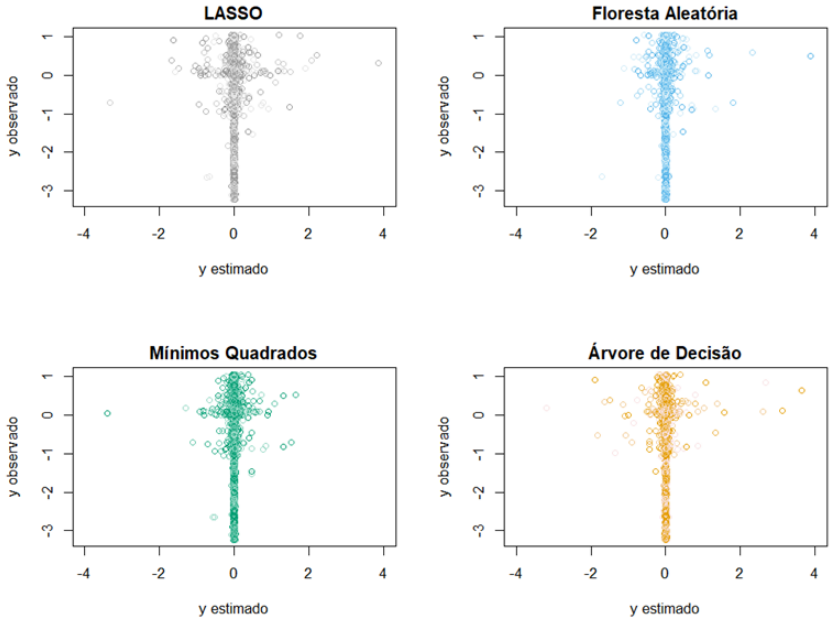


Figura 16 – Firewall de Internet (Versão 3) - Gráfico de dispersão entre o y observado simulado e o y estimado

CAPÍTULO

# 4

# CONCLUSÃO

Neste estudo investigamos cenários em que a obtenção de rótulos precisos é inviável, recorrendo, portanto, à opinião de múltiplos especialistas, cujas avaliações podem divergir significativamente. Para lidar com essas situações, propusemos dois métodos inovadores, rápidos e flexíveis para problemas de regressão em ambientes com rótulos ruidosos, levando em consideração a variabilidade das *expertises* dos especialistas de forma geral, pelo método WEAR, e em função de cada observação pelo método WEAR-INS.

Os métodos propostos se destacam pela construção inteligente dos pesos atribuídos a cada especialista, sendo a ponderação por observação fundamentada na variância empírica das avaliações, técnica amplamente utilizada em abordagens não paramétricas. Essa estratégia permite ajustar o peso de cada especialista de forma adaptativa, privilegiando aqueles cujas avaliações apresentam, num cenário não viesado, menor variabilidade em relação ao rótulo real. Assim, a média ponderada das opiniões resulta em uma estimativa mais robusta do rótulo verdadeiro.

Por meio de análises empíricas realizadas tanto em dados simulados quanto em dados precisos, evidenciamos que a abordagem proposta supera métodos concorrentes em termos de desempenho preditivo. Além disso, a flexibilidade do WEAR e WEAR-INS possibilita a aplicação de diferentes técnicas paramétricas e não paramétricas, permitindo a seleção do modelo com menor risco preditivo para cada cenário específico.

A experimentação que trouxe dados totalmente simulados apresenta 4 cenários, sendo que o primeiro teve o método WEAR-INS com desempenho superado e o WEAR e o método da média aritmética com MSE próximos. A segunda experimentação prova que há cenários onde ambos os métodos propostos são melhores do que o da média aritmética, sendo o modelo de Raykar com desempenho semelhante ao WEAR. Este cenário mostra que ao conseguir replicar a forma de Y próxima do real, os modelos paramétricos possuem maior desempenho do que os modelos não paramétricos. A terceira experimentação possui os quatro especialistas provindos

de distribuições normais com médias iguais e variâncias distintas, sendo que neste cenário os métodos que usam técnicas ponderadas possuem desempenho semelhante aos tradicionais. Ou seja, quando conhecido que os especialistas possuem comportamentos semelhantes, o modelo da média aritmética possui a vantagem de ter um processamento reduzido em comparação aos métodos propostos neste trabalho. Já a última experimentação dos dados totalmente simulados apresenta que em situações muito específicas o método de estimar as *expertises* não funciona adequadamente.

