



# Modelação do Processo de Afinação de Tintas através de Algoritmos de Machine Learning

CATARINA MORIM DA COSTA

setembro de 2023

POLITÉCNICO DO PORTO  
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

---

Modelação do Processo de Afinação de  
Tintas através de Algoritmos de  
*Machine Learning*

---

Catarina Morim da Costa

Mestrado em Engenharia e Gestão Industrial



DEPARTAMENTO DE ENGENHARIA MECÂNICA  
Instituto Superior de Engenharia do Porto

Setembro, 2023



*Esta dissertação satisfaz, parcialmente, os requisitos que constam da Ficha de Unidade Curricular de Tese/Dissertação, do 2º ano, do Mestrado em Engenharia e Gestão Industrial.*

**Candidato:** Catarina Morim da Costa, N.º 1180788, 1180788@isep.ipp.pt

**Orientação Científica:** Carlos Manuel Abreu Gomes Ferreira,  
cgf@isep.ipp.pt

**Empresa:** Gloma – Comércio, Indústria e Representações de Pinturas  
Ecológicas, Lda ®

**Orientador:** Sónia Martins, soniamartins@gloma.pt



DEPARTAMENTO DE ENGENHARIA MECÂNICA  
Instituto Superior de Engenharia do Porto  
Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto

Setembro, 2023



*À minha estrelinha guia,  
cuja luz continua a brilhar no meu coração!*



# Agradecimentos

Não poderia deixar de expressar a minha profunda gratidão a todas as pessoas que tornaram possível a concretização deste projeto de investigação. A jornada que aqui termina foi enriquecida pelo seu apoio e dedicação, pelo que não poderiam deixar de ser mencionados aqueles cujas contribuições foram essenciais para o meu crescimento.

Inicialmente, gostaria de agradecer ao Instituto Superior de Engenharia do Porto (ISEP) por proporcionar a oportunidade de realizar este projeto, permitindo-me aplicar eficazmente os conhecimentos adquiridos e explorá-los num contexto mais realista do mercado de trabalho. E, neste sentido, desejo expressar a minha sincera apreciação ao meu orientador científico, o Doutor Carlos Manuel Abreu Gomes Ferreira pela sua orientação, disposição constante para esclarecer dúvidas e por todo o apoio ao longo deste percurso, direcionando este projeto a “bom porto”. E, além disso, ao Diretor de Curso, Engenheiro Manuel Pereira Lopes, pela abordagem visionária que nos prepara para os desafios do futuro.

Por conseguinte, agradecer à Gloma® por aprovar a realização do projeto e mais precisamente à minha orientadora, Sónia Martins, por acompanhar todo o processo. Não podia esquecer de agradecer a toda a equipa da Gloma®, que se tornou um pilar fundamental na minha jornada profissional.

Neste caminho, não posso deixar de mencionar os amigos e companheiros de jornada que sempre se mostraram disponíveis para me ajudar e motivar, durante o percurso académico.

Aos meus pais e irmã, por acreditarem em mim e pelo apoio incondicional em cada etapa da minha vida.

Aos meus avós, que tornaram o meu objetivo num sonho deles.

A todos, o meu muito obrigada!



# Resumo

Considerando o constante esforço em aprimorar a eficiência dos processos industriais, esta pesquisa direciona o seu foco para a convergência entre Engenharia e Gestão Industrial e a colorimetria. No âmbito deste estudo, ressalta a necessidade premente de melhorar o processo de afinação de tintas, no qual a capacidade de prever a proporção ideal de cada base de tinta para alcançar uma tonalidade específica se configura como um desafio complexo e relevante.

As bases de tinta, essência da paleta cromática, oferecem uma gama diversificada de tonalidades por meio de combinações variadas. O processo de identificar as bases a integrar na mistura e as suas proporções é uma questão de destaque na colometria, frequentemente apresentando múltiplas soluções devido à natureza intrincada da cor.

Este estudo adota abordagens sustentáveis de *Machine Learning* para determinar fórmulas de tintas a partir de características específicas da cor e acabamento da tinta, oferecendo uma solução inovadora aos desafios da colorimetria. Simultaneamente, esta pesquisa promove a colaboração entre a Inteligência Artificial e o setor industrial, melhorando a eficiência em aplicações essenciais.

Assim, o objetivo primordial deste projeto reside em desenvolver, no mínimo, um modelo de previsão de fórmulas de tinta capaz de alcançar a cor especificada com base nos parâmetros  $L$ ,  $a$  e  $b$  do sistema de cor CIELab.

Esta dissertação adota a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) como estrutura orientadora. Por meio da exploração da sua natureza adaptativa, os algoritmos de *Machine Learning* demonstram um potencial substancial para melhorar não só a eficiência, mas também a precisão do processo de afinação de cores.

Este estudo empreendeu uma análise criteriosa de três distintas abordagens, com foco na avaliação da consistência e da precisão nas previsões realizadas. A primeira abordagem pautou-se na modelação clássica das fórmulas de tintas baseada em algoritmos únicos. No entanto, diferentes modelos destacaram-se na previsão de distintas bases de tinta e, por isso mesmo, na segunda abordagem, adotou-se uma estratégia que envolve a combinação dos modelos mais adequados para cada base com o propósito de melhorar a eficácia global das previsões. Por fim, na terceira abordagem, dividiu-se o conjunto de dados em *clusters*, para avaliar a precisão das previsões das fórmulas de cores, explorando a especialização dos modelos.

Entre os algoritmos avaliados, o *Artificial Neural Network with a pre-trained Decoder* (ANND) da primeira abordagem, apoiado na técnica de *Transfer Learning*, destacou-se como o mais promissor. Este demonstrou um desempenho superior tanto em termos de Erro Absoluto Médio (MAE) quanto de Raiz do Erro Quadrático Médio (RMSE), sobressaindo-se particularmente na previsão das bases de tinta mais complexas.

Adicionalmente, ao realizar uma análise individual para cada *cluster*, destacaram-se diferenças notáveis, onde cores mais claras mostraram-se mais facilmente modeladas, enquanto cores mais escuras apresentaram uma maior complexidade. Portanto, ao lidar com cores mais claras, a escolha de um modelo específico pode ser uma estratégia mais eficaz. Foram ainda realizados testes experimentais, em ambiente empresarial, que reforçaram a aplicabilidade deste estudo, evidenciando o seu potencial ao formular tintas conforme as especificações de cor e acabamento desejadas.

Esta dissertação, além de enriquecer o conhecimento na área de pesquisa, destaca inequivocamente o potencial das abordagens de *Machine Learning* na afinação de tintas. Assim, deixa uma contribuição importante na esfera da formulação de tintas, destacando a importância da análise de dados e da modelação como meios eficazes de enfrentar os desafios perenes desta indústria em constante evolução.

**Palavras-Chave:** Afinação de Tintas, Espaço de cor CIELab, *Machine Learning*, *Multi-Target Regression*, *Clustering*, Combinação de Modelos, *Transfer Learning*

# Abstract

Considering the ongoing effort to enhance industrial processes' efficiency, this research focuses on the convergence between Industrial Engineering and Management and colorimetry. At the heart of this study lies the pressing need to improve the paint formulation process, where the ability to predict the ideal proportion of each ink base to achieve a specific shade is a complex and relevant challenge.

Ink bases, the essence of the color palette, offer a diverse range of shades through various combinations. The process of identifying the bases to integrate into the mixture and their proportions is a prominent issue in colorimetry, often presenting multiple solutions due to the intricate nature of color.

This study adopts sustainable Machine Learning approaches to determine paint formulas based on specific color and finish characteristics, offering an innovative solution to colorimetry challenges. Simultaneously, this research promotes collaboration between Artificial Intelligence and the industrial sector, enhancing efficiency in essential applications.

Thus, the primary objective of this project is to develop, at least, one paint formula prediction model capable of achieving the specified color based on the CIELab color system's  $L$ ,  $a$ , and  $b$  parameters.

This dissertation adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology as a guiding framework. Through the exploration of its adaptive nature, Machine Learning algorithms demonstrate substantial potential to improve the efficiency and accuracy of the color tuning process.

This study undertook a thorough analysis of three distinct approaches, focusing on evaluating consistency and precision in predictions. The first approach relied on classic modeling of paint formulas, based on single algorithms. However, different models excelled in predicting different ink bases, leading to the adoption of a strategy in the second approach that involved combining the most suitable models for each base to enhance overall prediction effectiveness. Finally, in the third approach, the dataset was divided into clusters to assess the accuracy of color formula predictions, exploring model specialization.

As the most promising method, the Artificial Neural Network with a pre-trained Decoder (ANND) from the first approach, based on Transfer Learning, demonstrated superior performance in terms of both Mean Absolute Error (MAE) and Root Mean

Square Error (RMSE), particularly excelling in predicting the most challenging ink bases.

Additionally, through individual analysis for each cluster, notable differences emerged, with lighter colors proving easier to model, while darker colors exhibited greater complexity. Therefore, when dealing with lighter colors, selecting a specific model may be a more effective strategy.

Furthermore, experimental tests were conducted in a business environment, further confirming the applicability of this study and highlighting its potential in formulating paints according to desired color and finish specifications.

In addition to enriching knowledge in the research field, this dissertation unequivocally emphasizes the potential of Machine Learning approaches in paint formulation. Thus, it makes a significant contribution to the ink tuning sphere, underscoring the importance of data analysis and modeling as effective means to address the ongoing challenges in this ever-evolving industry.

**Keywords:** Ink Tuning, Color Space CIELab, Machine Learning, Multi-Target Regression, Clustering, Ensemble Learning, Transfer Learning

# Índice

<b>Lista de Figuras</b>	<b>vii</b>
<b>Lista de Tabelas</b>	<b>ix</b>
<b>Lista de Acrónimos</b>	<b>xi</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Problema de investigação, enquadramento e pertinência . . . . .	1
1.2 Questão e objetivos de investigação . . . . .	3
1.3 Opções metodológicas . . . . .	3
1.4 Apresentação da Empresa . . . . .	4
1.5 Inovações e Contribuições . . . . .	5
1.6 Estrutura do Trabalho . . . . .	5
<b>2 Revisão Bibliográfica</b>	<b>7</b>
2.1 Fundamentos de Tinta e do Processo de Afinação . . . . .	7
2.2 Colorimetria . . . . .	9
2.2.1 Espaço de Cor RGB . . . . .	10
2.2.2 Espaço de Cor HSV . . . . .	10
2.2.3 Espaço de Cor CIELab . . . . .	10
2.3 <i>Machine Learning</i> na Previsão das Fórmulas de Afinação de Tinta .	11
2.3.1 <i>Linear Regression</i> . . . . .	12
2.3.2 <i>MultiLayer Perceptron</i> . . . . .	13
2.3.3 <i>Autoencoder-based Transfer Learning</i> . . . . .	14
2.3.4 <i>Random Forest</i> . . . . .	15
2.3.5 <i>Extreme Gradient Boosting</i> . . . . .	16
2.3.6 <i>Clustering</i> e Aplicação dos Modelos aos <i>Clusters</i> . . . . .	17
2.3.7 <i>Ensemble Learning</i> . . . . .	17
2.4 Trabalhos Relacionados . . . . .	18
<b>3 Metodologia</b>	<b>23</b>
3.1 <i>Business Understanding</i> . . . . .	24
3.2 <i>Data Understanding</i> . . . . .	25
3.3 <i>Data Preparation</i> . . . . .	30

3.4	<i>Modeling</i> . . . . .	31
3.4.1	Combinação de Modelos . . . . .	33
3.4.2	Divisão dos dados em <i>Clusters</i> . . . . .	34
3.5	<i>Evaluation</i> . . . . .	36
<b>4</b>	<b>Resultados e Discussão</b>	<b>37</b>
4.1	Modelação Clássica baseada em Algoritmos Únicos . . . . .	38
4.2	Combinação dos Melhores Modelos para cada Base de Tinta . . . . .	39
4.3	Modelação baseada em <i>Clusters</i> do Espaço de Cor CIELab . . . . .	41
4.4	Discussão Geral dos Resultados . . . . .	43
4.5	Comparação da Modelação entre <i>Clusters</i> na Afinação de Tintas . . . . .	46
4.6	Uma Abordagem Prática da Afinação de Tintas com <i>Machine Learning</i> . . . . .	47
<b>5</b>	<b>Conclusões</b>	<b>51</b>
5.1	Trabalho Futuro . . . . .	53
	<b>Referências</b>	<b>54</b>
	<b>Apêndice A Estatísticas dos Dados</b>	<b>67</b>
	<b>Apêndice B Complemento aos Resultados e Discussão</b>	<b>71</b>
B.1	Conjunto de treino dos modelos da 1ª Abordagem . . . . .	71
B.2	Complemento à Comparação da Modelação entre <i>Clusters</i> . . . . .	72
	<b>Apêndice C Complemento às Considerações Finais</b>	<b>79</b>

# Lista de Figuras

1.1	Logotipo da Gloma® . . . . .	4
2.1	<i>Munsell Ordering System</i> . . . . .	9
2.2	Diferentes Tonalidades . . . . .	9
2.3	Diagrama de espaço de cor CIELab . . . . .	11
2.4	Representação de MLP . . . . .	14
2.5	Esquema de um <i>Perceptron</i> com $n$ dados de entrada . . . . .	14
2.6	Esquema de um <i>Autoencoder</i> . . . . .	15
2.7	Esquema de um RF . . . . .	16
2.8	<i>Clustering</i> em 4 grupos no espaço . . . . .	17
2.9	<i>Ensemble</i> de Modelos . . . . .	18
3.1	Metodologia CRISP-DM . . . . .	24
3.2	Gráfico do Conjunto de Dados . . . . .	27
3.3	Histograma da Base <i>Amarelo Vivo</i> . . . . .	28
3.4	Histograma da Base <i>Branco</i> . . . . .	28
3.5	Matriz de Correlação . . . . .	29
3.6	Arquitetura do <i>Autoencoder</i> . . . . .	32
3.7	Conjunto de dados dividido em quatro subgrupos . . . . .	35
4.1	Comparação dos resultados do desempenho dos modelos no conjunto de dados completo . . . . .	39
4.2	Comparação dos resultados do desempenho dos modelos nos <i>clusters</i> . . . . .	43
4.3	Comparação dos resultados do desempenho das diferentes abordagens . . . . .	45
A.1	Resultados das Estatísticas dos Dados - Parte 1 . . . . .	68
A.2	Resultados das Estatísticas dos Dados - Parte 2 . . . . .	69
B.1	Comparação dos resultados do desempenho dos modelos no <i>Cluster A</i> . . . . .	73
B.2	Comparação dos resultados do desempenho dos modelos no <i>Cluster B</i> . . . . .	75
B.3	Comparação dos resultados do desempenho dos modelos no <i>Cluster C</i> . . . . .	76
B.4	Comparação dos resultados do desempenho dos modelos no <i>Cluster D</i> . . . . .	78
C.1	Relatório do teste Cinza . . . . .	80
C.2	Relatório do teste Azul . . . . .	81

C.3 Relatório do teste Amarelo . . . . .	82
--	----

# Lista de Tabelas

3.1	Exemplo de uma observação . . . . .	26
3.2	Contagem de Observações por Tonalidade . . . . .	27
3.3	Processo de Padronização das Fórmulas de Tinta . . . . .	30
3.4	Parametrização de cada Modelo . . . . .	33
3.5	Características dos <i>Clusters</i> . . . . .	35
4.1	Métricas de Desempenho dos Modelos no Conjunto de Teste . . . . .	38
4.2	Escolha dos Modelos para integrarem o <i>Ensemble</i> . . . . .	40
4.3	Métricas de Desempenho da Combinação de Modelos no Conjunto de Teste . . . . .	41
4.4	Métricas de Desempenho dos Modelos aplicados aos <i>Clusters</i> no Conjunto de Teste . . . . .	42
4.5	Resumo das Métricas de Desempenho dos Modelos no Conjunto de Teste dos <i>Clusters</i> . . . . .	46
4.6	Resultados da Afinação de Tinta com base em Previsões do Modelo RF . . . . .	48
4.7	Comparação dos Resultados obtidos com as Cores Pretendidas . . . . .	48
5.1	Resumo dos Resultados das Abordagens . . . . .	52
B.1	Métricas de desempenho dos modelos no Conjunto de Treino . . . . .	72
B.2	Métricas de desempenho dos modelos no Conjunto de Teste do <i>Cluster</i> A . . . . .	73
B.3	Métricas de desempenho dos modelos no Conjunto de Teste do <i>Cluster</i> B . . . . .	74
B.4	Métricas de desempenho dos modelos no Conjunto de Teste do <i>Cluster</i> C . . . . .	76
B.5	Métricas de desempenho dos modelos no Conjunto de Teste do <i>Cluster</i> D . . . . .	77



# Lista de Acrónimos

$R^2$	Coeficiente de Correlação
ANN	<i>Artificial Neural Network</i>
ANND	<i>Artificial Neural Network with a pre-trained Decoder</i>
B	Azul
BG	Azul-Verde
CIE	<i>Comission Internationale de l'Eclaiage</i>
CNN	<i>Convolutional Neural Network</i>
CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i>
CV	<i>Cross-Validation</i>
DNN	<i>Deep Neural Network</i>
DT	<i>Decision Tree</i>
G	Verde
GY	Verde-Amarelo
HSV	<i>Hue, Saturation, Value</i>
ISEP	Instituto Superior de Engenharia do Porto
LR	<i>Linear Regression</i>
MAE	Erro Absoluto Médio
MLP	<i>MultiLayer Perceptron</i>
P	Púrpura
PB	Púrpura-Azul
R	Vermelho
RF	<i>Random Forest</i>

<b>RGB</b>	<i>Red, Green, Blue</i>
<b>RMSE</b>	Raiz do Erro Quadrático Médio
<b>RP</b>	Vermelho-Púrpura
<b>XGBoost</b>	<i>Extreme Gradient Boosting</i>
<b>Y</b>	Amarelo
<b>YR</b>	Amarelo-Vermelho

## Capítulo 1

# Introdução

A área da Engenharia e Gestão Industrial tem demonstrado uma crescente preocupação em desenvolver metodologias avançadas que aprimorem a eficiência e a sustentabilidade dos processos industriais. Tal abordagem implica estar em constante evolução, alinhada com as necessidades da indústria a fim de elevar os padrões de excelência e competitividade no domínio industrial.

O presente capítulo aborda o problema de investigação e as questões que esta dissertação se propõe a responder, com explicação do contexto envolvente e da sua relevância para a comunidade científica. Posteriormente, são abordados os objetivos e é formulada a questão de investigação. São ainda descritas as opções metodológicas que serão seguidas. Por fim, será feita a apresentação da empresa que acolheu este projeto e será delineada a estrutura global da dissertação.

### **1.1 Problema de investigação, enquadramento e pertinência**

Dado o contínuo empenho em melhorar e tornar mais eficientes os processos industriais, a presente dissertação tem como propósito abordar um desafio que reside na interseção entre a Engenharia e Gestão Industrial e a arte da coloração: a previsão eficaz da proporção das bases de tinta a misturar para obter uma cor específica. O processo de afinação de tintas e o próprio conceito de Cor são críticos e essenciais em vários domínios da investigação [1] e para a indústria de tintas [2], de plásticos

[3], têxtil [4], de pintura [5], de impressão [6] e outras áreas afins, tendo um impacto significativo na qualidade dos produtos finais e na satisfação dos clientes.

K. Fehrman & C. Fehrman [7] consideram as bases de tinta a essência da paleta cromática, permitindo a criação de uma ampla variedade de tonalidades por meio da combinação em proporções distintas. Essa questão de determinar as bases de tinta e suas proporções sempre teve um grande destaque no campo da colometria e, dada a natureza complexa da Cor, frequentemente surgem múltiplas soluções possíveis, já que diferentes combinações de tintas podem resultar na mesma tonalidade.

Até meados da década de 1960, a formulação de cores era fortemente dependente da intuição e do conhecimento acumulado de afinadores experientes [8]. Posteriormente, a previsão de receita de tinta foi apresentada comercialmente, tornando-se uma das aplicações industriais mais importantes da indústria da Cor [9]. Segundo P. R. Bélanger [8], um marco significativo surgiu quando o programa desenvolvido por Preston & Tough começou a calcular as concentrações necessárias para obter uma cor específica, considerando um conjunto particular de três bases. Também Allen, na empresa *American Cyanamid*, expandiu esse conceito ao criar um programa que trabalhava com um conjunto maior, composto por até doze bases de tinta, onde se procurava solucionar as equações de cor para todas as combinações de três bases.

Considerando a crescente utilização de *Machine Learning*, uma subárea da Inteligência Artificial, este projeto tem como principal intuito desenvolver um modelo que possibilite a afinação de tinta com base nos parâmetros de cor e no histórico de dados dos últimos cinco anos, visando aprimorar a gestão sustentável de recursos, quer ao nível de tinta, quer ao nível de tempo. Os modelos propostos procuram minimizar os erros na reprodução de cores em termos de diferenças colorimétricas e espectrais.

A relevância desse problema é reforçada pelo panorama atual do mercado, onde a disponibilidade de *softwares* para o cálculo da proporção de bases de tinta é reduzida, exigindo frequentemente investimentos consideráveis. Tal incentiva à criação de alternativas viáveis e acessíveis que possam otimizar a previsão de receitas de cores.

A motivação intrínseca consiste justamente na existência desta lacuna em relação a soluções acessíveis para prever a proporção exata das bases de tinta, representando não só um desafio prático como também uma oportunidade para a inovação e contribuição científica. Assim, esta dissertação surge como uma resposta inovadora, oferecendo uma alternativa viável, eficaz e de qualidade para a previsão de receitas de cores.

O desenvolvimento dessas abordagens sustentáveis é de grande importância para a economia das empresas e, portanto, a criação de uma ferramenta de previsão de fórmulas de cores e a estimativa das quantidades de bases de tintas é altamente relevante. Por esse motivo, esta pesquisa está na vanguarda da previsão de receitas de tintas, fornecendo *insights* valiosos e soluções práticas para desafios na indústria.

## 1.2 Questão e objetivos de investigação

Com o desenvolvimento deste trabalho, a presente dissertação procura dar resposta à seguinte questão: De que forma os parâmetros do espaço de cor CIELab e o acabamento da pintura se relacionam com a proporção de bases de tinta a misturar para obter a cor pretendida?

Deste modo, o objetivo geral centra-se em desenvolver, pelo menos, um modelo de previsão da proporção das dezasseis bases de tinta para obter a cor requisitada pelo cliente, a partir dos parâmetros  $L$ ,  $a$  e  $b$  do sistema CIELab obtidos pelo espectrofotómetro<sup>1</sup>.

Pretende-se que a resolução do problema seja alcançada através de *Machine Learning*, culminando no final com a avaliação dos resultados. Neste sentido, os objetivos específicos orientadores são:

- Interpretar o processo atualmente realizado para afinação de tintas e as circunstâncias envolventes;
- Identificar as necessidades e limitações do processo;
- Conceber modelos e otimizar os seus parâmetros;
- Avaliar os resultados obtidos e escolher o melhor modelo;
- Experimentar o modelo em cores de teste;
- Analisar a conformidade da cor com o modelo desenvolvido.

## 1.3 Opções metodológicas

Este capítulo tem como intuito a apresentação das opções metodológicas consideradas para a realização da dissertação. De encontro ao mencionado anteriormente, a abordagem de investigação passa pela perspetiva quantitativa, na medida em que se trata de um estudo objetivo e suportado por métodos científicos para prever a receita de tinta, testando resultados para os indicadores de desempenho mais apropriados ao problema e a posterior análise estatística [11].

Esta perspetiva é suportada por um modelo hipotético-dedutivo, assumindo que o problema de investigação alcança uma solução objetiva a partir de um método científico. No que concerne às técnicas de recolha de dados, estas incidem sobre a análise de documentos, desde dissertações de mestrado a artigos científicos da área em questão.

---

<sup>1</sup>Um espectrofotómetro é um aparelho utilizado para medir a quantidade de luz absorvida ou transmitida por uma substância em função do seu comprimento de onda. É amplamente utilizado em várias aplicações industriais e científicas [10].

No âmbito dos métodos de investigação, este projeto insere-se no paradigma do estudo de caso, caracterizado por uma abordagem intensiva e minuciosa de uma entidade claramente definida. Isso implica uma análise aprofundada de uma situação específica no contexto real [11]. Portanto, o trabalho inicia-se com uma base teórica sólida e uma descrição detalhada do problema em análise, visando uma contextualização eficaz. Em seguida, são desenvolvidos modelos de *Machine Learning*, fazendo uso dos dados históricos de afinação de cores, com o objetivo de alcançar o objetivo estabelecido.

## 1.4 Apresentação da Empresa

A Gloma – Comércio, Indústria e Representações de Pinturas Ecológicas, Lda ® é uma empresa especialista no tratamento de superfície, com uma ampla experiência no segmento dos pequenos acessórios, com sede em Vila do Conde.



Figura 1.1: Logotipo da Gloma®

Desde 1997, a empresa presta serviços de pintura para diversas indústrias, entre elas, Indústria Têxtil, Construção Civil, Metalúrgica e Cutelaria. A empresa alcança sucesso no mercado através da sua determinação e criatividade, superando constantemente os desafios a que é sujeita [12].

Além disso, a Gloma® é conhecida por oferecer serviços de alta qualidade, consoante as normas de qualidade internas, regulamentos legais aplicáveis e as normas internacionais de segurança para o serviço.

A cor, sendo a essência do negócio, desempenha um papel vital no âmbito do negócio da Gloma®, uma vez que a precisão na obtenção da cor desejada não é apenas um requisito técnico, mas uma manifestação tangível da qualidade e excelência que a Gloma® procura oferecer aos clientes.

Nesse contexto, o objetivo central desta dissertação ganha ainda mais destaque, na medida em que a pesquisa se alinha às necessidades e valores da empresa, enquanto também oferece uma solução para aprimorar a eficiência e a precisão do processo de afinação.

## 1.5 Inovações e Contribuições

Os esforços para aprimorar a previsão de afinação de tintas têm levado ao desenvolvimento de modelos de *Machine Learning*. Esses modelos oferecem flexibilidade e conseguem efetivamente compreender relações complexas entre as bases de tinta e as cores desejadas [13]. Ao aproveitar a adaptabilidade e a não linearidade dos algoritmos de *Machine Learning*, os modelos possuem um potencial significativo para melhorar a eficiência e a precisão dos processos de afinação.

Apesar da literatura nesta área ser escassa, a proposta inovadora deste estudo procura não apenas contribuir para o avanço do conhecimento no campo da colorimetria, mas também automatizar o desenvolvimento de fórmulas de tinta ao considerar as características de cor e os acabamentos desejados.

O trabalho aqui apresentado posiciona-se como uma solução para os desafios complexos que permeiam a indústria de formulação de tintas. Além disso, abre também novas perspectivas para futuros estudos e progressos, ao promover uma colaboração mais próxima entre o domínio da Inteligência Artificial e as aplicações tangíveis na indústria, melhorando processos e impulsionando a eficiência de um setor essencial em diversas áreas.

Antes deste estudo, foi conduzida uma pesquisa sobre as capacidades dos modelos de *Machine Learning* na previsão de afinação de tintas. Esse trabalho passou por um rigoroso processo de revisão e validação por especialistas da comunidade científica, culminando na sua aceitação para publicação. Os seus resultados constituem a base sólida sobre a qual se constrói esta dissertação. Com este compromisso contínuo com a excelência e a validação da investigação, espera-se contribuir ainda mais para o campo da colorimetria e da Inteligência Artificial..

## 1.6 Estrutura do Trabalho

Esta dissertação segue uma estrutura organizacional composta por cinco capítulos adicionais, além da Introdução. O próximo capítulo, intitulado *Revisão Bibliográfica*, apresenta o estado da arte das ciências da cor e da formulação de tinta, com foco particular nos métodos de *Machine Learning*. Após a fundamentação teórica, o capítulo *Metodologia* delinea o estudo de caso deste projeto e a estratégia metodológica adotada, proporcionando detalhes acerca dos dados utilizados e dos métodos aplicados.

No capítulo subsequente, denominado *Resultados e Discussão*, são apresentados, avaliados e comparados os resultados obtidos para cada abordagem explorada. Por último, o capítulo *Conclusões* sintetiza as etapas percorridas no decorrer do projeto, analisando os resultados alcançados e delineando as fundamentações subjacentes a

esses resultados. Adicionalmente, são propostas algumas sugestões para desenvolvimentos futuros. No seguimento destes capítulos, nos *Apêndices*, são disponibilizadas informações adicionais, incluindo gráficos, detalhes complementares e recursos que desempenharam um papel crucial no desenvolvimento do projeto.

## Capítulo 2

# Revisão Bibliográfica

O presente capítulo destina-se à revisão bibliográfica relativa ao problema abordado, estando dividido em quatro subtópicos. No primeiro, faz-se o enquadramento do assunto-base, nomeadamente os fundamentos da tinta e o processo de afinação, aliados à descrição de diferentes sistemas de cor e dos instrumentos de medição, no segundo subtópico. De seguida, são apresentados diferentes modelos e técnicas de *Machine Learning* para a previsão da receita de tinta. O quarto subtópico faz referência a trabalhos desenvolvidos por diferentes autores no âmbito deste problema, de forma a analisar a estratégia aplicada e os diferentes resultados.

### 2.1 Fundamentos de Tinta e do Processo de Afinação

Segundo a norma ISO 4618:2023 [14], uma tinta é definida como um material pigmentado, líquido, em pasta ou em pó, que, quando aplicado sobre um substrato, forma uma película opaca com propriedades protetoras, decorativas e/ou outras propriedades específicas.

Conforme mencionado por Laurence W. McKeen [15], a composição da tinta envolve quatro componentes primordiais:

- Cargas e pigmentos: As cargas conferem propriedades físicas específicas, como controlo do brilho, melhoria das propriedades mecânicas, resistência à corrosão, entre outros. Os pigmentos, geralmente insolúveis no meio em que são incorporados, conferem cor, poder de cobertura e opacidade à tinta. Embora

também existam corantes, substâncias solúveis que proporcionam coloração, mantendo a transparência, não se encontra no âmbito desta dissertação.

- Aditivos: A tinta pode ser enriquecida por aditivos para modificar e/ou controlar determinadas propriedades da tinta. Conforme o objetivo, pode-se conferir diferentes efeitos no produto final, nomeadamente a modificação da viscosidade, resistência a raios UV e propriedades antifúngicas, entre outros.
- Ligantes: O ligante tem como propósito manter as substâncias da mistura agregadas.
- Solventes: Os solventes são compostos geralmente líquidos, à temperatura e pressão atmosférica, que atuam como meio de transporte e conseguem dissolver outras substâncias sem alterar as suas propriedades químicas [16].

Dessa forma, a complexa mistura harmoniza-se para criar uma tinta funcional e versátil, capaz de conferir cor, propriedades físicas e características distintas aos produtos finais. À medida que o solvente se evapora, uma película sólida é formada graças ao ligante, contribuindo para a aderência, o acabamento, a durabilidade, a flexibilidade e a resistência.

Uma vez desenvolvidas as bases de tinta, cada uma com a sua tonalidade específica, tornam-se os elementos fundamentais para a criação de uma paleta diversificada de cores. Para alcançar diferentes cores, é necessário combinar essas bases. Esse processo, denominado afinação de tintas, envolve a criação de uma mistura específica de bases de tinta para obter a cor desejada, onde a experiência dos afinadores desempenha um papel vital. Uma pequena variação na quantidade de qualquer base pode levar a uma diferença perceptível na cor final do produto, sendo especialmente crítico em setores onde a cor é uma parte central da identidade da marca ou do produto. A capacidade de reproduzir cores consistentemente é importante para a qualidade dos produtos. Além disso, a eficiência na afinação de tintas tem implicações diretas na produtividade e no custo final [16].

Nos últimos anos, avanços tecnológicos têm impactado positivamente o processo de afinação de tintas. A crescente procura por soluções mais objetivas impulsiona a utilização de instrumentos de medição de cor e os avanços tecnológicos de análise colorimétricas, em detrimento da acuidade visual e da natureza subjetiva do processo de seleção da cor [17]. Espectrofotômetros, dispositivos de medição de cor, admitem medir com precisão a refletância espectral e a transmitância de um estímulo de cor [10], sendo que, comparativamente ao método manual, contribui com melhorias na investigação da cor, no controlo de qualidade e na eficácia/objetividade da comunicação.

## 2.2 Colorimetria

Cor é um conceito abstrato que, segundo Ishikawa-Nagai et al. [18], consiste no atributo da percepção visual produzida pelo feixe luminoso da combinação de componentes cromáticos e acromáticos. Efetivamente, a sua definição, pelo *Munsell Ordering System*, assenta na conjugação de três parâmetros fundamentais, nomeadamente, tonalidade (HUE), luminosidade (VALUE) e croma (CHROMA) [19], apresentando as cores no espaço tridimensional apresentado na figura 2.1.

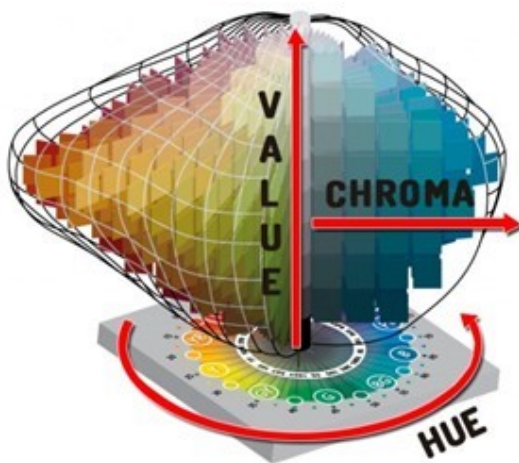


Figura 2.1: *Munsell Ordering System*  
Fonte: Lori Sawaya [20]



Figura 2.2: Diferentes Tonalidades  
Fonte: Pegalajar et al. [21]

Neste sistema, as tonalidades expressam-se na superfície do sólido tridimensional, enquanto a luminosidade varia ao longo do eixo vertical e a croma intensifica-se à medida que o raio aumenta [19]. Os estudos fisiológicos citados por Zeileis et al. [22] apoiam esta teoria já que a luz é codificada em três dimensões pelo olho humano.

A cor está estritamente correlacionada com a distribuição espectral da luz visível. Nesse sentido, o princípio da ciência da cor sustenta-se no método *Tristimulus XYZ*, uma representação matemática derivada da sensibilidade do olho humano, onde os sensores recebem a luz refletida de um objeto e transmitem a informação [19]. Essa informação resulta da combinação de dados luminosos, dos valores de refletância e das funções de correspondência de cor [23].

A necessidade de quantificar e comunicar cores univocamente deu origem à colorimetria, a ciência e o conjunto de tecnologias envolvidas na caracterização da percepção de cores pelos seres humanos em diversas áreas tecnológicas [24]. Para evitar ambiguidades nas descrições de cores, foram criados espaços de cor, como RGB, HSV e CIE Lab, que fornecem uma linguagem universal baseada em modelos matemáticos [16]. Esses espaços garantem uma compreensão comum e consistente da cor em diferentes contextos e aplicações, facilitando a comunicação das especificações de cor desejadas.

### 2.2.1 Espaço de Cor RGB

De acordo com Dissaneevate et al. [25], o sistema RGB é baseado na Teoria Tricromática da Cor de Young-Helmholtz<sup>1</sup>, uma vez que processa uma cor como o resultado da combinação aditiva de tonalidades na escala de cinzentos que se adaptam à intensidade da luz vermelha, verde e azul. Este é presumivelmente o mapa de representação visual mais conhecido, no entanto Stauffer et al. [27] considera-o de difícil interpretação e prejudicial, já que o espectro do espaço de cor RGB apresenta bruscas diferenças de brilho, cores vívidas e não é uniforme.

### 2.2.2 Espaço de Cor HSV

O espaço de cor HSV, também baseado na Teoria Tricromática da Cor de Young-Helmholtz, deriva dos conceitos de Tonalidade, Saturação e Intensidade, sendo que Dissaneevate et al. [25] e Li & Yuen [28] afirmam que este é um modelo estável, coerente e intuitivo para a percepção visual humana. Apesar disso, as cores HSV sofrem dos mesmos problemas que o espaço de cor RGB, ou seja, apresentam um mapa de três dimensões atônito, onde a intensidade não é uniforme sobre as tonalidades e saturações [27].

### 2.2.3 Espaço de Cor CIELab

De modo a quantificar a cor, em 1976, a *Comission Internationale de l'Eclairage* (CIE) assenta o espaço de cor nas três dimensões do referencial de coordenadas cartesianas  $L$ ,  $a$ ,  $b$  [29]. Em comparação aos restantes espaços de cor, o CIELab foi concebido para se aproximar da percepção humana [30], refletindo um sistema aproximadamente uniforme [31]. Devido a isso, é progressivamente mais utilizado e aceite em todos os tipos de sistemas de representação de cor, já que cada cor em particular tem uma localização única. O eixo vertical  $L$  corresponde à luminosidade, enquanto que os eixos  $a$  e  $b$ , que não apresentam limites, revelam as coordenadas de cromaticidade segundo a teoria da cor oponente de Hering<sup>2</sup>. No diagrama do espaço de cores, o eixo  $a$  indica a componente vermelho-verde de uma cor, enquanto as componentes amarela e azul são representadas no eixo  $b$ . A distância ao eixo central  $L$  representa a saturação da cor [32], como evidenciado na figura 2.3.

<sup>1</sup>A Teoria Tricromática da Cor de Young-Helmholtz postula que a percepção das cores é resultado da ativação de três tipos de cones oculares sensíveis a diferentes comprimentos de onda. Esses cones, associados às cores vermelho, verde e azul, quando estimulados em várias intensidades, permitem distinguir uma vasta variedade de cores [26].

<sup>2</sup>A Teoria do Processo do Oponente de Hering explica a percepção das cores humanas, contrastando com a Teoria Tricromática da Cor de Young-Helmholtz. Esta teoria propõe que a percepção da cor é mediada por pares antagônicos que se inibem. Esses pares criam seis cores independentes em oposição: *branco vs preto*, *vermelho vs verde* e *amarelo vs azul* [32].

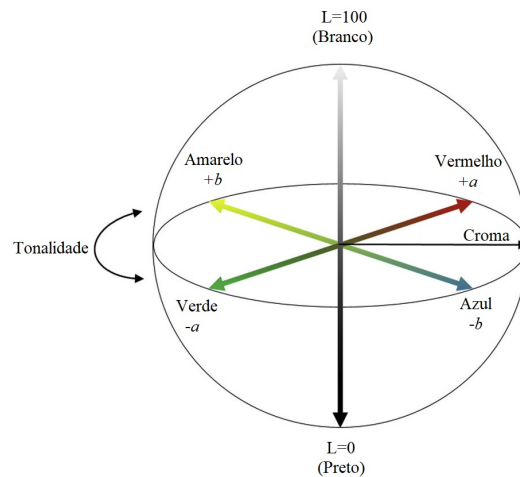


Figura 2.3: Diagrama de espaço de cor CIELab  
 Fonte: Ly et al. [33]

### 2.3 *Machine Learning* na Previsão das Fórmulas de Afição de Tinta

As técnicas de *Machine Learning* têm adquirido cada vez mais reconhecimento em investigações aplicadas, incluindo a área da colorimetria. Estes algoritmos são projetados para detectar automaticamente padrões complexos a partir dos dados de entrada, convertendo experiências passadas em conhecimento [34].

A previsão da receita de afiação de tintas é um desafio crítico nessa indústria, que requer exatidão e eficiência. Como tal, a utilização de modelos de *Machine Learning* tem-se mostrado uma via promissora para a previsão de receita de tinta [34]. Por isso, serão apresentados alguns conceitos fundamentais e analisados diferentes métodos usualmente utilizados nesta área de investigação.

Os algoritmos de *Machine Learning* podem ser divididos, em termos gerais, em duas categorias predominantes: *Supervised Learning* e *Unsupervised Learning* [35].

- *Supervised Learning*: os algoritmos são orientados por dados de treino com informações conhecidas. Por isso mesmo, procuram aprender um modelo capaz de prever resultados, usando o conhecimento prévio disponível [35][36];
- *Unsupervised Learning*: o algoritmo adapta o seu comportamento a partir das observações do meio envolvente, uma vez que não existem objetivos explícitos a serem alcançados. Esses algoritmos são especialmente úteis para identificar padrões e auxiliar na compreensão de conjuntos de dados complexos [34].

Dentro deste contexto complexo de tomada de decisões, que envolve uma ampla gama de fatores e critérios a serem considerados [37], a introdução do paradigma

*Multi-Output Learning* destaca-se como uma ferramenta essencial. Esta técnica permite a previsão simultânea de múltiplas saídas, que enriquece a compreensão das relações existentes e proporciona *insights* para auxiliar a tomada de decisão [38].

O objetivo de prever a proporção de cada base de tinta a misturar para obter uma determinada cor enquadra-se em *Multi-target Regression*, uma vez que requirita inúmeras variáveis de resposta contínua correlacionadas estatisticamente. Por esse motivo, a revisão bibliográfica focaliza-se especificamente nesse subcampo. Os métodos de *Multi-target Regression* permitem modelar efetivamente o conjunto de dados, aliando tanto as relações implícitas entre as variáveis preditoras e os *targets*, bem como as relações entre os próprios *targets* [39].

Seja  $X = R^d$ , um espaço de entrada  $d$ -dimensional de  $n$  instâncias e  $Y = R^m$  um espaço de saída  $m$ -dimensional. *Multi-Output Regression* pretende desenvolver uma função  $f$  a partir do conjunto de formação  $D = (x_1, y_1), \dots, (x_n, y_n)$ , mapeando cada entrada para múltiplas saídas [37] como nas equações 2.1 e 2.2.

$$f : \Omega x_1 \times \dots \times \Omega x_m \rightarrow \Omega y_1 \times \dots \times \Omega y_n \quad (2.1)$$

$$x = (x_1, \dots, x_m) \rightarrow y = (y_1, \dots, y_n) \quad (2.2)$$

sendo que  $\Omega x_i$  e  $\Omega y_k$  referem-se ao espaço de variáveis preditoras  $x_i, \forall i \in \{1, \dots, m\}$ , e *targets*  $y_k, \forall k \in \{1, \dots, n\}$  [40].

Os métodos de *Multi-target Regression* geralmente utilizam duas abordagens: Adaptação de Algoritmos e Transformação de Problemas. A primeira adapta os métodos de *Single-Target* para atender aos algoritmos de *Multi-Target* com um único modelo, explorando as dependências estatísticas dos resultados e aumentando a capacidade de previsão [40]. A segunda opção desenvolve modelos independentes para cada *target* e concatena os resultados [41].

Os métodos de *Machine Learning* referentes a regressão que serão analisados constituem uma seleção de modelos que ganharam visibilidade em ambientes de investigação aplicada, nos últimos anos. Portanto, têm maior intimidade com o objetivo especificado, entre eles, *Linear Regression*, *MultiLayer Perceptron*, RF, entre outros.

Serão apresentados os conceitos técnicos dos modelos integrados, seguidos de uma breve descrição. Em seguida, serão analisadas os métodos adotados em trabalhos relacionados para discutir as relações não lineares propostas nesta dissertação.

### 2.3.1 *Linear Regression*

*Linear Regression* (LR) é um método de *Supervised Learning* que explica a relação entre duas variáveis quando há uma correlação linear entre elas [42], estimando os coeficientes a partir do método dos mínimos quadrados [43]. É amplamente reconhecida e compreendida no campo das estatísticas e de *Machine Learning* [44].

Adicionalmente, apresenta um excelente desempenho comparativamente a outros métodos de aprendizagem estatística, graças à sua eficiência computacional com complexidade  $O(n)$ <sup>3</sup>. Essa eficiência faz da LR uma escolha frequente quando a qualidade das previsões é comparável a métodos mais complexos. O modelo é descrito na equação 2.3:

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_mx_m + \varepsilon \quad (2.3)$$

onde  $\hat{y}$  denota a estimativa da variável dependente,  $x_i, \forall i \in \{1, \dots, m\}$ , representa o valor da  $i$ -ésima variável independente,  $\beta_i, \forall i \in \{1, \dots, m\}$ , corresponde ao coeficiente de regressão respectivo e  $\varepsilon$  indica o termo de perturbação, a parte não explicada de  $y$  [45].

### 2.3.2 *MultiLayer Perceptron*

*MultiLayer Perceptron* (MLP) é um tipo de *Artificial Neural Network* (ANN) que reflete a organização do cérebro humano [46]. Caracterizado por um fluxo unidirecional, os dados atravessam o MLP desde a camada de entrada, passando pelas camadas ocultas até à camada de saída [47]. Este modelo é parte integrante do *Supervised Learning* [35].

Os MLP são projetados para aproximar qualquer função contínua [48] e, assim, implementar qualquer relação não linear entre os dados de entrada e os dados de saída [49]. Essa capacidade de aproximação é resultado da combinação de vários *perceptrons*, os neurónios fundamentais dos MLP [50], organizados em, pelo menos, três camadas totalmente conectadas [51], conforme a figura 2.4:

- Uma camada de entrada, que distribui os dados iniciais pela primeira camada oculta;
- Uma ou mais camadas ocultas de *perceptrons*. As camadas ocultas recebem como entrada a saída do *perceptron* da camada adjacente, sendo este o verdadeiro motor computacional do MLP [49];
- Uma camada de saída de *perceptrons*, que, igualmente, recebe como entrada a saída de cada *perceptron* da última camada oculta [50].