A experimentação com dados reais e especialistas simulados mostrou que os métodos propostos neste trabalho são mais adequados para estimar o valor real de $Y$, especialmente em cenários com múltiplos especialistas distintos e opiniões discrepantes. Nos dados Skin Segmentation e 3D Road Network (North Jutland, Denmark), os métodos WEAR e WEAR-INS demonstraram desempenho superior à média aritmética. No segundo conjunto de dados, embora o método WEAR-INS tenha se destacado como o mais eficiente, o método WEAR apresentou um desempenho semelhante ao da média aritmética. Além disso, na subseção 3.2.4 vemos um bom desempenho dos métodos propostos, mesmos em situações de especialistas muito semelhantes ou especialistas que não foram gerados em função do Y real. Todos os resultados apresentados reforçam a aplicabilidade dos métodos propostos em situações que demandam a consideração de múltiplas opiniões especializadas.

Apesar dos resultados promissores, algumas limitações devem ser consideradas. O método pode ser aprimorado para diferenciar o ruído específico dos especialistas do viés inerente ao modelo, além de flexibilizar a suposição de que as expectativas dos especialistas sejam necessariamente congruentes entre si e com o padrão-ouro.

Concluímos, portanto, que a abordagem proposta, ao penalizar observações com alta variabilidade e valorizar aquelas mais consistentes, constitui uma alternativa eficaz para a modelagem de dados com rótulos ruidosos, contribuindo para o avanço das técnicas de regressão em contextos complexos e de difícil rotulagem.

## 4.1 Trabalhos Futuros

- Explorar a flexibilização da suposição de igualdade de expectativas entre os especialistas e o padrão-ouro, visando aumentar a adaptabilidade do método a diferentes cenários.
- Desenvolver estratégias que permitam diferenciar o ruído proveniente dos especialistas do viés natural do modelo.

# REFERÊNCIAS


ABDULKAREEM, N. M.; ABDULAZEEZ, A. M. Science and business. **International Journal**, v. 5, n. 2, p. 128–142, 2021. Citado na página 53.

ALBUQUERQUE, P. H. M.; MEDINA, F. A. S.; SILVA, A. R. d. Regressão logística geograficamente ponderada aplicada a modelos de credit scoring. **Revista Contabilidade & Finanças**, SciELO Brasil, v. 28, p. 93–112, 2017. Citado na página 23.

ALGAN, G.; ULUSOY, I. Image classification with deep learning in the presence of noisy labels: A survey. **Knowledge-Based Systems**, Elsevier, v. 215, p. 106771, 2021. Citado na página 25.

CHITTARANJAN, G.; ARAN, O.; GATICA-PEREZ, D. Inferring truth from multiple annotators for social interaction analysis. In: **Neural Information Processing Systems (NIPS) Workshop on Modeling Human Communication Dynamics (HCD)**. [S.l.: s.n.], 2011. Citado na página 25.

CIOBOTARU, A.; CONSTANTINESCU, M. V.; DINU, L. P.; DUMITRESCU, S. Red v2: Enhancing red dataset for multi-label emotion detection. In: **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. [S.l.: s.n.], 2022. p. 1392–1399. Citado na página 24.

COONEY, P.; WHITE, A. Utilizing expert opinion to inform extrapolation of survival models. **arXiv preprint arXiv:2112.02288**, 2021. Citado na página 25.

CZEKAJ, L.; ZIEMBLA, W.; JEZIERSKI, P.; SWINIARSKI, P.; KOLODZIEJAK, A.; OGNIEWSKI, P.; NIEDBALSKI, P.; JEZIERSKA, A.; WESIERSKI, D. Labeler-hot detection of eeg epileptic transients. In: IEEE. **2019 27th European Signal Processing Conference (EUSIPCO)**. [S.l.], 2019. p. 1–5. Citado na página 24.

DONGES, N.; URWIN, M.; PIERRE, S. Random forest: a complete guide for machine learning. **Built In**, 2023. Citado na página 53.

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A. d.; CARVALHO, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. 2021. Citado na página 23.

FRÉNAY, B.; VERLEYSEN, M. Classification in the presence of label noise: a survey. **IEEE transactions on neural networks and learning systems**, IEEE, v. 25, n. 5, p. 845–869, 2013. Citado na página 25.

GULSHAN, V.; PENG, L.; CORAM, M.; STUMPE, M. C.; WU, D.; NARAYANASWAMY, A.; VENUGOPALAN, S.; WIDNER, K.; MADAMS, T.; CUADROS, J. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. **Jama**, American Medical Association, v. 316, n. 22, p. 2402–2410, 2016. Citado na página 24.

HASTIE, T. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009. Citado na página 47.