Conforme mostrado na figura 2.5, um *perceptron* recebe  $n$  dados de entrada  $x = x_1, \dots, x_n$ , cada um associado a um peso correspondente  $w = w_1, \dots, w_n$ . A soma ponderada dos dados é então direcionada a uma função de ativação, a qual gera a saída do *perceptron*. O processo de aprendizagem de Retro-propagação envolve a constante atualização dos pesos dos *perceptrons*, visando minimizar o erro [50]. Esse

---

<sup>3</sup>Complexidade  $O(n)$  denota que o tempo de execução do algoritmo é linearmente proporcional ao tamanho da entrada.

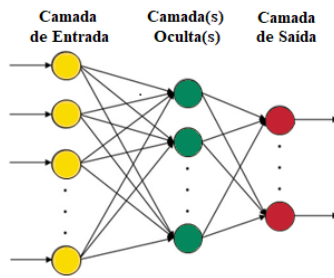


Figura 2.4: Representação de MLP

Fonte: Adaptado de Abiodun et al.

[52]

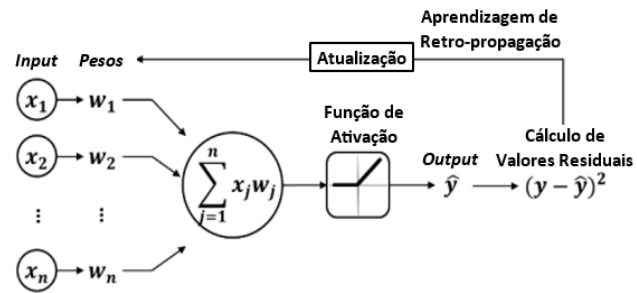


Figura 2.5: Esquema de um *Perceptron* com  $n$  dados de entrada

Fonte: Adaptado de Baek & Choi [53]

procedimento é repetido de maneira contínua até que o valor do erro de previsão seja inferior ao limiar estabelecido [46].

### 2.3.3 *Autoencoder-based Transfer Learning*

*Autoencoder-based Transfer Learning* é considerada uma variante que combina dois componentes principais: uma estrutura de *Autoencoder* fundamentada na ideia de *Transfer Learning*<sup>4</sup> e um modelo de *Machine Learning* que tem como responsabilidade mapear os dados de entrada para os dados de saída desse novo contexto [55].

O *Autoencoder*, por sua vez, é uma arquitetura de ANN que opera em duas fases distintas: *Encoder* e *Decoder*. Durante a fase de *Encoder*, os dados de entrada são geralmente transformados numa representação de dimensão inferior, enquanto que o *Decoder* funciona como o reflexo do *Encoder*, onde os dados originais são reconstruídos a partir da representação codificada, como evidenciado na figura 2.6.

Ao minimizar o erro de reconstrução entre os dados de entrada e os dados de saída, os *Autoencoders* aprendem padrões significativos intrínsecos aos dados [48].

Nesse contexto, ao incorporar o conceito de *Transfer Learning*, a informação latente obtida pelo *Autoencoder* é integrada ao conjunto de dados e usada como entrada para um outro modelo de *Machine Learning*. Nesse método, o modelo é empregado posteriormente com o propósito de correlacionar os dados de entrada aos dados de saída [57]. Isso concede ao novo modelo a capacidade de prever e generalizar com base no conhecimento extraído pelo *Autoencoder* e, portanto, a lidar eficazmente com o novo conjunto de dados.

Esta abordagem possibilita a adaptação resultante do modelo para lidar com o novo conjunto de dados, capitalizando o conhecimento previamente adquirido. Assim, segundo Tan et al. [55], ao combinar o poder do *Autoencoder* e de outro

<sup>4</sup>*Transfer Learning* é uma técnica que utiliza o conhecimento adquirido na resolução de um problema (domínio de origem) e aplica-o a um novo problema (domínio de destino), melhorando assim a precisão da previsão [54][55].

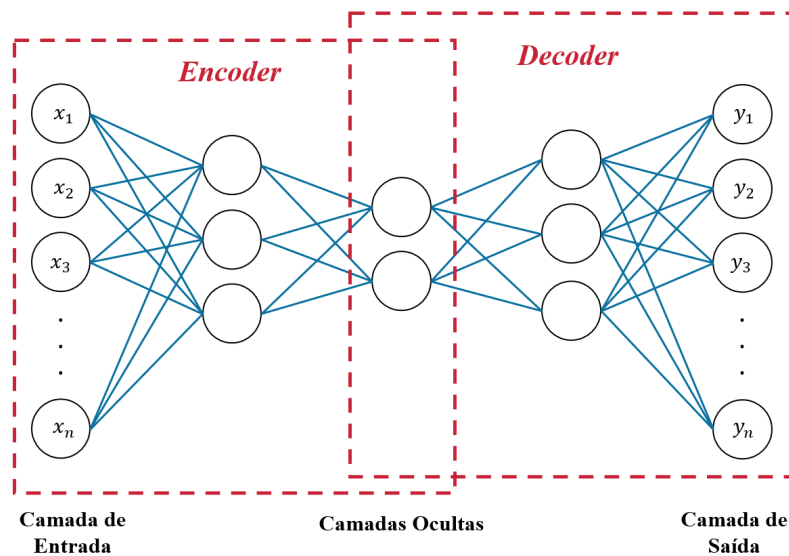


Figura 2.6: Esquema de um *Autoencoder*

Fonte: Adaptado de Hung et al. [56]

modelo, essa estratégia visa melhorar a eficácia das previsões em contextos onde as relações complexas entre os dados requerem uma representação latente para uma melhor generalização.

### 2.3.4 *Random Forest*

*Random Forest* (RF) representa uma outra classe de algoritmos não lineares de *Supervised Learning* amplamente utilizada, que combina várias *Decision Tree* (DT) para melhorar o desempenho preditivo [58]. Neste modelo, segue-se a técnica *Bagging* [59], que consiste em construir cada DT a partir de uma amostra independente de *bootstrap* dos dados<sup>5</sup>, permitindo a compreensão de diferentes padrões exclusivos presentes nos dados [60]. Cada nó da árvore é dividido pelo melhor entre um subconjunto aleatório de variáveis preditoras, aumentando a diversidade e robustez do modelo [61]. A previsão final para uma nova amostra é obtida ao calcular a média das previsões de todas as DT individuais [58], como evidenciado na figura 2.7.

As principais vantagens das RF referem-se à sua flexibilidade e resistência ao *overfitting*<sup>6</sup> sobre uma DT única, tornando-se menos sensíveis aos valores de entrada [64]. As RF são notáveis pela facilidade de utilização e capacidade de alcançar alta precisão e discriminação com apenas dois parâmetros-chave: o tamanho do subconjunto aleatório de variáveis e o número de DT que abrange [61]. Graças a

<sup>5</sup>*Bootstrap* é uma técnica de seleção aleatória de atributos, com reposição, para criar múltiplas amostras a partir de um conjunto de dados [60].

<sup>6</sup>*Overfitting* acontece quando um modelo se ajusta demais aos dados de treino, absorvendo até mesmo o ruído. Isso leva a um desempenho excelente nos dados de treino e a uma queda no desempenho quando aplicado a novos dados não vistos, o conjunto de teste, onde a sua capacidade de generalização é limitada [63] [64].

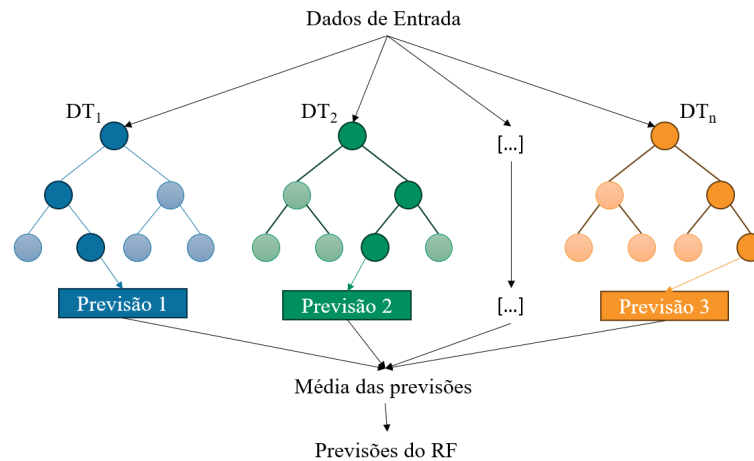


Figura 2.7: Esquema de um RF  
 Fonte: Adaptado de Segura et al. [62]

isso, são uma alternativa poderosa aos métodos tradicionais, destacando-se pela sua robustez ao ruído [64].

### 2.3.5 *Extreme Gradient Boosting*

O XGBoost, abreviação para *Extreme Gradient Boosting*, é um método avançado de *Supervised Learning* que tem conquistado destaque numa variedade de aplicações. Trata-se de um algoritmo de *ensemble* baseado em DT, projetado para otimizar o desempenho preditivo e a eficiência computacional [65].

Uma das principais inovações do XGBoost reside na otimização tanto do *bias* quanto da variância<sup>7</sup> dos modelos, procurando um equilíbrio entre ambos, a partir de uma técnica de *Boosting* [67]. Nesta técnica, várias DT são criadas sequencialmente, cada uma delas corrigindo os erros da anterior [68], a partir do ajuste automático dos pesos das observações durante o treino.

Uma característica distintiva e inovadora reside na estratégia eficiente de diminuição do gradiente para otimizar a função de perda [69]. Por outro lado, o XGBoost também introduz um termo de regularização que controla a complexidade do modelo e previne o *overfitting* [70]. Essas abordagens combinadas contribuem para a generalização e precisão, resultando num desempenho robusto e confiável em diversos cenários e aplicações.

<sup>7</sup> *Bias* e variância são dois aspetos cruciais na avaliação de modelos. O bias reflete erros sistemáticos devido a suposições simplificadas, enquanto a variância mostra sensibilidade a variações nos dados. Reduzir o *bias* do modelo envolve torná-lo mais complexo e capaz de capturar os padrões nos dados. Por outro lado, reduzir a variância envolve tornar o modelo mais robusto às variações nos dados e evitar o *overfitting* [66].

### 2.3.6 *Clustering* e Aplicação dos Modelos aos *Clusters*

Sendo uma técnica de *Data Mining*, o principal objetivo do *Clustering* consiste em agrupar os pontos em categorias, automática e homoganeamente, por proximidades e grau de semelhança. No entanto, os limites são muitas vezes desvirtuados, como se pode verificar pela figura 2.8 [71].

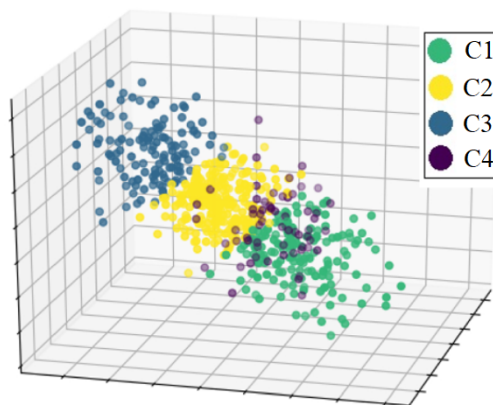


Figura 2.8: *Clustering* em 4 grupos no espaço

Fonte: Jiao et al. [72]

A divisão dos dados em agrupamentos permite aprimorar o desempenho e a precisão dos modelos, representando uma estratégia que ganhou relevância em aplicações de análise de dados. Segundo Du et al. [73], ao dividir os dados que compartilham características semelhantes, é possível personalizar os modelos para se adequarem a diferentes subconjuntos de dados. Este procedimento visa melhorar a adaptação dos modelos aos padrões exclusivos presentes em cada grupo, resultando em previsões mais precisas e contextualmente relevantes.

O *Self-Organizing Maps* consiste num algoritmo avançado de *Clustering* que utiliza técnicas de ANN para realizar transformações não lineares em dados multidimensionais. Sendo um algoritmo de *Unsupervised Learning* [35], tais transformações permitem analisar propriedades intrínsecas no conjunto de dados original [74]. Nagar et al. [75] organizam os dados de forma a destacar padrões, tornando-os uma ferramenta poderosa para visualização, análise exploratória e identificação de *insights* em conjuntos de dados complexos.

### 2.3.7 *Ensemble Learning*

*Ensemble Learning* envolve a adaptação de dois ou mais modelos aos mesmos dados, seguida pela combinação das previsões de cada modelo [76]. Essa combinação tem como objetivo alcançar um desempenho superior, proporcionando previsões mais precisas e robustas do que as obtidas por um único método analítico isolado [77].

A *Ensemble Learning* é um método útil para melhorar a previsibilidade e reduzir a variância de algoritmos estocásticos [76]. Deste modo, os modelos-base colaboram de forma sinérgica, com o objetivo de aprimorar conjuntamente o desempenho e contribuir para a robustez do modelo a fim de enfrentar a complexidade inerente aos dados, como procura demonstrar a figura 2.9.

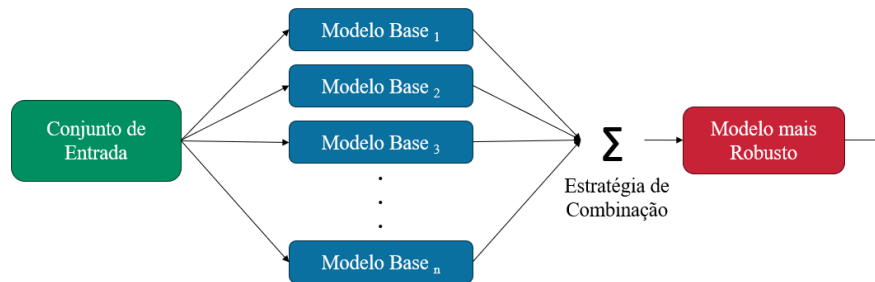


Figura 2.9: *Ensemble* de Modelos

Fonte: Adaptado de Chen et al. [78]

Do ponto de vista de *Machine Learning*, as considerações levantadas por Dietterich [79] apontam algumas vantagens desta técnica. Estas considerações incluem: (i) a limitação dos dados de treino em fornecer informações suficientes para selecionar um único algoritmo como ótima solução e (ii) o desafio de lidar com problemas complexos, procurando um desempenho excelente em todos os *targets*, pode estar além das capacidades de qualquer algoritmo individual. Por conseguinte, esta técnica emerge como uma alternativa atrativa, uma vez que fornece um meio de contornar as limitações inerentes à representação no espaço de hipóteses, possibilitando a exploração de soluções mais resilientes e eficazes, conforme mencionado por Wang et al. [80].

## 2.4 Trabalhos Relacionados

A previsão das concentrações de bases de tintas é um processo intrinsecamente não linear e de grande importância para diversas indústrias, uma vez que proporções incorretas podem afetar a satisfação com o produto final. Por isso mesmo, a precisão neste procedimento é crucial para garantir que as tonalidades sejam alcançadas de modo consistente. Esta seção explora estudos relevantes com o propósito de compreender métodos e técnicas eficazes para uma previsão de qualidade neste domínio.

Apesar da limitada disponibilidade de literatura nesta área específica, é notável que as abordagens existentes se concentram predominantemente em duas vertentes: a aplicação de metaheurísticas [81] e o uso de técnicas de Inteligência Artificial, com especial incidência em ANN [82].

Em relação às metaheurísticas, o estudo conduzido por Chaouch et al. [81] destaca-se ao empregar o *Ant Colony Optimization Algorithm* para avaliar as concentrações previstas. A análise dos resultados revelou um elevado grau de correspondência entre as concentrações previstas e as tonalidades desejadas, com a maioria dos desvios de cor observados abaixo do limite de 0.7 unidades. Da mesma forma, os resultados do *Genetic Algorithm* para previsão de receita de cor para produção de fios tingidos demonstraram uma boa correspondência colorimétrica entre as cores-alvo, com diferenças de cor médias de 0.69 unidades [83]. Esses resultados ressaltam a eficácia intrínseca de ambas as metaheurísticas, uma vez que uma reprodução de cor bem-sucedida requer desvios de cor inferiores a 2 unidades [84], segundo os critérios especificados em ambos os estudos.

No domínio das técnicas de Inteligência Artificial, os primeiros avanços na interseção da colorimetria com *Machine Learning* foram pioneiramente explorados por Bishop, Usher et al. [82], que destacaram a viabilidade das ANN na indústria da cor, representando um passo significativo. No mesmo sentido, vários estudos subsequentes demonstraram importantes resultados na previsão de cores por meio desse método. Os estudos de Huang et al. [85] concentraram-se na previsão de pigmentos para aguarela, onde a rede foi treinada para minimizar os resíduos na previsão. Mei-Yun Chen et al. [86], por outro lado, desenvolveram uma Paleta Inteligente usando uma *Deep Neural Network* (DNN) para a previsão de receitas de mistura de pigmentos. Ambos os estudos mostraram que aproximadamente 85% dos dados tinham diferenças de cor inferiores a 5 unidades nas pontuações de teste, o que está abaixo do limite para ser determinado como uma incompatibilidade [85] [86]. Isto indica uma interessante congruência entre as previsões geradas pelas ANN e as cores reais obtidas. O algoritmo de Golob et al. [87], também baseado em ANN, foi considerado eficaz e rápido em termos computacionais para a determinação da concentração de corantes, apesar de exigir um esforço de formação e parametrização maior. Portanto, esses resultados ressaltaram a notável habilidade do modelo em identificar padrões nas interações entre os pigmentos.

A procura por métodos avançados de previsão de cores também se estendeu à comparação com modelos matemáticos bem-estabelecidos, como o sistema de previsão de *Kubelka-Munk* (K-M)<sup>8</sup>. Westland [4] conduziu um estudo abrangente que comparou o desempenho das ANN com esse modelo matemático no domínio da colorimetria. Os resultados revelaram que o *Machine Learning* superou o modelo teórico, na medida em que ANN obtiveram um desvio de cor de cerca de 1 unidade, enquanto o modelo K-M apresentou um desvio de cor de 1.5 unidades. Estudos adicionais de

---

<sup>8</sup>O sistema K-M é uma técnica matemática que descreve a absorção e reflexão da luz por materiais coloridos, contribuindo para determinar a quantidade necessária de um pigmento específico para alcançar as cores desejadas em produtos finais [88]. Vale ressaltar que a maioria dos *softwares* atualmente disponíveis no mercado, destinados a objetivos semelhantes aos desta dissertação, fundamenta-se nesta teoria [89].

Militký [89] corroboraram essas conclusões, ressaltando também a simplificação em relação à complexidade matemática do modelo teórico.

A comparação direta dos resultados deste estudo com trabalhos relacionados é inviável devido a diferenças significativas nos conjuntos de dados utilizados e aos distintos problemas abordados. Os conjuntos de dados dos estudos mencionados envolvem diferentes tonalidades de bases de tinta, o que torna os modelos desenvolvidos incompatíveis entre si. Além disso, os problemas abordados também variam consideravelmente, dependendo do contexto específico da sua aplicação, como a tinta usada em grandes superfícies [24], a tinta para tingir tecidos [83], a tinta para impressão ou estampados. As diferenças também se estendem às características das bases de tinta, pigmentos e corantes para aguarela. Também a confidencialidade dos processos de afinação das tintas por parte das empresas dificulta o fornecimento de dados específicos e relevantes para a área de estudo. Outra diferença reside na natureza da comparação dos resultados. Enquanto os estudos anteriores centram-se nas diferenças entre as cores obtidas e a referência, esta pesquisa centra-se na análise das proporções fornecidas na fórmula da tinta.

Na secção de trabalhos relacionados, a pesquisa expande o seu âmbito além dos estudos estritamente relacionados com a colorimetria, abrangendo uma vasta gama de trabalhos que exploram a aplicação de técnicas de *Machine Learning* na previsão de proporções de elementos em diferentes contextos. Esta abordagem tem como propósito identificar conhecimentos e métodos que possam ser adaptados e aplicados ao contexto específico deste estudo. Desta forma, a revisão da literatura inclui pesquisas relacionadas com a previsão de composições e misturas em diversos domínios, tais como química, indústria alimentar e outras áreas onde a precisão desempenha um papel fundamental.

Saxena & Parasher [90] desenvolveram um estudo que visa prever a mistura ideal de ração para animais, de forma a manter um equilíbrio entre nutrientes, água e preço, garantindo a saúde dos animais e controlando os custos. Neste estudo, utilizou-se ANN para treinar modelos matemáticos e prever as proporções ideais de ingredientes na ração, a partir das propriedades dos alimentos. Os resultados demonstram a superioridade deste modelo em relação a modelos já existentes, com um Coeficiente de Correlação ( $R^2$ ) de 0,9998, refletindo uma alta precisão nas previsões. Por conseguinte, foram desenvolvidas também ANN mais complexas, com 5 camadas ocultas, que mostraram melhorias significativas, reduzindo erros e aumentando o fator de correlação.

Numa outra área de estudo, a indústria alimentar, onde a precisão na medição dos componentes é essencial para a qualidade do produto final, Zhou et al. [91] apresentaram o desenvolvimento de um sistema projetado para avaliação da proporção de alimentos em pó na mistura. No contexto da *Multi-target Regression*, este sistema utiliza *Convolutional Neural Network* (CNN) para processar os espectros dos

alimentos em pó. Os resultados alcançados revelaram uma melhoria significativa em relação ao modelo de referência linear, com redução do Raiz do Erro Quadrático Médio (RMSE) e aumento do  $R^2$ .

Na área da geologia também se encontrou estudos análogos. Chang et al. [92] utilizam técnicas de *Machine Learning*, com foco em ANN, para prever e otimizar a composição de ligas de alta entropia, com níveis excepcionais de dureza. Esta abordagem também integrou o algoritmo de *Simulated Annealing*, uma metaheurística reconhecida, para aprimorar a composição das ligas e o método de *bootstrap*, que permite quantificar a incerteza dos resultados. Esta pesquisa demonstra com sucesso que é possível obter novas composições para as ligas em questão, com dureza superior aos valores alcançados pelos modelos empíricos existentes.

De acordo com a análise de Bannigan et al. [93], o uso de modelos de *Machine Learning*, incluindo DT, ANN e DNN, tem-se destacado no desenvolvimento de formulações de medicamentos, uma área de extrema importância na indústria farmacêutica. Essa integração apresenta como benefício a capacidade de otimizar o processo em termos de eficiência, levando a uma redução no tempo e nos recursos necessários para desenvolver fármacos de alta qualidade. Assim, a aplicação dessas técnicas possibilita a exploração de novos componentes e formulações inovadoras que podem melhorar a eficácia dos medicamentos.

Na Indústria de Petróleo e Gás, de acordo com as conclusões de Adeniyi et al. [94], o emprego de ANN e de *AdaBoost* para modelar a composição do gás, resultante da reforma a vapor do bio-óleo de biomassa, revelou-se de grande relevância nos setores de produção de energia e química. Ambos os algoritmos alcançaram bons resultados, com melhor desempenho do ANN em alguns *targets* e do *AdaBoost* noutros, capturando, no total, 99.9% da variabilidade dos dados.

Já Yang et al. [95] abordam a previsão da proporção de gás-óleo, a partir de um modelo de *Machine Learning* treinado com dados de reservatórios bem estabelecidos. Utilizaram-se, entre outros, RF e uma variante específica da LR, a *Linear Regression Regularized with Elastic Net*. Os resultados demonstram uma concordância significativa entre as previsões e as medições reais em amostras correspondentes. Isso fornece informações valiosas para a tomada de decisão e a compreensão da distribuição de fluidos no reservatório.

Os estudos aqui abordados, embora não diretamente vinculados à colorimetria, enriquecem significativamente a presente pesquisa ao oferecerem perspectivas diferentes. Os contributos transcendentais aqui presentes podem ser proveitosamente aplicados para ampliar a compreensão acerca da previsão das proporções de elementos em contextos diversos. Esta ampla visão promove uma abordagem verdadeiramente interdisciplinar, fortalecendo a capacidade de aprimorar a precisão das estimativas no âmbito da colorimetria.



## Capítulo 3

# Metodologia

Esta dissertação segue a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) [96]. Esse modelo compreende seis etapas para planejar e executar um projeto de *Data Mining*, como evidenciado na figura 3.1: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* e *Deployment*. A escolha desta metodologia para este projeto é altamente pertinente, pois alinha-se perfeitamente com os objetivos específicos desta pesquisa. O CRISP-DM é um processo iterativo que permite acomodar a descoberta e atualização às necessidades da investigação, dada a complexidade inerente no contexto deste projeto.

Durante a fase de *Business Understanding*, as necessidades específicas da indústria são identificadas. Os dados relevantes para o projeto são analisados na etapa de *Data Understanding*. Posteriormente, os dados são preparados para a *Modeling*, onde diferentes algoritmos de *Machine Learning* são explorados. Na fase de *Evaluation*, o modelo que melhor se adequa aos objetivos é identificado. Por fim, o modelo final é implementado para facilitar o acesso aos resultados obtidos. Neste estudo, a fase de *Deployment* não foi realizada, uma vez que o foco principal foi o desenvolvimento e a avaliação do modelo. Decisões de implementação serão tomadas após otimização e validação por um especialista.

Para conduzir este projeto, optou-se pela utilização da plataforma *Konstanz Information Miner Analytics* (KNIME), uma vez que é uma ferramenta versátil que permite implementar diversos algoritmos e prever decisões com base em registos históricos. Além disso, também se diferencia por partilhar uma *interface* gráfica

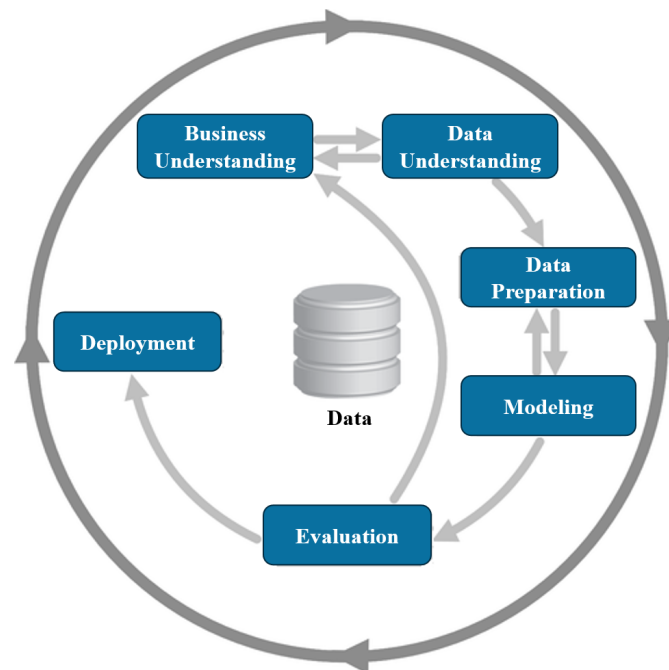


Figura 3.1: Metodologia CRISP-DM

Fonte: Huber et al. [97]

interativa que possibilita experimentar e avaliar intuitivamente o desempenho de vários algoritmos.

Dado o cenário complexo da previsão de dados, o KNIME é um *software* livre para integração de ferramentas de *Machine Learning* com a linguagem de código aberto [98], que pode ser usado para trabalhar com parâmetros dos sistemas de cor [99].

### 3.1 *Business Understanding*

Sendo a Gloma® uma empresa de pintura de acessórios, é frequente receber encomendas que solicitam cores personalizadas. Os clientes fornecem amostras de cores específicas, esperando que a empresa reproduza a cor correspondente. Nos últimos anos, a Gloma® iniciou um processo de digitalização deste fluxo de trabalho, começando pela aquisição de um espectrofotômetro de elevada precisão. Este dispositivo é agora utilizado para verificar a conformidade das cores com base nas especificações dos clientes. No entanto, o processo de mistura de tintas não sofreu alterações para integrar plenamente este equipamento. Embora eficaz, o processo definido requer o envolvimento de colaboradores experientes.

Dado este cenário, o processo de afinação de tintas revela um elevado potencial para evolução tecnológica. Com a crescente procura por cores personalizadas que requerem a participação de profissionais qualificados, torna-se importante procurar soluções mais eficientes e sustentáveis.

O intuito desta dissertação passa por desenvolver um modelo de *Machine Learning* que apoie os técnicos na afinação de tintas. O processo envolve a análise das amostras por meio do espectrofotômetro para obter os parâmetros da cor que são posteriormente utilizados pelos técnicos para determinar a proporção precisa de cada uma das bases de tinta disponíveis que deve ser misturada para alcançar a cor desejada.

A implementação de um modelo de *Machine Learning* visa melhorar significativamente a eficiência da afinação de tintas, reduzindo a dependência de colaboradores altamente qualificados e aumentando a capacidade de atender à crescente procura por cores personalizadas. Esta iniciativa representa um passo relevante na melhoria dos processos, resultando em benefícios palpáveis tanto para a empresa como para os seus clientes.

### 3.2 *Data Understanding*

Considerando a colaboração da empresa, que autorizou o compartilhamento dos dados essenciais para a condução deste projeto, foi fornecido um conjunto de 3639 fórmulas de tinta. Assim, por meio do espectrofotômetro, foram obtidos e acrescentados os parâmetros  $L$ ,  $a$  e  $b$  do sistema CIELab da cor correspondentes a cada uma dessas fórmulas<sup>1</sup>. Durante um período de 5 meses, a colaboração ativa dos membros da equipa resultou na recolha adicional de 6 367 observações. Desta forma, o conjunto de dados original foi ampliado para 10 006 observações reunidas, estabelecendo-se a base para a concretização deste projeto, cujo objetivo principal é a criação de um modelo de *Machine Learning* capaz de calcular as proporções necessárias de tintas a serem misturadas, a fim de alcançar uma cor específica.

Este estudo empregou um conjunto de dados composto por quatro variáveis independentes, nomeadamente as componentes  $L$ ,  $a$ ,  $b$  e *Acabamento*, em conjunto com dezasseis variáveis dependentes. Essas últimas abrangem uma diversidade de tonalidades vibrantes, tais como *Amarelo Vivo*, *Laranja Claro*, *Ocre*, *Amarelo Alaranjado*, *Laranja Escuro*, *Vermelho Carmim*, *Vermelho*, *Bordeaux*, *Castanho*, *Azul*, *Verde*, *Verde Claro*, *Branco* e *Preto*, que são as bases de tinta a misturar, além das componentes *Extender Mate* e *Extender Brilho*. Por razões económicas, a ação do *extender* não influencia a tonalidade da mistura, mas atua como um agente de volume para a tinta.

Todas as variáveis foram tratadas de forma quantitativa, com exceção da variável *Acabamento*, que apresenta caráter qualitativo, acrescentando assim uma dimensão de complexidade à análise. Vale destacar que cada observação neste conjunto de

---

<sup>1</sup>Todas as leituras no espectrofotômetro foram realizadas sob as mesmas condições e com a estabilidade do instrumento, a fim de garantir resultados consistentes e minimizar influências externas. Isso é crucial para assegurar a validade e precisão das análises das características de cor e dos modelos de previsão.

dados representa uma cor única, proporcionando uma contribuição singular para o estudo. A tabela 3.1 ilustra um exemplo de uma observação, onde o acabamento e os parâmetros de cor CIELab são considerados como variáveis preditoras, enquanto as bases de tinta são tratadas como variáveis *target*.

Tabela 3.1: Exemplo de uma observação

Variáveis		Fórmula de Cor
Variáveis Preditoras	Acabamento	Brilho
	$L$	11.33
	$a$	93.91
	$b$	77.72
Variáveis <i>Target</i>	Mate	0
	Brilho	21.98
	Amarelo Vivo	0
	Laranja Claro	7.18
	Ocre	0
	Amarelo Alaranjado	57.78
	Laranja Escuro	0
	Vermelho Carmim	0
	Vermelho	0
	Bordeaux	0
	Castanho	0
	Azul	0
	Verde	0
	Verde Claro	0
	Branco	13
Preto	0	

Relativamente à sua composição, as tintas utilizadas são à base de solvente de poliuretano, escolhidas devido à natureza dos materiais a serem revestidos, que incluem plástico, metal, madeira, entre outros [100], diferenciando-se dos tipos de tinta usados em estudos relacionados. Como se pode constatar pela quantidade de variáveis dependentes, estas tintas exibem uma vasta gama de pigmentações. Essa diversidade de pigmentos é necessária para atender à crescente procura por cores personalizadas, sendo uma característica intrínseca do processo de afinação que o torna mais complexo e desafiante.

Na figura 3.2, é possível visualizar o conjunto de dados total e as respetivas cores em cada ponto do gráfico. Dessa forma, emerge claramente um esquema, em três dimensões, semelhante ao espaço de cor CIELab, ilustrado na figura 2.3. Para uma visualização completa dos dados, foram fornecidas duas perspetivas complementares, as subfiguras da esquerda e da direita, que capturam diferentes aspetos do mesmo gráfico.

Pelo gráfico obtido, também se consegue aferir sobre a distribuição dos dados. A tabela 3.2 apresenta o número de observações por cada tonalidade expressa no *Munsell Ordering System*, da figura 2.2 da secção 2.2. É evidente que a distribuição não é uniforme, com determinados segmentos de cor contendo um número significativamente maior de observações em comparação com outros, como o caso do *Amarelo*, que possui aproximadamente cinco vezes mais observações do que o *Púrpura*. Embora a preferência seja por uma distribuição equitativa, a realidade do

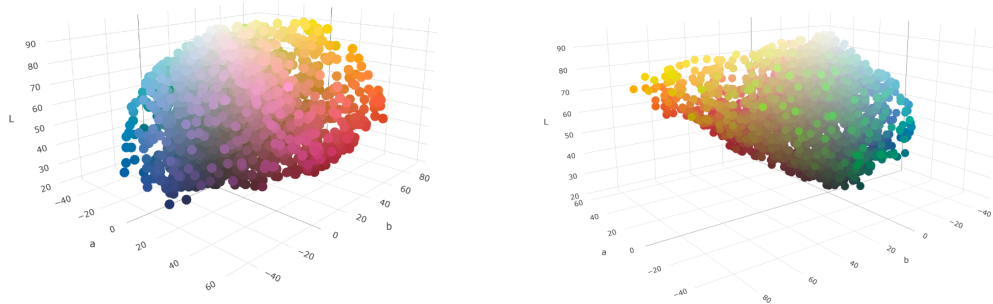


Figura 3.2: Gráfico do Conjunto de Dados

conjunto de dados pode impor algumas limitações. No entanto, a consciencialização sobre a distribuição dos dados auxilia na interpretação cuidadosa dos resultados, levando em consideração a representatividade limitada de certos grupos.

Tabela 3.2: Contagem de Observações por Tonalidade

Tonalidades	Contagem das Observações
Vermelho	1 446
Amarelo-Vermelho (YR)	1 444
Amarelo (Y)	1 994
Verde-Amarelo (GY)	754
Verde (G)	657
Azul-Verde (BG)	591
Azul (B)	851
Púrpura-Azul (PB)	1 327
Púrpura (P)	414
Vermelho-Púrpura (RP)	528

A partir de uma análise exploratória dos dados, é possível constatar que as categorias “Brilho” e “Mate” da variável *Acabamento* possuem uma relação intrínseca com as características de *Extender Brilho* e *Extender Mate*, sendo mutuamente exclusivas. Por isso mesmo, no conjunto de dados fornecidos, observa-se que 4 638 registos correspondem ao acabamento “Mate”, enquanto as 5 368 restantes são associadas ao acabamento “Brilho”. Para o acabamento ativo, a proporção do *extender* varia entre 10% a 40%. Além disso, ambos os *extenders* exibem valores centrados na média<sup>2</sup>, indicando uma distribuição com menos valores extremos, já que a *kurtosis* é negativa e próxima de zero<sup>3</sup>. Quanto à *skewness*, indica simetria com leve inclinação para valores menores<sup>4</sup>.

Quanto às variáveis *a* e *b*, cujos valores de *skewness* e *kurtosis* são positivos<sup>5</sup>, torna-se evidente que a maioria dos valores são da menor escala, com apenas algumas ocorrências de valores elevados. Já a variável *L* apresenta valores de *skewness* e

<sup>2</sup>A média para o Extender Brilho é de 11.94, enquanto o Extender Mate apresenta uma média de 12.00.

<sup>3</sup>Para a kurtosis, o Extender Brilho regista -0.48, ao passo que o Extender Mate apresenta -0.96.

<sup>4</sup>Os valores de *skewness* para *Extender Brilho* é 0.72 e para *Extender Mate* é 0.67.

<sup>5</sup>A variável *a* apresenta *skewness* de 0.97 e 1.46 de *kurtosis*. A variável *b* apresenta de *skewness* de 0.97 e 2.53 de *kurtosis*.

de *kurtosis* negativos e próximos de zero<sup>6</sup>, indicando que a distribuição dos dados nessa variável é relativamente equilibrada e uniforme, concentrando-se menos nos extremos.

No apêndice A, são apresentados os dados estatísticos das variáveis em estudo, fornecendo detalhes adicionais para uma análise mais aprofundada.

Entre as bases mais requisitadas, *Branco* encontra-se presente em aproximadamente 95.6% das observações, seguido por *Preto* com 75.5%, *Ocre* com 64.3% e *Vermelho* com 61.7%. Em contrapartida, existem bases com uma taxa de utilização consideravelmente baixa, como o *Verde Claro*, com apenas 0.95% de utilização, o *Amarelo Alaranjado*, com 3.61%, e o *Amarelo Vivo*, com 3.98%. A notar que as bases mais frequentes exibem uma variabilidade mais pronunciada, refletida em valores menores tanto para o *skewness* quanto para a *kurtosis*. Essa característica revela a tendência dessas bases de tinta a uma distribuição mais uniforme, menos assimétrica e menos concentrada comparativamente às restantes bases. As figuras 3.3 e 3.4 representam histogramas de bases pouco e muito frequentes, respectivamente, com o intuito de evidenciar as discrepâncias mencionadas.

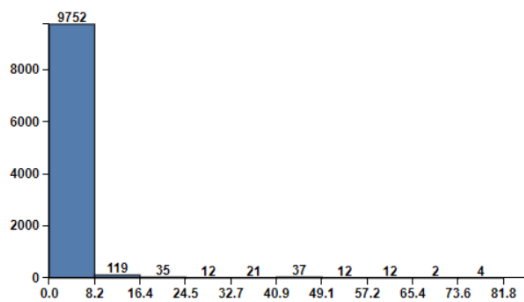


Figura 3.3: Histograma da Base *Amarelo Vivo*

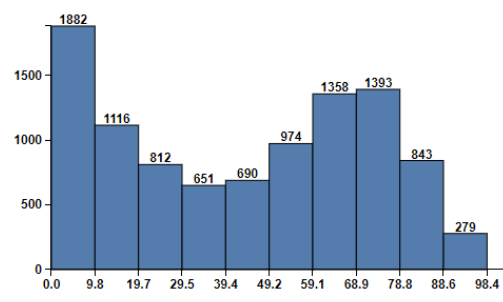


Figura 3.4: Histograma da Base *Branco*

A análise da matriz de correlação na figura 3.5 revela relações significativas entre as variáveis do conjunto de dados. Uma das observações significativas, como supracitado, é a forte relação entre o atributo *Acabamento* e o tipo de *extender* usado, seja *Extender Mate* (-93.9%) ou *Extender Brilho* (91.0%), indicando uma influência marcante desses fatores nas características das amostras. A correlação negativa entre as variáveis *Extender Mate* e *Extender Brilho* (-85.5%) sugere que essas duas opções são mutuamente exclusivas.

Observa-se que valores mais altos na coordenada *a* está associado a uma proporção maior de cores quentes, como *Laranja Claro* (42.7%), *Vermelho* (48.6%) e *Bordeaux* (26.0%), enquanto valores menores estão ligados a cores mais frias, como *Verde* (-54.7%), *Azul* (-39.3%), *Preto* (-29.2%), *Branco* (-27.5%). O comportamento semelhante é observado na coordenada *b*, onde valores mais elevados estão relacionados a cores com tons de *Laranja Claro* (45.2%), *Ocre* (32.6%), *Amarelo Alaranjado*

<sup>6</sup>A variável *L* apresenta *skewness* de -0.069 e *kurtosis* de -1.283.

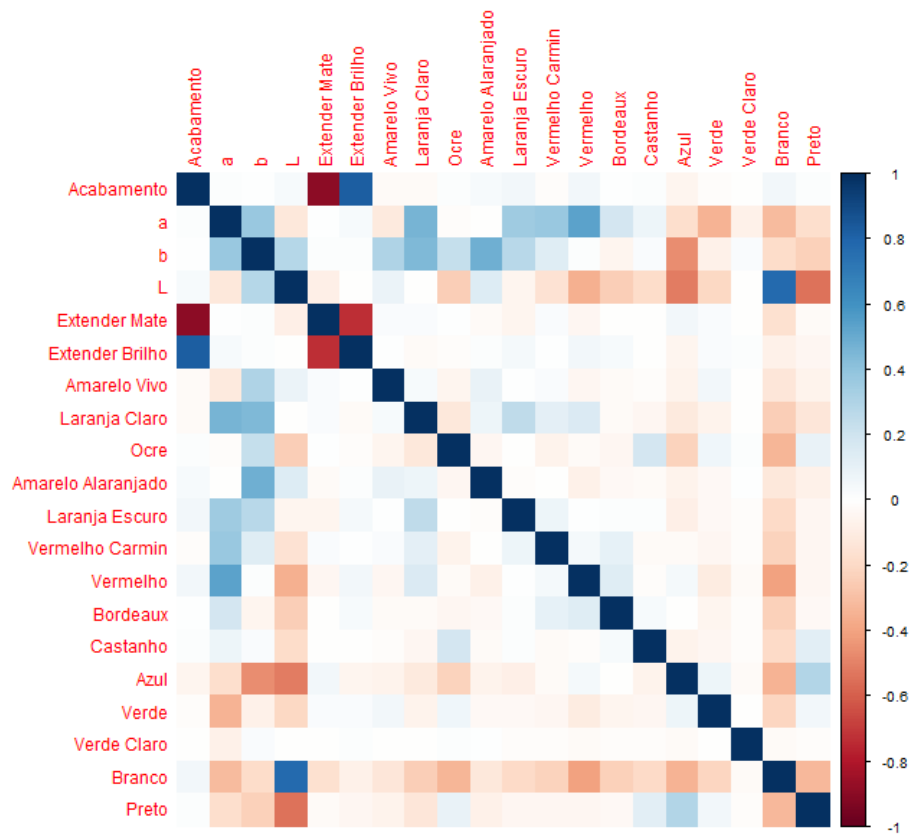


Figura 3.5: Matriz de Correlação

(30.8%) e *Laranja Escuro* (29.2%), enquanto valores menores estão também ligados a cores com tons de *Azul* (-75.6%) e *Preto* (-38.0%).

A coordenada *L* revela uma correlação positiva com cores claras, como *Branco* (76.5%) e uma correlação negativa com cores escuras, como *Preto* (-57.4%), *Vermelho* (-25.3%) e *Azul* (-38.2%). Esses resultados sugerem que a luminosidade está relacionada à intensidade da cor.

Além disso, é importante destacar algumas relações inversas entre as cores. Entre eles, é observada uma correlação inversa entre a base *Laranja Claro* e as bases *Ocre* (-29.1%), *Azul* (-30.1%) e *Preto* (-42.3%). Da mesma forma, *Ocre* apresenta uma relação inversa com as bases *Azul* (-41.2%) e *Preto* (-32.0%), *Vermelho* e *Verde* (-36.9%) e entre *Preto* e *Azul* (-27.1%).

No que diz respeito ao *Acabamento*, não se identificou nenhuma influência dessa variável na seleção das bases de tinta ou mesmo na proporção em que são combinadas. Então, do mesmo modo, também não possui correlação significativa com as coordenadas do espaço de cor. Apesar disso, destaca-se as correlações positivas entre a variável *b* com as variáveis *a* (45.2%) e *L* (28.9%). Isso pode ser justificado pela distribuição de cores nas regiões intermediárias do espaço de cor, já que existem segmentos de cor com mais observações que outros.

### 3.3 Data Preparation

Nesta fase, a base de dados passou por tratamentos específicos para garantir a completude e consistência dos dados. Primeiramente, os valores ausentes foram removidos, assegurando dados confiáveis e de alta qualidade.

No processo de afinação de cores, é frequentemente necessário ajustar manualmente a cor obtida pela adição cuidadosa de diferentes bases de tinta para alcançar a tonalidade desejada. Este procedimento, no entanto, afeta a proporção *Extender Mate* ou *Extender Brilho*, já que à medida que se adiciona mais base de tinta, a proporção do *extender* diminui gradualmente e, por isso, estas variáveis apresentavam um desvio padrão elevado (na ordem das 10 unidades). Uma vez que o *extender* age principalmente como um agente de volume e não influencia diretamente a cor, é possível ajustar os valores das fórmulas desde que a proporção das bases de tinta seja mantida constante. Assim, para padronizar as fórmulas de cor, fixou-se o valor do *extender* para 35% e os restantes 65% foram distribuídos pelas bases de tinta de forma proporcional, como se pode comprovar pelo exemplo ilustrado na tabela 3.3.