IZBICKI, R.; STERN, R. B. Learning with many experts: model selection and sparsity. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, Wiley Online Library, v. 6, n. 6, p. 565–577, 2013. Citado na página 25.

JINDAL, I.; PRESSEL, D.; LESTER, B.; NOKLEBY, M. An effective label noise model for dnn text classification. **arXiv preprint arXiv:1903.07507**, 2019. Citado na página 25.

KAHNEMAN, D.; SIBONY, O.; SUNSTEIN, C. R. **Noise: A flaw in human judgment**. [S.l.]: Hachette UK, 2021. Citado na página 24.

NARIMANZADEH, H.; BADIE-MODIRI, A.; SMIRNOVA, I. G.; CHEN, T. H. Y. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. **Proceedings of the ACM on Human-Computer Interaction**, ACM New York, NY, USA, v. 7, n. CSCW2, p. 1–29, 2023. Citado na página 24.

NGUYEN, Q.; SHIKINA, T.; TERUYA, D.; HOTTA, S.; HAN, H.-D.; NAKAJO, H. Leveraging expert knowledge for label noise mitigation in machine learning. **Applied Sciences**, MDPI, v. 11, n. 22, p. 11040, 2021. Citado na página 25.

PAIXÃO, G. M. de M.; SANTOS, B. C.; ARAUJO, R. M. de; RIBEIRO, M. H.; MORAES, J. L. de; RIBEIRO, A. L. Machine learning na medicina: Revisão e aplicabilidade. **Arquivos Brasileiros de Cardiologia**, SciELO Brasil, v. 118, n. 1, p. 95, 2022. Citado na página 23.

RAYKAR, V. C.; YU, S.; ZHAO, L. H.; VALADEZ, G. H.; FLORIN, C.; BOGONI, L.; MOY, L. Learning from crowds. **Journal of Machine Learning Research**, v. 11, n. 4, 2010. Citado nas páginas 24, 25, 39, 40, 42, 44, 45, 47, 50, 52, 53, 56 e 58.

RODRIGUES, F.; LOURENCO, M.; RIBEIRO, B.; PEREIRA, F. C. Learning supervised topic models for classification and regression from crowds. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 39, n. 12, p. 2409–2422, 2017. Citado na página 25.

SANTOS, M. R. d.; IZBICKI, R. Expertise-based weighting for regression models with noisy labels. **arXiv preprint arXiv:2305.07430**, 2023. Citado na página 26.

SMYTH, P.; FAYYAD, U.; BURL, M.; PERONA, P.; BALDI, P. Inferring ground truth from subjective labelling of venus images. The MIT Press, 1995. Citado na página 25.

SOUSA, T. A. O. d. Modelo de risco para provisão judicial: método quantitativo para aplicação em instituição financeira. 2022. Citado na página 23.

TANNO, R.; SAEEDI, A.; SANKARANARAYANAN, S.; ALEXANDER, D. C.; SILBERMAN, N. Learning from noisy labels by regularized estimation of annotator confusion. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2019. p. 11244–11253. Citado na página 25.

WANG, Y.; SUN, X.; FU, Y. Scalable penalized regression for noise detection in learning with noisy labels. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2022. p. 346–355. Citado na página 25.

WASSERMAN, L. **All of nonparametric statistics**. [S.l.]: Springer Science & Business Media, 2006. Citado na página 26.

XIAO, H.; XIAO, H.; ECKERT, C. Learning from multiple observers with unknown expertise. In: SPRINGER. **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2013. p. 595–606. Citado na página 25.

XU, X.; FRANK, E. Logistic regression and boosting for labeled bags of instances. In: SPRINGER. **Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings 8**. [S.l.], 2004. p. 272–281. Citado na página 24.

YAN, Y.; ROSALES, R.; FUNG, G.; DY, J. G. Modeling multiple annotator expertise in the semi-supervised learning scenario. In: **Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)**. [S.l.: s.n.], 2010. Citado nas páginas 24 e 25.

ZHANG, J. Knowledge learning with crowdsourcing: a brief review and systematic perspective. **IEEE/CAA Journal of Automatica Sinica**, IEEE, v. 9, n. 5, p. 749–762, 2022. Citado na página 25.

ZHENG, Y.; LI, G.; LI, Y.; SHAN, C.; CHENG, R. Truth inference in crowdsourcing: Is the problem solved? **Proceedings of the VLDB Endowment**, VLDB Endowment, v. 10, n. 5, p. 541–552, 2017. Citado na página 25.


ICMC
USP
SÃO CARLOS