Tabela 3.3: Processo de Padronização das Fórmulas de Tinta

Variáveis	Fórmula Original	Procedimento	Fórmula Padronizada
Acabamento	Brilho	Fixação do <i>extender</i> para 35% e distribuição dos restantes 65% pelas bases de tinta	Brilho
a	11.33		11.33
b	93.91		93.91
L	77.72		77.72
Mate	0		0
Brilho	21.98		35
Amarelo Vivo	0		0
Laranja Claro	7.18		5.98
Ocre	0		0
Amarelo Alaranjado	57.83		48.19
Laranja Escuro	0		0
Vermelho Carmim	0		0
Vermelho	0		0
Bordeaux	0		0
Castanho	0		0
Azul	0		0
Verde	0		0
Verde Claro	0		0
Branco	13		10.83
Preto	0		0

De forma a facilitar a interpretação dos dados, a variável *Acabamento* foi também convertida para um formato numérico. Posteriormente, os dados foram normalizados usando o *Min-Max Scaler*, para garantir que todos os valores das variáveis estejam dentro da faixa de 0 a 1. Esse procedimento facilita a comparação das variáveis, garantindo que todas contribuam igualmente nas análises subsequentes, conforme recomendado por Mirza et al. [96].

### 3.4 *Modeling*

A fase subsequente do estudo reside na concepção de um modelo cujo propósito é determinar as proporções ideais das dezasseis bases de tinta disponíveis. Efetivamente, este desafio é acentuado pela discrepância significativa entre as quatro variáveis de entrada e o número substancialmente maior de saídas. Como tal, para identificar o algoritmo mais eficaz para prever fórmulas de tinta, foram exploradas três distintas abordagens:

1. Modelação Clássica baseada em Algoritmos Únicos;
2. Combinação dos Melhores Modelos para cada Base de Tinta;
3. Modelação baseada em *Clusters* do Espaço de Cor CIELab.

Para tal intento, uma variedade de algoritmos de *Machine Learning* comumente empregados foi desenvolvida e testada. Desta forma, serão apresentados os algoritmos adotados e suas respectivas configurações. Os resultados finais serão sintetizados para uma posterior análise no capítulo 4.

***Linear Regression:*** LR foi utilizado porque é um bom algoritmo para obter uma previsão de referência, já que é um dos algoritmos de regressão mais simples de *Machine Learning* [42].

***MultiLayer Perceptron:*** MLP foi escolhido devido ao uso comum para resolver problemas semelhantes, como demonstrado na secção 2.4, a partir de dados não-lineares. Para atender aos requisitos neste contexto, optou-se por configurar duas camadas ocultas, compostas por 8 e 12 nós, respetivamente [49].

***Artificial Neural Network with a pre-trained Decoder:*** O modelo em questão, o qual se nomeou *Artificial Neural Network with a pre-trained Decoder* (ANND), é o resultado da fusão entre o *Autoencoder-based Transfer Learning* e o MLP. Essa abordagem foi selecionada com base na comprovada competência do *Transfer Learning* em melhorar a capacidade preditiva e de generalização de um modelo, enquanto o MLP foi escolhido como o modelo base devido à sua ampla utilização em pesquisas relacionadas [49].

Neste estudo, o *Autoencoder* opera de forma diferente e especial. Inicialmente, os valores de entrada do *Autoencoder* consistem em 16 variáveis, representando as bases de tinta. Essas variáveis são reduzidas até 4 pelo *Encoder*, que procura se aproximar das variáveis de entrada originais. Em seguida, o *Decoder* visa reconstruir de volta as 16 variáveis iniciais. A figura 3.6 ilustra a arquitetura do *Autoencoder* de forma visual, compreendendo 5 camadas ocultas, compostas por 12, 8, 4 e depois novamente 8 e 12 nós para reconstrução dos dados.

Após essa etapa, o *Decoder* é extraído e incorporado ao conjunto de dados original, uma vez que reflete a essência do modelo, ou seja, partir das características

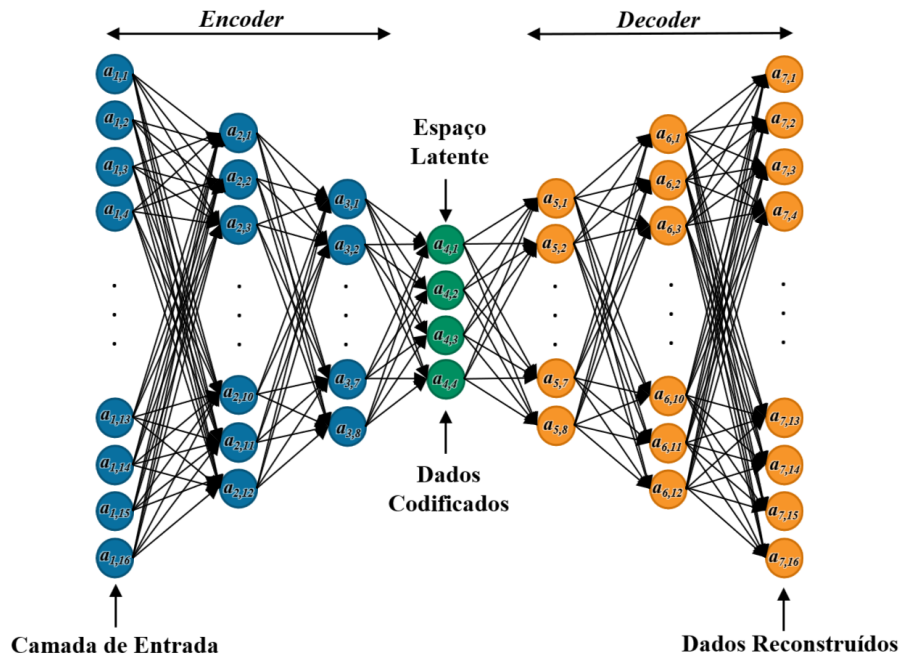


Figura 3.6: Arquitetura do *Autoencoder*

de entrada (atributos da cor) e alcançar os dados de saída (bases de tinta). Deste modo, é adotado o conceito de *Transfer Learning*[57], onde a informação latente obtida pelo *Decoder* é integrada à base de dados e utilizada como entrada para um MLP, com as mesmas configurações mencionadas anteriormente.

**Random Forest:** O algoritmo RF foi escolhido como parte deste estudo pela sua estratégia diferenciada de resolução de problemas, que difere significativamente das abordagens já referidas. A sua natureza baseada em DT é capaz de captar interações complexas entre os atributos, sendo uma mais-valia ao lidar com problemas onde relações não são facilmente modeladas por equações lineares ou ANN [61].

**Extreme Gradient Boosting:** XGBoost foi escolhido devido ao equilíbrio entre precisão e eficiência que proporciona, tornando-o adequado para conjuntos de dados complexos como o presente. Em comparação com o RF, o XGBoost tende a ter uma *performance* mais rápida e consome menos recursos [65].

No âmbito da *Multi-Target Regression*, o LR, RF e XGBoost optaram pela estratégia de Transformação de Problemas, que envolve o desenvolvimento de modelos independentes para cada *target*, com os resultados posteriormente concatenados. Em contrapartida, o MLP e o ANND adotaram a estratégia de Adaptação de Algoritmos, onde um único modelo é treinado para atender a todos os *targets*.

Foi conduzido um estudo de sensibilidade para avaliar parâmetros dos algoritmos em análise e, portanto, na tabela 3.4, é apresentada a gama de valores explorada para cada parâmetro. Em seguida, foram executados os algoritmos considerando

Tabela 3.4: Parametrização de cada Modelo

Modelo	Parâmetros	Intervalo de Pesquisa	Configuração de Parâmetros
MLP e Autoencoder	<i>Epochs</i>	1000-10000	10000
	<i>Training Batch Size</i>	100-1000	1000
	<i>Activation Function</i>	<i>Sigmoid, Linear, Tahn, ReLU, SeLU, Softplus</i>	<i>ReLU</i>
	<i>Optimizer</i>	<i>Adadelta, Adagrad, Adam, Adamax, Nadam, RMSProp</i>	<i>Adam</i>
	<i>Learning Rate</i>	0,001-0,5	0,4
	<i>Beta 1</i>	0.1-0.9	0.3
	<i>Beta 2</i>	0.1-0.9	0.2
	<i>Epsilon</i>	0.1-0.4	0.2
RF	<i>Learning Rate Decay</i>	0.0-0.1	0,0
	<i>Max_Depth</i>	5-30	30
XGBoost	<i>Num_of_Estimator</i>	100-600	600
	<i>Boosting Rounds</i>	100-1000	100
	<i>Base Score</i>	0.0-1.0	0,5
	<i>Eta</i>	0.0-1.0	0,2
	<i>Lambda</i>	0.0-1.0	0.1
	<i>Alpha</i>	0.0-1.0	0.1
	<i>Gamma</i>	0.0-1.0	0.1

diversas configurações de parâmetros, com o objetivo de identificar a combinação com o desempenho mais eficaz.

Claro está que o desempenho dos modelos é aprimorado à medida que se aumenta o valor de alguns parâmetros, como o número de *Epochs* no MLP e no *Autoencoder* e a *Max\_Depth* no RF. No entanto, essa melhoria progressiva veio acompanhada de um aumento no esforço computacional. Para contornar esse problema, foram estabelecidos limites, de forma a procurar um equilíbrio entre a qualidade da solução e a viabilidade computacional. Nas três abordagens adotadas, é importante ressaltar que os parâmetros dos modelos permanecem constantes para fins de comparação posteriormente.

Para avaliar o desempenho dos modelos, foi integrado *k-fold Cross-Validation* (CV), dividindo os dados de treino em  $k - 1$  subconjuntos e um subconjunto para teste. A técnica CV tem como objetivo estimar a validade do modelo e a capacidade de aprendizagem. Neste estudo, o parâmetro  $k$  foi definido como 10, encontrando um equilíbrio entre eficiência computacional e uma estimativa precisa de desempenho [101]. Adicionalmente, não se realizou a divisão em conjunto de validação, devido à limitada quantidade de dados disponíveis para treino dos modelos. Introduzir um conjunto de validação reduziria ainda mais o tamanho do conjunto de treino, tornando a comparação injusta, especialmente para modelos que requisitam um volume maior de dados para um bom desempenho.

### 3.4.1 Combinação de Modelos

Na segunda abordagem, a estratégia concentra-se na formação de um *ensemble* de modelos, que procura otimizar a precisão e eficácia das previsões para cada um dos

*targets* no conjunto de dados, já que diferentes algoritmos se destacam em distintas bases de tinta.

A premissa subjacente a essa abordagem é que a combinação de modelos bem-sucedidos para cada *target*, escolhidos criteriosamente com base no seu desempenho no conjunto de treino, tem o potencial de resultar num *ensemble* poderoso e altamente adaptável, que visa aprimorar a qualidade das previsões para cada uma dessas variáveis [102].

Essa estratégia de modelação enfatiza a adaptabilidade e personalização, garantindo que cada modelo seja selecionado com base na sua eficácia comprovada no conjunto de treino. O resultado é um *ensemble* que reúne o conhecimento especializado de diferentes modelos, com o intuito de melhorar a capacidade de prever com precisão cada *target* específico.

### 3.4.2 Divisão dos dados em *Clusters*

Na terceira abordagem, a segmentação do conjunto de dados possibilita o desenvolvimento de modelos exclusivos para aprimorar a precisão e a contextualização das previsões em cada agrupamento, ao considerar as características e relações específicas aos dados dentro de cada grupo.

Desta forma, no contexto da aplicação do *Self Organizing Map* para divisão dos dados, foram selecionados os parâmetros essenciais para o processo. O parâmetro *convergeEpochs* foi fixado em 1000, enquanto *height* recebeu uma atribuição de 2 e o *LearningRate* foi ajustado para 0.4. Adicionalmente, *orderingEpochs* foi definido com um valor de 2000, enquanto o parâmetro *width* foi especificado como 2. Esta escolha foi motivada pela simplificação da representação dos dados, visando identificar padrões macroscópicos sem prejudicar a quantidade de dados disponíveis em cada grupo, em vez de se concentrar em nuances muito específicas.

Os dados foram subdivididos, automaticamente, em quatro subconjuntos, como se pode constatar pela figura 3.7, alcançando uma abordagem equilibrada. Também evidencia que a distribuição dos dados não se encontra estritamente associada à proximidade no gráfico. No entanto, ao analisar a tabela 3.5, confirma-se que as diferenças mais pronunciadas entre os grupos residem, principalmente, no tipo de acabamento e nos valores da coordenada *L*. Quanto ao *acabamento*, destaca-se que os *Clusters* A e C possuem um acabamento mate, em contraste com os *Clusters* B e D, nos quais o acabamento é brilho. Uma análise mais detalhada da coordenada *L* revela que os *Clusters* C e D tendem a apresentar valores mais baixos de luminosidade, indicando cores mais escuras, enquanto os *Clusters* A e B se caracterizam por valores mais elevados, denotando cores mais claras.

Embora as variações nas coordenadas *a* e *b* não sejam tão marcantes, é possível identificar pequenas divergências entre os agrupamentos. Relativamente às especificidades, no *Cluster* D, a mediana da coordenada *a* destaca-se por ser mais elevada.

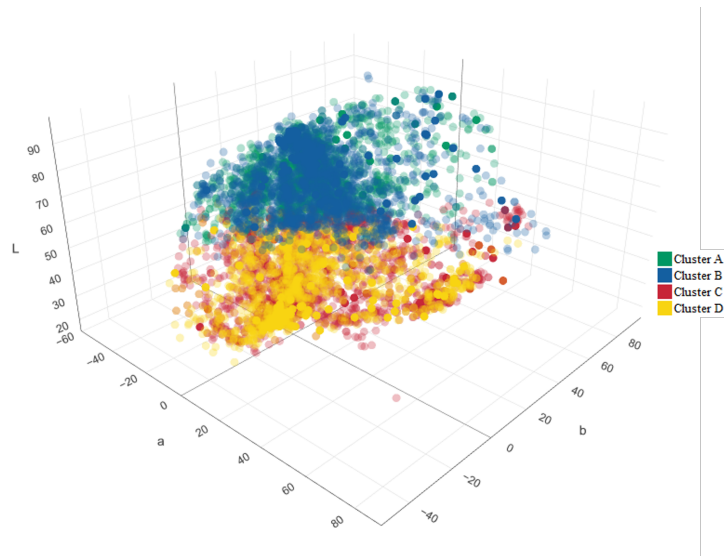


Figura 3.7: Conjunto de dados dividido em quatro subgrupos

Já no *Cluster C*, a mediana da coordenada *b* é inferior em comparação com os demais. Adicionalmente, nos últimos subgrupos, emerge uma perceptível discrepância entre as médias e medianas registadas. Este fenómeno aponta para a presença de uma dispersão expressiva e assimétrica nos dados, onde algumas observações exibem valores atípicos substancialmente superiores.

Tabela 3.5: Características dos *Clusters*

	<i>Cluster A</i>			<i>Cluster B</i>			<i>Cluster C</i>			<i>Cluster D</i>		
	L	a	b	L	a	b	L	a	b	L	a	b
Mínimo	53.1	-51.8	-34.4	53.9	-56.0	-33.8	20.0	-48.4	-47.8	17.5	-47.6	-44.2
Máximo	96.4	85.8	91.5	97.2	58.1	93.9	60.1	61.1	54.4	66.0	60.4	63.9
Média	75.1	4.0	13.1	77.0	2.2	11.5	37.4	9.3	2.4	40.6	12.2	5.0
Mediana	75.3	1.0	8.7	77.4	0.9	7.9	36.5	1.2	0.7	39.6	2.8	3.3
Acabamento	Mate			Brilho			Mate			Brilho		

Além disso, é relevante destacar a distribuição dos dados nos segmentos: o *Cluster A* abrange 2 977 observações (30%), o *Cluster B* engloba 2 525 observações (25%), o *Cluster C* contém 2 391 observações (24%), e o *Cluster D* compreende 2 113 observações (21%).

No âmbito deste estudo, a divisão dos dados em quatro grupos oferece então uma perspetiva intrigante e uma base para avaliar como a segmentação dos dados pode afetar os modelos de previsão aplicados. Com isto, pretende-se compreender a influência do agrupamento de dados e analisar diferenças e melhorias na capacidade de previsão.

### 3.5 Evaluation

A avaliação de modelos de regressão envolve a determinação do erro ao contrastar valores conhecidos com valores estimados. Além de ter em consideração o esforço computacional, a precisão dos modelos de regressão foi avaliada usando duas métricas estatísticas comuns no conjunto de teste, o Erro Absoluto Médio (MAE) e a Raiz do Erro Quadrático Médio (RMSE). Ambas as métricas foram escolhidas pelo facto de serem amplamente aceites e empregues para avaliar a maioria dos estudos relacionados [73] [79] [81] [83] [103] [104] [91] [94].

Ao avaliar modelos de regressão, o MAE e o RMSE fornecem *insights* valiosos. O MAE expressa a média do erro total do modelo e avalia o quão próxima a previsão está dos valores reais. Por sua vez, o RMSE quantifica a discrepância global entre os valores reais e as previsões. A minimização destas métricas é essencial para otimizar os resultados e aprimorar a qualidade das previsões. A aplicação destas métricas estatísticas permite a escolha do modelo de regressão mais apropriado para obter resultados mais precisos [104]. As equações dos parâmetros estatísticos são mostradas nas equações 3.1 e 3.2.

$$MAE = \frac{\sum_{k=0}^{n-1} |y_k - \hat{y}_k|}{n} \quad (3.1)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{n}} \quad (3.2)$$

onde  $n$  representa o número de observações,  $\hat{y}_k$  é a  $k$ -ésima observação e  $y_k$  é a  $k$ -ésima previsão.

É importante notar que o MAE mede o erro médio absoluto entre as previsões do modelo e os valores reais. Por outro lado, o RMSE considera o erro quadrático médio, atribuindo maior peso aos erros maiores. Esta diferença pode explicar por que algumas métricas RMSE são relativamente mais altas do que as correspondentes métricas MAE, indicando que os modelos podem ter cometido previsões substancialmente desviadas em alguns casos.

## Capítulo 4

# Resultados e Discussão

A procura pela formulação ideal de mistura de tintas tem-se revelado um desafio constante na indústria, onde a eficácia das previsões desempenha um papel fundamental na obtenção de cores desejadas. No presente capítulo, leva-se a cabo a análise e avaliação dos resultados obtidos através da aplicação de diferentes abordagens. O objetivo principal é identificar a estratégia mais eficaz que permite prever com maior precisão a composição necessária para atingir uma cor específica.

Ao longo deste estudo, foram exploradas três abordagens distintas, cada uma representando uma perspectiva única na tarefa de previsão de fórmulas de tintas:

1. Modelação Clássica baseada em Algoritmos Únicos: Foram aplicados modelos diretamente aos conjuntos de dados, através da adaptação de cada um deles às características específicas dos dados.
2. Combinação dos Melhores Modelos para cada Base de Tinta: A segunda abordagem envolve a criação de um *ensemble* de modelos, onde múltiplos modelos individuais colaboram para proporcionar previsões mais robustas e abrangentes. Esta estratégia visa aproveitar a diversidade de modelos para melhorar a precisão geral das previsões.
3. Modelação baseada em *Clusters* do Espaço de Cor CIELab: Procurou-se agrupar os dados em conjuntos semelhantes para, posteriormente, captar a relação entre as variáveis dentro de cada *cluster*, a fim de tornar os modelos mais adequados às características dos dados.

Um ponto importante a destacar consiste no facto de que todos os modelos foram desenvolvidos sob a mesma configuração, no sentido de garantir uma comparação justa entre as diferentes abordagens. A uniformidade nas configurações assegura que a *performance* seja avaliada com base na eficácia intrínseca de cada método aplicado, permitindo uma análise confiável e comparável.

Ao longo deste capítulo, são apresentados os resultados detalhados, realçando os modelos mais promissores em cada cenário.

## 4.1 Modelação Clássica baseada em Algoritmos Únicos

Nesta primeira abordagem, foram aplicados os modelos mencionados anteriormente ao conjunto de dados para avaliar o desempenho na previsão das proporções de tinta. A síntese de resultados apresentada na tabela 4.1 reúne os valores de MAE e de RMSE para cada modelo em relação às diferentes bases de tinta, como referido na secção 3.5.

Tabela 4.1: Métricas de Desempenho dos Modelos no Conjunto de Teste

	Conjunto de teste									
	LR		MLP		ANND		RF		XGBoost	
<i>Target</i>	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.19	1.00	0.03	0.70	0.00	0.03	0.03	0.60	0.03	0.59
Brilho	0.21	1.88	0.03	1.16	0.09	0.13	0.02	0.48	0.02	0.45
Amarelo Vivo	1.37	4.07	2.07	5.28	0.41	2.56	0.35	2.25	0.48	2.47
Laranja Claro	2.56	5.39	1.96	4.97	0.23	0.69	0.74	2.71	0.88	2.86
Ocre	6.77	10.31	7.15	7.85	0.95	1.52	1.86	4.17	2.12	4.35
Amarelo Alaranjado	1.55	4.18	0.82	3.08	0.29	1.81	0.29	1.88	0.37	1.96
Laranja Escuro	1.46	3.87	0.86	3.31	0.43	2.53	0.38	1.75	0.50	1.88
Vermelho Carmim	1.81	4.50	1.55	4.69	0.64	3.70	0.44	1.81	0.57	2.02
Vermelho	4.03	7.29	5.12	6.02	1.42	2.76	1.33	3.43	1.59	3.69
Bordeaux	2.01	5.33	1.92	6.43	0.47	2.27	0.49	2.35	0.59	2.50
Castanho	1.12	3.31	1.39	3.09	0.59	3.44	0.34	1.58	0.43	1.65
Azul	4.78	7.94	3.41	4.75	0.47	1.19	0.83	2.80	0.94	2.83
Verde	2.10	4.64	1.51	5.82	0.61	2.29	0.44	1.87	0.55	2.05
Verde Claro	0.17	1.25	0.12	1.25	0.08	1.26	0.08	1.12	0.09	1.06
Branco	7.20	9.42	6.78	8.37	1.09	2.06	2.89	5.83	3.22	5.94
Preto	2.86	5.27	2.12	4.22	0.46	0.89	0.74	2.37	0.90	2.44
Total	40.18	79.67	36.85	70.96	8.23	29.12	11.28	37.00	13.27	38.74

O ANND, RF e XGBoost emergiram como modelos de realce para múltiplos *targets*, demonstrando consistentemente MAE e RMSE inferiores quando comparados a LR e MLP, em todas as execuções. Isso significa que estes modelos demonstram maior habilidade em captar a intrincada complexidade das relações entre os parâmetros de cor e as concentrações de tinta.

Na última linha da tabela 4.1, é apresentado o desempenho global de todos os modelos, resultante da soma dos valores de MAE e RMSE de cada *target*, respectivamente. Entre os modelos analisados, destaca-se o ANND, que exhibe as métricas de menor valor, como evidenciado na figura 4.1, realçando-se como o modelo mais congruente com os dados em análise. No entanto, é crucial ressaltar que o desempenho pode variar de acordo com o *target* específico.

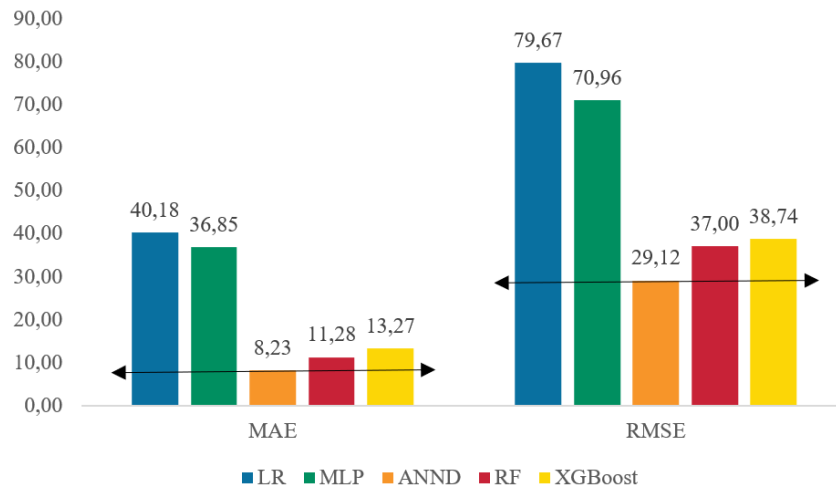


Figura 4.1: Comparação dos resultados do desempenho dos modelos no conjunto de dados completo

Os resultados do ANND demonstram uma notável vantagem em termos de precisão preditiva quando comparados com os demais modelos. As métricas MAE e RMSE apresentaram diferenças substanciais: o ANND exibiu reduções de 80% e 63% em relação ao LR, 78% e 59% em relação ao MLP, 27% e 21% em relação ao RF, e 38% e 25% em relação ao XGBoost, respectivamente.

Sublinha-se, também, a significativa variação no desempenho dos modelos entre diferentes *targets*. Enquanto *targets* como *Extender Mate* e *Extender Brilho* ostentam valores relativamente reduzidos de MAE e RMSE para a maioria dos modelos, já outros *targets*, como *Vermelho* e *Castanho*, revelam desafios mais proeminentes na previsão. Estas variações apontam para a complexidade diversificada dos vínculos entre os parâmetros de cor e as concentrações de tinta em diferentes contextos.

## 4.2 Combinação dos Melhores Modelos para cada Base de Tinta

Dado que nenhum modelo único é o melhor consistentemente em todos os *targets*, como é observado nos resultados apresentados na tabela 4.1, pode ser interessante considerar a formação de um *ensemble* que combina os pontos fortes de vários modelos, já que diferentes modelos são especialistas em *targets* distintos.

Desta forma, a segunda abordagem concentra-se na construção de um *ensemble* dos modelos que tiveram melhor desempenho a prever cada uma das bases de tinta. Os modelos a combinar serão escolhidos com base nos resultados do desempenho no conjunto de treino, como referido na subsecção 3.4.1. Esses dados encontram-se expostos na tabela B.1 que se encontra no apêndice B.1. Ao combinar as qualidades de cada modelo, procura-se alcançar um sistema mais robusto e abrangente.

Tabela 4.2: Escolha dos Modelos para integrarem o *Ensemble*

<i>Target</i>	MAE	Modelo	RMSE	Modelo	Modelo escolhido para <i>ensemble</i>
Mate	0.00	ANND	0.03	ANND	<b>ANND</b>
Brilho	0.00	MLP	0.00	MLP	<b>MLP</b>
Amarelo Vivo	0.13	RF	0.78	RF	<b>RF</b>
Laranja Claro	0.23	ANND	0.71	ANND	<b>ANND</b>
Ocre	0.70	RF	1.54	ANND	<b>RF</b>
Amarelo Alaranjado	0.11	RF	0.72	RF	<b>RF</b>
Laranja Escuro	0.14	RF	0.67	RF	<b>RF</b>
Vermelho Carmim	0.16	RF	0.65	RF	<b>RF</b>
Vermelho	0.51	RF	1.33	RF	<b>RF</b>
Bordeaux	0.19	RF	0.90	RF	<b>RF</b>
Castanho	0.13	RF	0.58	RF	<b>RF</b>
Azul	0.31	RF	1.05	RF	<b>RF</b>
Verde	0.16	RF	0.68	RF	<b>RF</b>
Verde Claro	0.03	RF	0.39	RF	<b>RF</b>
Branco	1.08	RF	2.04	ANND	<b>ANND</b>
Preto	0.28	RF	0.88	ANND	<b>RF</b>

Com base nos resultados obtidos a partir do conjunto de treino apresentado na tabela B.1 e na síntese de dados evidenciada na tabela 4.2, observa-se que o ANND demonstrou um desempenho exemplar em relação às variáveis *Extender Mate* e *Laranja Claro*. Por outro lado, o RF exibiu um desempenho superior nas variáveis *Amarelo Vivo*, *Amarelo Alaranjado*, *Laranja Escuro*, *Vermelho Carmim*, *Vermelho*, *Bordeaux*, *Castanho*, *Azul*, *Verde* e *Verde Claro*, enquanto que o *Extender Brilho* apresentou melhor desempenho no MLP.

Entretanto, notou-se uma falta de concordância nas métricas de desempenho para as variáveis *Ocre*, *Branco* e *Preto*, sublinhadas a cor diferente na tabela. Com o intuito de abordar esta discordância, foi realizada uma comparação detalhada das métricas de desempenho, onde se concluiu que as variáveis *Ocre* e *Preto* seriam melhor modeladas pelo RF. Esta decisão fundamentou-se na observação de que estas variáveis apresentaram valores mais baixos na soma das métricas. Dado que ambas as métricas são igualmente relevantes, a escolha recaiu sobre o modelo que oferecesse um equilíbrio otimizado entre o MAE e RMSE. Seguindo a mesma lógica, a variável *Branco* foi, em contrapartida, associada ao modelo ANND.

Os resultados destacados na tabela 4.3 ressaltam o desempenho do *ensemble* criado, verificando-se que o sistema desenvolvido apresenta MAE de 9.20 unidades e RMSE de 30.87. A análise destes resultados visa proporcionar uma compreensão aprofundada do potencial do *ensemble* de modelos em superar as limitações individuais de cada modelo.

Tabela 4.3: Métricas de Desempenho da Combinação de Modelos no Conjunto de Teste

<i>Target</i>	Conjunto de Teste		
	Modelo	MAE	RMSE
Mate	ANND	0.13	0.24
Brilho	MLP	0.05	0.56
Amarelo Vivo	RF	0.36	2.20
Laranja Claro	ANND	0.39	1.02
Ocre	RF	1.85	4.15
Amarelo Alaranjado	RF	0.29	1.86
Laranja Escuro	RF	0.38	1.70
Vermelho Carmim	RF	0.42	1.63
Vermelho	RF	1.33	3.42
Bordeaux	RF	0.50	2.37
Castanho	RF	0.34	1.56
Azul	RF	0.83	2.79
Verde	RF	0.43	1.83
Verde Claro	RF	0.07	1.04
Branco	ANND	1.10	2.13
Preto	RF	0.73	2.37
Total	<i>Ensemble</i>	9.20	30.87

Mesmo ao combinar as capacidades de diversos modelos, é claro que a previsão da base *Vermelho* continua a representar um desafio, como já observado na abordagem anterior. Por outro lado, notou-se uma melhoria significativa na previsão da proporção da base *Castanho*, enquanto que a variável *Ocre* demonstrou ser mais desafiadora nesta iteração.

### 4.3 Modelação baseada em *Clusters* do Espaço de Cor CIELab

Na terceira abordagem, o conjunto de dados foi segmentado em quatro grupos distintos, como mencionado na subsecção 3.4.2, permitindo desenvolver modelos específicos para cada grupo. Essa estratégia visa aprimorar a capacidade de compreender as características e relações particulares aos dados, melhorando assim a precisão e a contextualização das previsões em cada grupo.

Ao conduzir uma comparação entre os modelos, neste contexto específico, torna-se evidente que o RF, XGBoost e ANND surgem como os modelos mais robustos nesta abordagem, enquanto o MLP e LR não demonstram o mesmo nível de desempenho.

Uma análise dos resultados das métricas na tabela 4.4 revela variações consideráveis nos valores de MAE e RMSE entre os diferentes modelos, quando os dados são divididos. O RF exibe métricas de menor valor, conforme ilustrado na figura 4.2, consolidando-se como o modelo mais consistente com os dados em análise.

As métricas MAE e RMSE registam diferenças substanciais, com o RF a apresentar reduções de 69% e 50% em relação ao LR, 72% e 65% em relação ao MLP, 21% e 29% em relação ao ANND e 18% e 6% em relação ao XGBoost, respetivamente.

Tabela 4.4: Métricas de Desempenho dos Modelos aplicados aos *Clusters* no Conjunto de Teste

<i>Target</i>	Conjunto de teste									
	LR		MLP		ANND		RF		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.02	0.53	4.95	10.76	0.17	1.05	0.02	0.65	0.03	0.67
Brilho	0.01	0.38	18.74	25.61	0.11	0.94	0.02	0.50	0.02	0.53
Amarelo Vivo	1.36	4.06	0.79	4.86	0.69	4.58	0.37	2.30	0.52	2.61
Laranja Claro	2.46	5.29	1.96	6.53	0.92	3.00	0.72	2.61	0.91	2.85
Ocre	6.02	9.24	7.11	12.85	1.00	2.66	1.78	3.98	2.10	4.24
Amarelo Alaranjado	1.33	3.80	0.89	5.33	0.70	4.52	0.30	1.94	0.42	2.14
Laranja Escuro	1.32	3.69	0.94	4.57	0.51	2.60	0.38	1.76	0.51	1.95
Vermelho Carmim	1.62	4.37	0.86	4.71	0.67	3.65	0.43	1.65	0.56	1.95
Vermelho	3.64	6.89	4.27	9.78	1.15	2.51	1.34	3.49	1.62	3.73
Bordeaux	1.83	5.20	0.98	5.58	0.87	3.61	0.52	2.50	0.64	2.51
Castanho	1.05	3.28	0.55	2.85	0.43	2.62	0.36	1.67	0.44	1.71
Azul	3.69	6.83	4.38	11.57	2.31	7.32	0.78	2.63	0.92	2.74
Verde	1.98	4.47	1.37	5.51	1.00	3.96	0.43	1.79	0.54	1.97
Verde Claro	0.17	1.25	0.08	1.33	0.08	1.25	0.08	1.07	0.10	1.07
Branco	5.45	7.61	14.41	17.26	1.83	3.78	2.51	4.74	2.92	4.99
Preto	2.49	4.91	3.04	6.31	1.12	2.11	0.71	2.32	0.92	2.40
Total	34.44	71.77	65.33	135.40	13.56	50.18	10.76	35.60	13.18	38.06

Na análise das previsões para cada base de tinta, destacam-se alguns padrões distintivos. Especificamente, as variáveis *Amarelo Alaranjado* e *Verde Claro* exibem valores relativamente baixos de MAE e RMSE, sugerindo um desempenho mais consistente. Por outro lado, as variáveis *Ocre*, *Vermelho* e *Branco* apresentam-se como desafios consideráveis, como se tem evidenciado também nas outras abordagens, pelos valores elevados de MAE e RMSE registrados nos múltiplos modelos. Isso aponta para a dificuldade que esses modelos enfrentam ao tentar prever com precisão esses *targets*. No entanto, é intrigante destacar o desempenho do modelo ANND, particularmente em relação a esses *targets* mais desafiantes, que evidencia a sua adaptação a cenários complexos e a sua habilidade em capturar nuances nos dados.

Além disso, também se manifesta uma melhoria no desempenho dos RF, XGBoost e LR relativamente à primeira abordagem. Esses ganhos podem ser atribuídos à estratégia de segmentação em *clusters*, a qual demonstra ser interessante nesse contexto. No entanto, quanto aos modelos MLP e ANND, como requisitam uma maior quantidade de dados para serem treinados, a divisão em grupos aparenta afetar adversamente o seu desempenho.

Nesse sentido, é pertinente denotar que, em particular, os valores da métrica RMSE obtidos pelo MLP e ANND são significativamente superiores, com uma diferença de 29.79 e 21.06 unidades, respectivamente, aos valores equivalentes da primeira abordagem. O RMSE, ao enfatizar os erros maiores, sugere que esses modelos podem ter realizado previsões substancialmente divergentes em algumas cores, quando

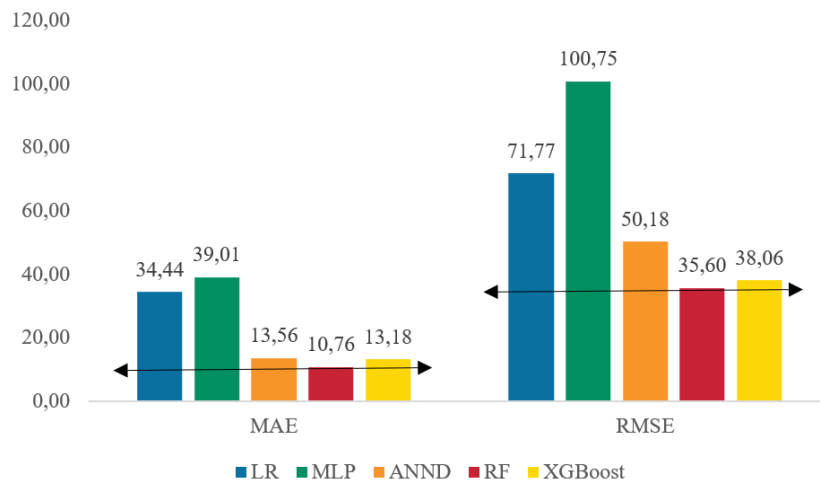


Figura 4.2: Comparação dos resultados do desempenho dos modelos nos *clusters*

aplicados a *clusters*.

## 4.4 Discussão Geral dos Resultados

Em resumo, os resultados desta secção destacam a complexidade das relações entre as técnicas de *Machine Learning* e os resultados obtidos nos conjuntos de treino e de teste. Cada abordagem possui as próprias vantagens e limitações inerentes e, por isso mesmo, compreender e analisar cuidadosamente essas características é fundamental para a seleção informada da estratégia mais adequada em conformidade com os objetivos específicos do estudo.

Ao comparar os resultados das diferentes estratégias, algumas observações evidentes podem ser destacadas. Primeiramente, é evidente que certos *targets*, como *Ocre*, *Vermelho*, *Castanho* e *Branco* apresentaram-se como maiores desafios devido à complexidade dos dados e às relações envolvidas, sendo também bases com maior taxa de utilização e com maior variância, como referido na subsecção 3.2. Quanto às variáveis *Ocre*, *Vermelho* e *Castanho*, a presença de pigmentos fortes e dominantes nessas bases pode causar interações complexas, afetando a previsão. Essas bases, sendo altamente utilizadas, podem ter variações significativas nas proporções de mistura, dependendo da aplicação. Já *Branco* é frequentemente usado para clarear diversas tonalidades, o que as torna especialmente desafiadora de prever devido à sua natureza versátil.

Isso sugere que a complexidade das relações entre os parâmetros de cor e as proporções de tinta pode diferir consideravelmente entre diferentes *targets*, quer seja no conjunto de dados completo, quer seja em cada *cluster*, explicando que os *targets* mais desafiadores apresentam valores mais elevados de MAE e RMSE, mesmo nos modelos mais eficazes.

Além disso, é importante destacar a diferença significativa nos resultados entre o ANND e o MLP (que não utiliza *Transfer Learning*), realçando o valor acrescentado da incorporação do *Decoder* pré-treinado no mapeamento dos valores de entrada. Isso é particularmente intrigante, uma vez que ambas as redes compartilham a mesma arquitetura. Conforme mencionado na secção 2.4, estudos anteriores frequentemente utilizaram ANN para prever concentrações de base de tinta. No entanto, nesta pesquisa e neste conjunto de dados, o MLP não conseguiu captar adequadamente as complexidades das relações entre os parâmetros de cor CIELab e as concentrações de base de tinta.

Este estudo adotou diversos modelos, alguns incorporando a técnica de Transformação de Problemas e outros aplicando a Adaptação de Algoritmos. Contudo, observa-se que, após uma análise cuidadosa, não se conseguiu discernir nenhuma diferença significativa nos resultados entre essas abordagens, demonstrando-se ambas igualmente eficazes.

Por outro lado, a análise de cada uma das três abordagens desenvolvidas abrange as peculiaridades de cada estratégia, considerando a consistência e a precisão das previsões.

Na primeira abordagem, o ANND destaca-se ao apresentar os valores mais baixos de MAE e RMSE em comparação com outros modelos. Essa consistência reforça a adequação aos dados analisados, com MAE de 8.23 e RMSE de 29.12 unidades.

Na segunda abordagem, os resultados enfatizam o desempenho do *ensemble* desenvolvido, avaliado com MAE de 9.20 e um RMSE de 30.87 unidades. Esta perspectiva foi concebida para capitalizar as virtudes do RF, do ANND e do MLP. Na formação deste sistema, os modelos para cada *target* foram selecionados com base no desempenho no conjunto de treino. No entanto, comparando os resultados do conjunto de treino da tabela B.1 com os resultados do conjunto de teste da tabela 4.1, são evidenciadas discrepâncias que sugerem a presença de *overfitting*, especialmente no caso do modelo RF. Esta dinâmica explica a razão pela qual, na primeira abordagem, o ANND se destacou amplamente, enquanto que, na segunda, apenas foi selecionado por três *targets*.

Quanto à terceira abordagem, o modelo RF apresentou valores reduzidos de MAE e RMSE em relação a outros modelos, quando o conjunto de dados está dividido. Neste caso, o modelo apresentou MAE de 10.76 e RMSE de 35.60 unidades.

Assim, em relação à seleção da melhor estratégia, é crucial levar em consideração a adaptabilidade dos modelos aos dados e a coerência das suas previsões ao selecionar a abordagem mais vantajosa. Entre as diferentes estratégias avaliadas, o ANND da primeira abordagem sobressai-se, como evidenciado na figura 4.3, sustentado pela notável habilidade de generalização e adaptação a *targets* desafiantes, sendo por isso a resposta à questão de investigação definida. Em contraste, os demais modelos e estratégias, embora possuam méritos próprios, podem não se ter ajustado

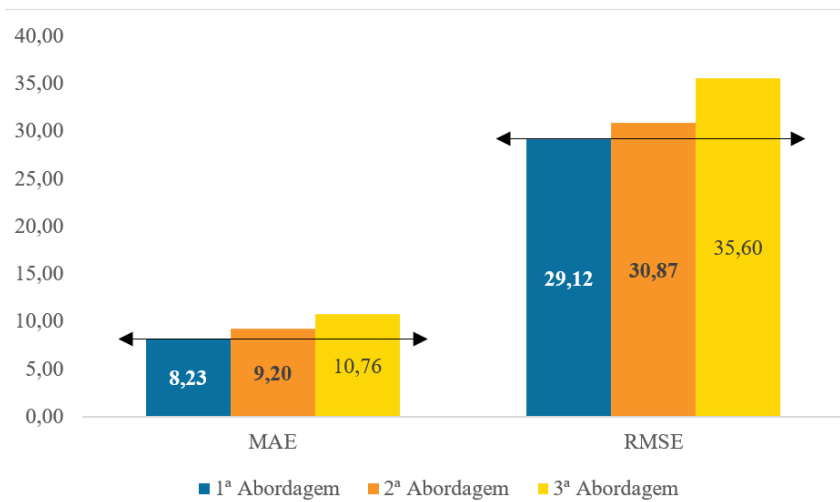


Figura 4.3: Comparação dos resultados do desempenho das diferentes abordagens

adequadamente às complexidades inerentes aos atributos específicos ou às relações não lineares entre os dados.

Ao longo de todo o projeto e das técnicas desenvolvidas, o ANND manteve o seu destaque. Além da primeira abordagem, o ANND foi um dos modelos selecionados para integrar o *ensemble* na segunda abordagem. Na terceira abordagem, onde o modelo RF teve o melhor desempenho global, o ANND ainda conseguiu se sobressair ao demonstrar-se eficaz e altamente adaptável na previsão de atributos mais desafiantes nos diferentes conjuntos.

Já relativamente ao LR, o modelo assume uma relação linear entre os dados de entrada e os dados de saída. Tendo em conta que os dados deste problema exibem relações não lineares, o LR enfrenta dificuldades para se adaptar a esses padrões complexos. Contudo, mesmo com essa limitação, o LR foi considerado um modelo de referência, fornecendo uma base para a comparação com outros modelos mais avançados.

No caso do MLP, este modelo frequentemente requer uma quantidade substancial de dados para aprender de maneira eficaz. Dado o tamanho do conjunto de dados de treino, os resultados podem não ter alcançado o nível de desempenho desejado. Além disso, o facto de, na terceira abordagem, ainda ter dividido o conjunto de dados, pode ter sido um fator iníquo. Este ponto também se aplica ao ANND, já que também integra um MLP, ou seja, aumentando o conjunto de dados, iria potencializar astronomicamente o desempenho de ambos os modelos, levando a resultados mais robustos e precisos.

Em relação ao RF e ao XGBoost, considerando que os dados são desbalanceados, como abordado na secção 3.2, estes modelos apresentam a desvantagem de oferecer maior ênfase aos segmentos de cor maioritários, o que prejudica a previsão dos segmentos de cor minoritários.

No entanto, é importante destacar que os resultados obtidos neste estudo são surpreendentes, pois contradizem, a um nível técnico, as expectativas convencionais. Tanto a segunda quanto a terceira abordagens, baseadas em métodos de modelação mais específicos, geralmente são associadas a um desempenho superior comparativamente aos modelos de base, de acordo com a literatura. Apesar disso, estes resultados demonstram que a abordagem clássica é capaz de alcançar previsões de fórmulas de tinta mais precisas, reforçando sua relevância dentro deste contexto. Já ao nível do negócio, estes resultados representam o ponto de partida para uma transformação significativa na automação deste processo, com o propósito de proporcionar ao afinador uma ferramenta objetiva e eficaz.

## 4.5 Comparação da Modelação entre *Clusters* na Afinação de Tintas

Recordando que a distribuição dos dados no espaço CIELab é não uniforme e que foi identificada uma disparidade na quantidade de exemplos entre os grupos, conforme detalhado na secção 3.2, optou-se por realizar uma análise individual da modelação de cada segmento de dados. Desta forma, pretende-se compreender se o desempenho global dos modelos é afetado por características específicas em grupos particulares, promovendo uma exploração detalhada de cada segmento.

Tendo em consideração os *clusters* introduzidos na terceira abordagem, foram atribuídos modelos específicos a cada grupo com as próprias particularidades de cor e composição de tinta. A análise comparativa destes resultados fornecerá *insights* sobre a capacidade de adaptação dos modelos às características únicas de cada grupo. O apêndice B.2 compreende a análise aprofundada de cada *cluster* específico, incluindo a avaliação da adequação dos modelos em relação aos dados e aos respetivos *targets*, além de destacar detalhes relevantes com potencial impacto na análise.

Em forma de resumo, a tabela 4.5 constitui uma representação concisa e informativa dos resultados das métricas de desempenho global dos modelos em análise. Essa tabela condensa a avaliação abrangente dos modelos em relação às métricas MAE e RMSE, fornecendo uma visão geral do panorama das *performances* individuais, que se encontram disponíveis no apêndice B.2, nas tabelas B.2, B.3, B.4 e B.5.

Tabela 4.5: Resumo das Métricas de Desempenho dos Modelos no Conjunto de Teste dos *Clusters*

<i>Clusters</i>	Conjunto de teste									
	LR		MLP		RF		ANND		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
A	16.77	37.03	15.04	36.06	<b>6.12</b>	<b>21.04</b>	7.79	36.39	7.35	22.12
B	15.25	32.25	15.27	30.99	<b>6.09</b>	<b>19.89</b>	7.35	34.40	7.29	21.12
C	56.07	92.26	45.31	78.29	17.06	46.03	<b>14.53</b>	<b>36.99</b>	20.60	49.36
D	57.79	96.48	50.12	89.55	<b>15.68</b>	<b>45.79</b>	25.19	70.74	19.92	46.56

No contexto dos *clusters* A, B e D, o modelo RF apresentou um bom desempenho nesses segmentos. Observa-se que este método se sobressai ao demonstrar consistentemente valores reduzidos de MAE e RMSE em relação a outros modelos, refletindo a adaptabilidade do modelo à natureza dos dados nestes segmentos. Por outro lado, no *cluster* C, o ANND emerge como o modelo de destaque, exibindo um domínio sobre as métricas de desempenho.

Uma observação fundamental nesta análise reside na diferença de modelação entre os *clusters* A e B, em comparação com os *clusters* C e D. Esta distinção evidencia-se ao considerar as métricas MAE e RMSE, que indicam que os agrupamentos que englobam as cores mais claras, valores mais elevados de  $L$ , são significativamente mais facilmente modelados em comparação com os conjuntos que abrangem as cores mais escuras, respetivamente, onde se verifica uma duplicação substancial do erro. Por outro lado, nos *clusters* C e D, os modelos referidos emergem com uma vantagem relativa, embora seja importante ressaltar que todos os modelos enfrentam desafios significativos, devido aos valores consideravelmente mais elevados de MAE e RMSE. Esses desafios podem ser atribuídos às características específicas mencionadas na subsecção 3.4.2, onde se detetou uma dispersão expressiva e assimétrica dos dados nestes segmentos.

Notavelmente, apesar da distinção entre o tipo de acabamento nos agrupamentos, a análise revelou que a variável de acabamento não desempenha uma influência significativa na modelação das cores em relação às bases de tinta. Essa observação corrobora o que já havia sido identificado na secção 3.2, onde se constatou uma ausência de correlação significativa entre esses fatores.

Portanto, a análise destaca uma distinção clara na facilidade de modelação entre esses grupos, com implicações significativas na seleção de modelos apropriados para cada contexto. Consequentemente, quando se lida com cores pertencentes aos primeiros grupos, a escolha de um modelo específico pode ser uma estratégia mais eficaz, especialmente em virtude da maior aceitabilidade do erro a eles associado.

## 4.6 Uma Abordagem Prática da Ajusta de Tintas com *Machine Learning*

No âmbito da validação da aplicabilidade prática das abordagens desenvolvidas ao nível do negócio, um teste experimental foi conduzido para avaliar a eficácia das previsões geradas por *Machine Learning* em ambiente empresarial. Neste teste, o modelo funciona como um apoio ao afinador para calcular, automaticamente, a proporção de cada base de tinta a misturar para obter a cor desejada, a partir dos parâmetros CIELab, consoante as características cromáticas especificadas. A subsequente mistura das quantidades previstas das bases de tinta, de acordo com as proporções indicadas pelas previsões, proporcionou um cenário de ajusta realista,

com o objetivo de produzir uma amostra concreta do resultado obtido. A cor resultante passou por uma análise laboratorial rigorosa, na qual os parâmetros de cor CIELab foram medidos por um espectrofotômetro. Este procedimento permitiu uma comparação direta entre a cor pretendida e os resultados empíricos alcançados. Ao combinar as bases de tinta alinhadas com as previsões geradas, surgiram resultados tangíveis que refletem a convergência entre os dados analíticos e a aplicação prática.

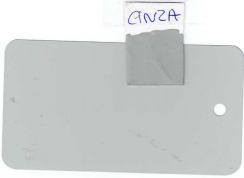


Os resultados deste processo empírico foram meticulosamente tabulados e organizados na tabela 4.6, refletindo os valores previstos por um dos modelos trabalhados, o RF da primeira abordagem, e os resultados obtidos para cada *target* em estudo.

Tabela 4.6: Resultados da Afinação de Tinta com base em Previsões do Modelo RF

Parâmetros CIELab (L; a; b)		1º Teste	2º Teste	3º Teste
Acabamento		Mate	Mate	Mate
Fórmula de Tinta	Mate	35.00	35.00	35.00
	Brilho	0.00	0.00	0.00
	Amarelo Vivo	0.00	0.00	22.20
	Laranja Claro	0.00	0.00	9.50
	Ocre	3.00	0.00	32.40
	Amarelo Alaranjado	0.00	0.00	0.00
	Laranja Escuro	0.00	0.13	0.00
	Vermelho Carmim	0.00	0.00	0.00
	Vermelho	0.10	0.00	0.00
	Bordeaux	0.00	0.50	0.00
	Castanho	0.15	0.00	0.00
	Azul	0.00	15.60	0.00
	Verde	0.00	0.35	0.00
	Verde Claro	0.00	0.00	0.00
	Branco	56.60	0.00	0.00
	Preto	2.00	0.00	0.00

Os resultados da aplicação do método prático foram registrados, fornecendo uma medida tangível da precisão das previsões do RF. A tabela 4.7 apresenta os parâmetros das cores obtidas após a mistura das bases de tintas, bem como os parâmetros das cores pretendidas.

Tabela 4.7: Comparação dos Resultados obtidos com as Cores Pretendidas

	1º Teste	2º Teste	3º Teste
Parâmetros CIELab da cor pretendida	(78.41; -1.70; 0.80)	(51.05; -3.54; -41.45)	(80.92; 6.68; 79.77)
Parâmetros CIELab da cor obtida	(60.69; -0.35; 1.01)	(52.16; -7.83; -30.60)	(68.18; 13.72; 57.10)
$\Delta E$	13.91	3.37	12.51
Variação Cromática			

As representações dos testes presentes na tabela 4.7 ilustram visualmente a comparação entre as cores obtidas (cartões pequenos) e as cores pretendidas (chapas

maiores), evidenciando as variações cromáticas entre os três conjuntos. No apêndice C, são apresentados os relatórios dos testes desenvolvidos.

Para avaliar a qualidade das previsões realizadas, recorreu-se ao cálculo do valor  $\Delta E$  entre as cores obtidas e as cores pretendidas. O valor  $\Delta E$  é uma métrica quantitativa que expressa a diferença entre duas cores em termos de percepção humana, proporcionando uma avaliação objetiva das previsões [105]. Essa análise estatística é fundamental para compreender até que ponto os algoritmos conseguem reproduzir as complexas relações entre os parâmetros de cor e as proporções de tinta.

No caso da segunda cor, o valor de  $\Delta E$  de 3.37 sugere uma previsão relativamente precisa, indicando uma pequena discrepância entre a cor prevista e a cor obtida, que é percebida num relance<sup>1</sup>. Por outro lado, os valores mais elevados de  $\Delta E$  para a primeira e terceira cor (13.91 e 12.51, respetivamente) apontam uma diferença perceptível entre a cor obtida e a cor pretendida.

No entanto, é importante referir que os valores de  $\Delta E$  não podem ser avaliados isoladamente. Eles devem ser considerados em conjunto com outros fatores, como a complexidade das cores em questão e as limitações inerentes do processo de mistura de tintas. Além disso, é fundamental reconhecer que a previsão de cores é um desafio complexo devido à natureza multidimensional das cores e à variedade de fatores que podem influenciar as percepções visuais.

Em resumo, este teste prático representa um passo crucial em direção à validação das estratégias propostas. Ao vincular diretamente os resultados teóricos às aplicações reais, consegue-se avaliar não apenas a capacidade preditiva, mas também a utilidade concreta dos modelos num contexto de produção de tintas. Este procedimento integrado funde a investigação analítica com a aplicação tangível, proporcionando uma compreensão da eficácia dos modelos de *Machine Learning* na previsão das concentrações de tinta com base nos parâmetros CIELab.

Entretanto, a aplicação prática revelou variações nos resultados em termos de  $\Delta E$ , o que não compromete as inferências derivadas da análise anterior dos resultados. Essa variação realça meramente a complexidade inerente ao desafio de prever cores com alta precisão, onde fatores como a interação entre diferentes bases de tinta podem influenciar os resultados.

Ainda que haja espaço para melhorias, particularmente em relação às cores que apresentam valores mais elevados de  $\Delta E$ , os resultados do teste prático ressaltam a utilidade e a promissora aplicabilidade do estudo na indústria de ajustagem de tintas. Essas considerações apontam para uma direção de desenvolvimento que visa explorar ainda mais o seu potencial e a sua adaptação específica para os desafios práticos enfrentados pela indústria.

---

<sup>1</sup>A percepção padrão é classificada em cinco categorias, nomeadamente  $\Delta E \in [0; 1]$ : Imperceptível ao olho humano,  $\Delta E \in ]1; 2]$ : Perceptível através de observação atenta,  $\Delta E \in ]2; 10]$ : Perceptível de relance,  $\Delta E \in ]10; 50]$ : As cores são mais semelhantes do que o oposto e  $\Delta E \in ]50; 100]$ : As cores são bastante diferentes [84].



## Capítulo 5

# Conclusões

A Inteligência Artificial é uma ferramenta poderosa, capaz de revelar descobertas significativas a partir de diferentes conjuntos de dados. No entanto, como evidenciado por este estudo, nem todos os algoritmos são igualmente adequados para abordar uma ampla variedade de problemas.

Esta dissertação constituiu uma exploração abrangente e inovadora no domínio da formulação de tintas, com ênfase particular na aplicação de métodos avançados de *Machine Learning* e na análise metódica dos dados associados.

Dado o papel crucial das bases de tinta na criação de uma ampla gama de tons por meio de combinações em diversas proporções, encontrar a mistura perfeita de bases para atingir uma cor específica é um desafio. Portanto, o objetivo primordial consistiu em desenvolver, pelo menos, um modelo capaz de antever fórmulas de tintas a partir de características específicas de cor e acabamento. Para atingir esse propósito, foram investigadas três abordagens diferentes, nomeadamente a modelação clássica baseada em algoritmos únicos, a combinação dos melhores modelos para cada base de tinta e a modelação baseada em *clusters* do espaço de cor CIE-Lab. Cada uma passou por uma análise pormenorizada, visando a avaliação da sua consistência e precisão nas previsões.

Ao determinar a abordagem mais promissora, é fundamental avaliar a adaptabilidade dos modelos aos dados e a consistência de suas previsões. Como se pode verificar pela tabela 5.1, o ANND da primeira abordagem, resultado da fusão entre o *Autoencoder-based Transfer Learning* e o MLP, apresentou um desempenho superior tanto em relação ao MAE quanto ao RMSE. Além disso, o algoritmo também se

Tabela 5.1: Resumo dos Resultados das Abordagens

	1ª Abordagem		2ª Abordagem		3ª Abordagem	
MAE	8.23	ANND	9.20	Combinação de modelos	10.76	RF
RMSE	29.12		30.87		35.60	

destacou como uma escolha interessante em todo o projeto, uma vez que foi um dos modelos eleitos para o *ensemble* da segunda abordagem e demonstrou uma notável capacidade em lidar com *targets* desafiantes na terceira abordagem. Desta forma e contrariamente às expectativas da literatura, a abordagem clássica baseada em algoritmos únicos superou a combinação de modelos e a modelação por grupos de dados.

Posteriormente, estes resultados e a não uniformidade da distribuição dos dados no espaço de cor CIELab motivaram uma análise detalhada da modelação de cada segmento de dados. Observou-se que a modelação apresentava diferenças significativas entre os *Clusters A* e *B* em comparação com os *Clusters C* e *D*, refletindo-se notavelmente nas métricas MAE e RMSE. Tal indica que os grupos de cores mais claras eram mais facilmente modelados do que os grupos de cores mais escuras, onde o erro substancialmente aumentava. Esses desafios podem ser atribuídos às características específicas desses grupos, tornando a modelação mais complexa. Portanto, ao lidar com cores mais claras, a escolha de um modelo específico pode ser uma estratégia mais eficaz.

Adicionalmente, foram afinadas tintas em ambiente empresarial utilizando um modelo de *Machine Learning*, o RF. Nestes testes, o modelo assume a função de uma ferramenta auxiliar para o afinador, permitindo o cálculo automático das proporções ideais de cada base de tinta, de modo a obter a tonalidade desejada. Desta forma, foi possível avaliar a eficácia das previsões em cenários realistas, com base no  $\Delta E$ . A aplicação concreta deste estudo evidenciou o seu potencial tangível, atendendo às especificações desejadas em termos de cor e acabamento. Neste sentido, ao nível do negócio, estes resultados marcam o ponto de partida para uma transformação significativa na automação deste processo, com o propósito fundamental de proporcionar ao afinador uma ferramenta mais eficaz e objetiva.

Em resumo, este estudo não apenas contribui para o avanço do conhecimento na área de investigação, mas também demonstrou claramente o potencial das abordagens de *Machine Learning* na afinação de tintas. Portanto, esta dissertação deixa uma indelével marca no campo da formulação de tintas, destacando o poder da análise de dados e da implementação de estratégias inteligentes para enfrentar os desafios dessa indústria em constante evolução.

## 5.1 Trabalho Futuro

Considerando o progresso alcançado neste estudo, várias e empolgantes oportunidades para pesquisas futuras emergem, que podem servir como bases para investigações subsequentes.

Primeiramente, é imperativo considerar a expansão do conjunto de dados. Isso evitará possíveis sub-representações de determinados segmentos de cor e garantirá que haja um volume adequado de observações para treinar os modelos numa base mais sólida.

Adicionalmente, a exploração de técnicas de *Deep Learning*, como DNN, representa um caminho promissor para aprimorar a previsão de fórmulas de tinta, abrindo espaço para investigações mais profundas e complexas, como efetuado por Mei-Yun Chen et al. [86].

Além disso, a integração de *Transfer Learning* com outros modelos demonstraram bom desempenho neste estudo e, portanto oferece um campo fértil para pesquisas adicionais. Isso pode envolver a adaptação de conhecimento prévio de modelos treinados para os diferentes modelos estudados, visando uma análise mais ampla do seu desempenho e potencial. No estudo conduzido por Tan et al. [55], também é aplicada, ao final do processo, uma técnica conhecida como *Manifold Constraints*. Essas restrições têm como objetivo forçar o algoritmo a transferir apenas os conhecimentos relevantes do domínio de origem, enquanto ignora aqueles que podem introduzir inconsistências na estrutura do *Manifold* do domínio de destino. Considerando os resultados promissores dessa abordagem, um trabalho futuro intrigante seria explorar a sua aplicação.

Uma área importante de exploração seria a imposição de restrições sobre o número de bases de tinta em cada afinação com o objetivo de mitigar os efeitos de metamerismo. Este fenômeno afeta a forma como as cores são percebidas em diferentes condições de iluminação e observação. Ao abordar essa questão, seria viável desenvolver soluções mais práticas e alinhadas com as complexidades da indústria de formulação de tintas, promovendo uma uniformidade maior nas cores.

Desta forma, os resultados alcançados, juntamente com as direções futuras delineadas, refletem o compromisso contínuo com a inovação e a procura pela excelência. À medida que essas direções futuras são exploradas, têm o potencial de aprimorar a eficiência e a precisão da afinação de tintas, consolidando o impacto deste projeto no campo da indústria de tintas.



# Referências

- [1] R. X. V. Sandoval and M. Pedersen, “Cmyk-cielab color space transformation using machine learning techniques,” *London Imaging Meeting (LIM)*, pp. 73–77, 2021. [Citado na página 1]
- [2] L. Yang, *Ink-Paper Interaction: A study in ink-jet color reproduction*. Univ., 2003. [Citado na página 1]
- [3] J. Yuan, G. Chen, H. Li, H. Prautzsch, and K. Xiao, “Accurate and computational: A review of color reproduction in full-color 3d printing,” *Materials Design*, vol. 209, p. 109943, 2021. [Citado na página 2]
- [4] S. Westland, “Artificial neural networks and colour recipe prediction,” in *Proceedings of the International Conference and Exhibition: Colour Science*, pp. 225–233, 1998. [Citado nas páginas 2 e 19]
- [5] A. Akanuma and D. Stamate, “A neural network approach to estimating color reflectance with product independent models,” in *Artificial Neural Networks and Machine Learning – ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part III*, (Berlin, Heidelberg), p. 803–806, Springer-Verlag, 2022. [Citado na página 2]
- [6] Y. Qu, *Color Prediction and Separation Models in Printing : Minimizing the Colorimetric and Spectral Differences employing Multiple Characterization Curves*. PhD thesis, Linköping University, Department of Science and Technology, Media and Information Technology. Linköping University, The Institute of Technology, 2013. [Citado na página 2]
- [7] C. Fehrman and K. Fehrman, *Color: The Secret Influence (Fourth Edition)*. Cognella, Incorporated, 2018. [Citado na página 2]
- [8] P. R. Bélanger, “Linear-programming approach to color-recipe formulations,” *Journal of the Optical Society of America*, vol. 64, pp. 1541–1544, 1974. [Citado na página 2]
- [9] J. M. Bishop, M. J. Bushnell, A. Usher, and S. Westland, “Genetic optimisation of neural network architectures for colour recipe prediction,” in *Artificial Neural Nets and Genetic Algorithms* (R. F. Albrecht, C. R. Reeves, and N. C. Steele, eds.), (Vienna), pp. 719–725, Springer Vienna, 1993. [Citado na página 2]

- [10] K. M. Lehmann, C. Igiel, I. Schmidtman, and H. Scheller, “Four color-measuring devices compared with a spectrophotometric reference system,” *Journal of Dentistry*, vol. 38, pp. e65–e70, 2010. *Journal of Color and Appearance in Dentistry*. [Citado nas páginas 3 e 8]
- [11] C. P. Coutinho, *Metodologia de investigação em ciências sociais e humanas*. Leya, 2014. [Citado nas páginas 3 e 4]
- [12] GLOMA, “25 anos gloma®.” <https://gloma.pt/>. [Citado na página 4]
- [13] K. Kusuma, M. Larsen, J. C. Quiroz, M. Gillies, A. Burnett, J. Qian, and M. Torok, “The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review,” *Journal of Psychiatric Research*, vol. 155, pp. 579–588, 2022. [Citado na página 5]
- [14] “Tintas e vernizes - termos e definições (iso 4618:2014),” 2022. [Citado na página 7]
- [15] L. W. McKeen, “3 - the components of paint,” in *Fluorinated Coatings and Finishes Handbook (Second Edition)* (L. W. McKeen, ed.), *Plastics Design Library*, pp. 51–58, Oxford: William Andrew Publishing, second edition ed., 2016. [Citado na página 7]
- [16] A. I. T. R. de Magalhães, “Otimização do sistema aquoso de afinação automática,” Master’s thesis, Faculdade de Engenharia da Universidade do Porto, Porto, 2017. [Citado nas páginas 8 e 9]
- [17] J. Schanda and I. Illumination, *Colorimetry: Understanding the CIE System*. Wiley-Interscience, 03 2007. [Citado na página 8]
- [18] S. Ishikawa-Nagai, A. Yoshida, M. Sakai, J. Kristiansen, and J. D. Da Silva, “Clinical evaluation of perceptibility of color differences between natural teeth and all-ceramic crowns,” *Journal of Dentistry*, vol. 37, pp. e57–e63, 2009. *Journal of Color and Appearance in Dentistry*. [Citado na página 9]
- [19] R. Korifi, Y. Le Dréau, J.-F. Antinelli, R. Valls, and N. Dupuy, “Ciel\*a\*b\* color space predictive models for colorimetry devices – analysis of perfume quality,” *Talanta*, vol. 104, pp. 58–66, 2013. [Citado na página 9]
- [20] L. Sawaya, “Chroma – explained in plain english with case studies.” <https://thelandofcolor.com/chroma-explained-in-plain-english-with-case-studies/>. [Citado na página 9]

- [21] M. Pegalajar, L. Ruiz, M. Sánchez-Marañón, and L. Mansilla, “A munsell colour-based approach for soil classification using fuzzy logic and artificial neural networks,” *Fuzzy Sets and Systems*, vol. 401, pp. 38–54, 2020. Fuzzy Measures, Integrals and Quantification in Artificial Intelligence Problems – An Homage to Prof. Miguel Delgado. [Citado na página 9]
- [22] A. Zeileis, K. Hornik, and P. Murrell, “Escaping rglab: Selecting colors for statistical graphics,” *Computational Statistics Data Analysis*, vol. 53, no. 9, pp. 3259–3270, 2009. [Citado na página 9]
- [23] I. L. Weatherall and B. D. Coombs, “Skin color measurements in terms of cielab color space values,” *Journal of Investigative Dermatology*, vol. 99, no. 4, pp. 468–473, 1992. [Citado na página 9]
- [24] S. F. A. P. Gil, “Criação de uma base de dados destinada à formulação automática de cores para uma tinta de fachadas, aquosa e de formulação genérica, com uma gama nova de pastas corantes de base aquosa,” Master’s thesis, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Lisboa, 2013. [Citado nas páginas 9 e 20]
- [25] S. Dissaneevate, T. Wongsirichot, P. Siritwat, N. Jintanapanya, U. Boonyakarn, W. Janjindamai, A. Thatrimontrichai, and G. Maneenil, “A mobile computer-aided diagnosis of neonatal hyperbilirubinemia using digital image processing and machine learning techniques,” *International Journal of Innovative Research and Scientific Studies*, vol. 5, no. 1, pp. 10–17, 2022. [Citado na página 10]
- [26] A. Grzybowski and K. Kupidura-Majewski, “What is color and how it is perceived?,” *Clinics in Dermatology*, vol. 37, no. 5, pp. 392–401, 2019. The Color of Skin. [Citado na página 10]
- [27] R. Stauffer, G. Mayr, M. Dabernig, and A. Zeileis, “Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations,” *Bulletin of the American Meteorological Society*, vol. 96, p. 140710055335002, 07 2014. [Citado na página 10]
- [28] C.-h. Li and P. C. Yuen, “Regularized color clustering in medical image database,” *IEEE transactions on medical imaging*, vol. 19, pp. 1150–5, 12 2000. [Citado na página 10]
- [29] C. Gómez-Polo, M. Portillo Muñoz, M. C. Lorenzo Luengo, P. Vicente, P. Galindo, and A. M. Martín Casado, “Comparison of two color-difference formulas using the bland-altman approach based on natural tooth color space,” *The*

- Journal of Prosthetic Dentistry*, vol. 115, no. 4, pp. 482–488, 2016. [Citado na página 10]
- [30] R. Warner, “Measurement of meat quality | measurements of water-holding capacity and color: Objective and subjective,” in *Encyclopedia of Meat Sciences (Second Edition)* (M. Dikeman and C. Devine, eds.), pp. 164–171, Oxford: Academic Press, second edition ed., 2014. [Citado na página 10]
- [31] B. Hill, T. Roger, and F. W. Vorhagen, “Comparative analysis of the quantization of color spaces on the basis of the cielaab color-difference formula,” *ACM Trans. Graph.*, vol. 16, p. 109–154, apr 1997. [Citado na página 10]
- [32] J. Seymour, “Why does the cielaab a\* axis point toward magenta instead of red?,” *Color Research & Application*, vol. 45, no. 6, pp. 1040–1054, 2020. [Citado na página 10]
- [33] B. C. K. Ly, E. B. Dyer, J. L. Feig, A. L. Chien, and S. Del Bino, “Research techniques made simple: Cutaneous colorimetry: A reliable technique for objective skin-color measurement,” *Journal of Investigative Dermatology*, vol. 140, no. 1, pp. 3–12.e1, 2020. [Citado na página 11]
- [34] B. M. F. Teixeira, “Análise e previsão de acidentes rodoviários usando data mining,” Master’s thesis, Instituto Superior de Engenharia do Porto - Instituto Politécnico do Porto, Porto, 2017. [Citado na página 11]
- [35] W. Bao, N. Lianju, and K. Yue, “Integration of unsupervised and supervised machine learning algorithms for credit risk assessment,” *Expert Systems with Applications*, vol. 128, pp. 301–315, 2019. [Citado nas páginas 11, 13 e 17]
- [36] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, “Survey on multi-output learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2409–2429, 2020. [Citado na página 11]
- [37] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, “Survey on multi-output learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2409–2429, 2020. [Citado nas páginas 11 e 12]
- [38] W. Liu, Y. Yang, Q. Xu, and Y. Xia, “Multi-target prediction model of urban distribution system rainfall-caused outage based on spatiotemporal fusion,” *International Journal of Electrical Power Energy Systems*, vol. 146, p. 108640, 2023. [Citado na página 12]
- [39] G. Melki, A. Cano, V. Kecman, and S. Ventura, “Multi-target support vector regression via correlation regressor chains,” *Information Sciences*, vol. 415–416, pp. 53–69, 2017. [Citado na página 12]

- 
- [40] H. Borchani, G. Varando, C. Bielza, and P. Larranaga, “A survey on multi-output regression,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, 07 2015. [Citado na página 12]
- [41] S. Barbon Junior, S. M. Mastelini, A. P. A. Barbon, D. F. Barbin, R. Calvini, J. F. Lopes, and A. Ulrici, “Multi-target prediction of wheat flour quality parameters with near infrared spectroscopy,” *Information Processing in Agriculture*, vol. 7, no. 2, pp. 342–354, 2020. [Citado na página 12]
- [42] A. F. Siegel and M. R. Wagner, “Chapter 11 - correlation and regression: Measuring and predicting relationships,” in *Practical Business Statistics (Eighth Edition)* (A. F. Siegel and M. R. Wagner, eds.), pp. 313–370, Academic Press, eighth edition ed., 2022. [Citado nas páginas 12 e 31]
- [43] Y. Zhao, “Chapter 5 - regression,” in *R and Data Mining* (Y. Zhao, ed.), pp. 41–50, Academic Press, 2013. [Citado na página 12]
- [44] X. Brusset, D. La Torre, and J. Broekaert, “Chapter 6 - algorithms, analytics, and artificial intelligence: harnessing data to make supply chain decisions,” in *The Digital Supply Chain* (B. L. MacCarthy and D. Ivanov, eds.), pp. 93–110, Elsevier, 2022. [Citado na página 12]
- [45] J. Schneider, M. Dziubany, A. Schmeink, G. Dartmann, K.-U. Gollmer, and S. Naumann, “Chapter 8 - predicting energy consumption using machine learning,” in *Big Data Analytics for Cyber-Physical Systems* (G. Dartmann, H. Song, and A. Schmeink, eds.), pp. 167–186, Elsevier, 2019. [Citado na página 13]
- [46] A. Singhal and D. K. Sharma, “Chapter 3 - voice signal-based disease diagnosis using iot and learning algorithms for healthcare,” in *Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain* (C. Chakraborty, S. K. Pani, M. Abdul Ahad, and Q. Xin, eds.), Intelligent Data-Centric Systems, pp. 59–81, Academic Press, 2023. [Citado nas páginas 13 e 14]
- [47] S. Abinaya and M. K. Devi, “Chapter 12 - enhancing crop productivity through autoencoder-based disease detection and context-aware remedy recommendation system,” in *Application of Machine Learning in Agriculture* (M. A. Khan, R. Khan, and M. A. Ansari, eds.), pp. 239–262, Academic Press, 2022. [Citado na página 13]
- [48] S. Abirami and P. Chitra, “Chapter fourteen - energy-efficient edge based real-time healthcare support system,” in *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases* (P. Raj and P. Evangelina, eds.), vol. 117 of *Advances in Computers*, pp. 339–368, Elsevier, 2020. [Citado nas páginas 13 e 14]

- [49] A. Meyer-Baese and V. Schmid, “Chapter 7 - foundations of neural networks,” in *Pattern Recognition and Signal Analysis in Medical Imaging (Second Edition)* (A. Meyer-Baese and V. Schmid, eds.), pp. 197–243, Oxford: Academic Press, second edition ed., 2014. [Citado nas páginas 13 e 31]
- [50] T. Menzies, E. Kocagüneli, L. Minku, F. Peters, and B. Turhan, “Chapter 24 - using goals in model-based reasoning,” in *Sharing Data and Models in Software Engineering* (T. Menzies, E. Kocagüneli, L. Minku, F. Peters, and B. Turhan, eds.), pp. 321–353, Boston: Morgan Kaufmann, 2015. [Citado na página 13]
- [51] J. Noh, T. Badloe, C. Lee, J. Yun, S. So, and J. Rho, “Chapter 1 - inverse design meets nanophotonics: From computational optimization to artificial neural network,” in *Intelligent Nanotechnology* (Y. Zheng and Z. Wu, eds.), Materials Today, pp. 3–32, Elsevier, 2023. [Citado na página 13]
- [52] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, “State-of-the-art in artificial neural network applications: A survey,” *Heliyon*, vol. 4, no. 11, p. e00938, 2018. [Citado na página 14]
- [53] J. Baek and Y. Choi, “Deep neural network for predicting ore production by truck-haulage systems in open-pit mines,” *Applied Sciences*, vol. 10, no. 5, 2020. [Citado na página 14]
- [54] S. Luo, X. Huang, Y. Wang, R. Luo, and Q. Zhou, “Transfer learning based on improved stacked autoencoder for bearing fault diagnosis,” *Knowledge-Based Systems*, vol. 256, p. 109846, 2022. [Citado na página 14]
- [55] C. Tan, F. Sun, B. Fang, T. Kong, and W. Zhang, “Autoencoder-based transfer learning in brain-computer interface for rehabilitation robot,” *International Journal of Advanced Robotic Systems*, vol. 16, p. 172988141984086, 03 2019. [Citado nas páginas 14 e 53]
- [56] Hung, Chung-Wen, W.-T. Li, W.-L. Mao, , and P.-C. Lee, “Design of a chamfering tool diagnosis system using autoencoder learning method,” *Energies*, vol. 12, p. 3708, 09 2019. [Citado na página 15]
- [57] C. Costa and C. A. Ferreira, “Modeling the ink tuning process using machine learning.” Artigo submetido à conferência *IDEAL 2023 (24th International Conference on Intelligent Data Engineering and Automated Learning)*., 2023. [Citado nas páginas 14 e 32]
- [58] P. Gedeck, C. Kramer, and P. Ertl, “4 - computational analysis of structure-activity relationships,” in *Progress in Medicinal Chemistry* (G. Lawton and D. Witty, eds.), vol. 49, pp. 113–160, Elsevier, 2010. [Citado na página 15]

- [59] R. Srivastava, S. Kumar, and B. Kumar, “7 - classification model of machine learning for medical data analysis,” in *Statistical Modeling in Machine Learning* (T. Goswami and G. Sinha, eds.), pp. 111–132, Academic Press, 2023. [Citado na página 15]
- [60] P. Kulkarni, V. Sreekanth, A. R. Upadhyaya, and H. C. Gautam, “Which model to choose? performance comparison of statistical and machine learning models in predicting pm2.5 from high-resolution satellite aerosol optical depth,” *Atmospheric Environment*, vol. 282, p. 119164, 2022. [Citado na página 15]
- [61] Y. Xia, “Chapter eleven - correlation and association analyses in microbiome study integrating multiomics in health and disease,” in *The Microbiome in Health and Disease* (J. Sun, ed.), vol. 171 of *Progress in Molecular Biology and Translational Science*, pp. 309–491, Academic Press, 2020. [Citado nas páginas 15 e 32]
- [62] D. Segura, E. Khatib, and R. Barco, “Dynamic packet duplication for industrial urllc,” *Sensors*, vol. 22, p. 587, 01 2022. [Citado na página 16]
- [63] A. Ziegler and I. König, “Mining data with random forests: Current options for real-world applications,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, 01 2014. [Citado na página 15]
- [64] T. Adetiloye and A. Awasthi, “Chapter 8 - predicting short-term congested traffic flow on urban motorway networks,” in *Handbook of Neural Computation* (P. Samui, S. Sekhar, and V. E. Balas, eds.), pp. 145–165, Academic Press, 2017. [Citado nas páginas 15 e 16]
- [65] R. Santhanam, N. Uzir, S. Raman, and S. Banerjee, “Experimenting xgboost algorithm for prediction and classification of different datasets,” *National Conference on Recent Innovations in Software Engineering and Computer Technologies (NCRISECT) 2017*, 03 2017. [Citado nas páginas 16 e 32]
- [66] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, “A high-bias, low-variance introduction to machine learning for physicists,” *Physics Reports*, vol. 810, pp. 1–124, 2019. A high-bias, low-variance introduction to Machine Learning for physicists. [Citado na página 16]
- [67] H. M. Rizeei, B. Pradhan, M. A. Saharkhiz, and S. Lee, “Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique,” *Journal of Hydrology*, vol. 579, p. 124172, 2019. [Citado na página 16]

- [68] A. Gómez-Ríos, J. Luengo, and F. Herrera, “A study on the noise label influence in boosting algorithms: Adaboost, gbm and xgboost,” in *Hybrid Artificial Intelligent Systems* (F. J. Martínez de Pisón, R. Urraca, H. Quintián, and E. Corchado, eds.), (Cham), pp. 268–280, Springer International Publishing, 2017. [Citado na página 16]
- [69] E. Sahin, “Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest,” *SN Applied Sciences*, vol. 2, 07 2020. [Citado na página 16]
- [70] A. Ibrahim Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, “Extreme gradient boosting (xgboost) model to predict the groundwater levels in selangor malaysia,” *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1545–1556, 2021. [Citado na página 16]
- [71] A. Scarinci and F. Marineli, “O modelo ondulatório da luz como ferramenta para explicar as causas da cor,” *Revista Brasileira de Ensino de Física*, vol. 36, pp. 1–14, 03 2014. [Citado na página 17]
- [72] W. Jiao, X.-j. Hu, L.-p. Tu, C.-l. Zhou, Z. Qi, Z.-y. Luo, L.-z. Zeng, X.-x. Ma, C.-H. Pai, H.-y. Fu, Y. Wang, J. Wang, and J.-t. Xu, “Tongue color clustering and visual application based on 2d information,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, 11 2019. [Citado na página 17]
- [73] B. Du, P. D. Lund, and J. Wang, “Improving the accuracy of predicting the performance of solar collectors through clustering analysis with artificial neural network models,” *Energy Reports*, vol. 8, pp. 3970–3981, 2022. [Citado nas páginas 17 e 36]
- [74] B. Roter, N. Ninkovic, and S. Dordevic, “Clustering superconductors using unsupervised machine learning,” *Physica C: Superconductivity and its Applications*, vol. 598, p. 1354078, 2022. [Citado na página 17]
- [75] D. Nagar, K. Pannerselvam, and P. Ramu, “A novel data-driven visualization of n-dimensional feasible region using interpretable self-organizing maps (isom),” *Neural Networks*, vol. 155, pp. 398–412, 2022. [Citado na página 17]
- [76] P. Schneider and F. Xhafa, “Chapter 8 - machine learning: ML for ehealth systems,” in *Anomaly Detection and Complex Event Processing over IoT Data Streams* (P. Schneider and F. Xhafa, eds.), pp. 149–191, Academic Press, 2022. [Citado nas páginas 17 e 18]

- [77] S. Simske, “Chapter 1 - introduction, overview, and applications,” in *Meta-Analytics* (S. Simske, ed.), pp. 1–98, Morgan Kaufmann, 2019. [Citado na página 17]
- [78] C. Chen, X. Xie, M. Yang, I. Seok, Z. Guo, Q. Jiang, G. Wangila, H. Zhang, and Q. Liu, “Recent advances in solar energy full spectrum conversion and utilization,” *ES Energy Environment*, 01 2021. [Citado na página 18]
- [79] X. Lü, T. Lu, C. J. Kibert, and M. Viljanen, “A novel dynamic modeling approach for predicting building energy performance,” *Applied Energy*, vol. 114, pp. 91–103, 2014. [Citado nas páginas 18 e 36]
- [80] Z. Wang and R. S. Srinivasan, “A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models,” *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 796–808, 2017. [Citado na página 18]
- [81] S. Chaouch, A. Moussa, I. Ben Marzoug, and N. Ladhari, “Colour recipe prediction using ant colony algorithm: principle of resolution and analysis of performances,” *Coloration Technology*, vol. 135, pp. 349–360, 06 2019. [Citado nas páginas 18, 19 e 36]
- [82] J. Bishop, M. Bushnell, A. Usher, and S. Westland, “Neural networks in the colour industry,” *Applications of Artificial Intelligence in Engineering VI*, pp. 423–434, 01 1991. [Citado nas páginas 18 e 19]
- [83] S. Chaouch, A. Moussa, I. Ben Marzoug, and N. Ladhari, “Application of genetic algorithm to color recipe formulation using reactive and direct dyestuffs mixtures,” *Color Research Application*, vol. 45, 05 2020. [Citado nas páginas 19, 20 e 36]
- [84] S. Girdthep, O. Hanmeng, N. Triamnak, N. Chailek, and N. Wanichacheva, “Theoretical solvent selection for nanostructured surface fabrication of reusable and colourimetric visual-eye sensor strips with rhodamine derivative-encapsulated polymeric membranes for highly sensitive and selective detection of hg<sup>2+</sup>,” *Polymer Testing*, vol. 97, p. 107151, 2021. [Citado nas páginas 19 e 49]
- [85] Y.-B. Huang, M.-Y. Chen, and M. Ouhyoung, “Perceptual-based cnn model for watercolor mixing prediction,” in *ACM SIGGRAPH 2018 Posters*, SIGGRAPH '18, New York, NY, USA: Association for Computing Machinery, 2018. [Citado na página 19]
- [86] M.-Y. Chen, Y.-B. Huang, S.-P. Chang, and M. Ouhyoung, “Prediction model for semitransparent watercolor pigment mixtures using deep learning with a

- dataset of transmittance and reflectance,” *arXiv preprint arXiv:1904.00275*, 2019. [Citado nas páginas 19 e 53]
- [87] D. Golob, D. P. Osterman, and J. Zupan, “Determination of pigment combinations for textile printing using artificial neural networks,” *Fibres & Textiles in Eastern Europe*, vol. 16, no. 3, p. 68, 2008. [Citado na página 19]
- [88] L. Yang and S. J. Miklavcic, “Revised kubelka–munk theory. iii. a general theory of light propagation in scattering and absorptive media,” *J. Opt. Soc. Am. A*, vol. 22, pp. 1866–1873, Sep 2005. [Citado na página 19]
- [89] J. Militký, “3 - fundamentals of soft models in textiles,” in *Soft Computing in Textile Engineering* (A. Majumdar, ed.), Woodhead Publishing Series in Textiles, pp. 45–102, Woodhead Publishing, 2011. [Citado nas páginas 19 e 20]
- [90] P. Saxena and Y. Parasher, “Application of artificial neural network (ann) for animal diet formulation modeling,” *Procedia Computer Science*, vol. 152, pp. 261–266, 2019. International Conference on Pervasive Computing Advances and Applications- PerCAA 2019. [Citado na página 20]
- [91] L. Zhou, L. Tan, C. Zhang, N. Zhao, Y. He, and Z. Qiu, “A portable nir-system for mixture powdery food analysis using deep learning,” *LWT*, vol. 153, p. 112456, 2022. [Citado nas páginas 20 e 36]
- [92] Y.-J. Chang, C.-Y. Jui, W.-J. Lee, and A.-C. Yeh, “Prediction of the composition and hardness of high-entropy alloys by machine learning,” *JOM*, vol. 71, 07 2019. [Citado na página 21]
- [93] P. Bannigan, M. Aldeghi, Z. Bao, F. Häse, A. Aspuru-Guzik, and C. Allen, “Machine learning directed drug formulation development,” *Advanced Drug Delivery Reviews*, vol. 175, p. 113806, 2021. [Citado na página 21]
- [94] A. Adeniyi, J. O. Ighalo, and G. Marques, “Utilisation of machine learning algorithms for the prediction of syngas composition from biomass bio-oil steam reforming,” *International Journal of Sustainable Energy*, vol. 40, pp. 1–16, 08 2020. [Citado nas páginas 21 e 36]
- [95] T. Yang, I. Arief, M. Niemann, M. Houbiers, K. Meisingset, A. Martins, and L. Froelich, “A machine learning approach to predict gas oil ratio based on advanced mud gas data,” *SPE Europec featured at 81st EAGE Conference and Exhibition*, 06 2019. [Citado na página 21]
- [96] B. Mirza, X. Li, K. Lauwers, B. Reddy, A. Muller, C. Wozniak, and S. Djali, “A clinical site workload prediction model with machine learning lifecycle,” *Healthcare Analytics*, vol. 3, p. 100159, 2023. [Citado nas páginas 23 e 30]

- [97] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, “Dmme: Data mining methodology for engineering applications – a holistic extension to the crisp-dm model,” *Procedia CIRP*, vol. 79, pp. 403–408, 2019. 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 18-20 July 2018, Gulf of Naples, Italy. [Citado na página 24]
- [98] A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum, and M. R. Berthold, “Knime for reproducible cross-domain analysis of life science data,” *Journal of Biotechnology*, vol. 261, pp. 149–156, 2017. Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the German Network for Bioinformatics Infrastructure. [Citado na página 24]
- [99] F. Villarroel Ordenes and R. Silipo, “Machine learning for marketing on the knime hub: The development of a live repository for marketing applications,” *Journal of Business Research*, vol. 137, pp. 393–410, 2021. [Citado na página 24]
- [100] H. Janik, M. Sienkiewicz, and J. Kucinska-Lipka, “9 - polyurethanes,” in *Handbook of Thermoset Plastics (Third Edition)* (H. Dodiuk and S. H. Godman, eds.), pp. 253–295, Boston: William Andrew Publishing, third edition ed., 2014. [Citado na página 26]
- [101] T. Song, L. Ding, L. Yang, J. Ran, and L. Zhang, “Comparison of machine learning models for performance evaluation of wind-induced vibration piezoelectric energy harvester with fin-shaped attachments,” *Ocean Engineering*, vol. 280, p. 114630, 2023. [Citado na página 33]
- [102] E. Sanchez, S. Pintos, and N. Queipo, “Toward an optimal ensemble of kernel-based approximations with engineering applications,” in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, vol. 36, pp. 2152–2158, 01 2006. [Citado na página 34]
- [103] X. Liu, H. Tang, Y. Ding, and D. Yan, “Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings,” *Energy and Buildings*, vol. 273, p. 112408, 2022. [Citado na página 36]
- [104] N. F. Salehuddin, M. Binti Omar, R. Ibrahim, and K. Bingi, “A neural network-based model for predicting saybolt color of petroleum products,” *Sensors*, vol. 22, p. 2796, 04 2022. [Citado na página 36]
- [105] A. T. d. M. Ferreira and G. F. Vieira, “Metamerismo: análise da influência do iluminante no comportamento da cor de dentes naturais e materiais restauradores estéticos,” Master’s thesis, Universidade de São Paulo, 2002. [Citado na página 49]



## Apêndice A

# Estatísticas dos Dados

A compreensão das nuances e características intrínsecas dos dados é um alicerce fundamental em qualquer projeto que envolva análise e modelação. Neste contexto, a análise estatística das variáveis desempenha um papel importante, fornecendo *insights* críticos sobre a natureza dos dados.

O presente apêndice apresenta uma análise extensa a partir das figuras A.1 e A.2, abordando diversos aspetos, como a distribuição das variáveis, tendências ao longo do tempo e relações complexas entre os dados. Além disso, disponibiliza informações cruciais, como os valores mínimos, máximos, médios de cada variável, bem como o desvio padrão, a variância, a *skewness* e a *kurtosis*. Adicionalmente, enriquece essa compreensão com representações visuais, cada variável é acompanhada de um histograma que traduz graficamente a distribuição dos seus valores.

Por meio desta análise estatística abrangente, é possível identificar padrões muitas vezes invisíveis à primeira vista, identificam-se valores discrepantes e, o mais crucial, compreende-se a estrutura subjacente dos dados, de modo a efetuar uma modelação coerente na secção principal deste estudo.

Este apêndice serve como um suporte sólido para a análise principal sobre a qual todo o estudo foi construído, garantindo que os resultados obtidos sejam robustos, confiáveis e contextualmente relevantes. A compreensão profunda dos dados assegura decisões e conclusões derivadas da análise principal mais informadas, contribuindo assim para a validade e a qualidade do trabalho em questão.

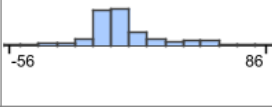


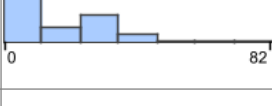
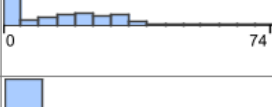
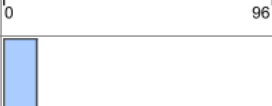
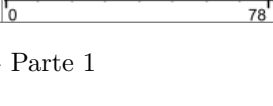
S	Column	D	Min	D	Max	D	Mean	D	Std. deviation	D	Variance	D	Skewness	D	Kurtosis	Histogram
	a		-56.04		85.76		6.874		18.298		334.824		0.965		1.459	
	b		-47.81		93.91		8.147		19.821		392.88		0.973		2.531	
	L		17.5		97.24		58.11		21.21		449.858		-0.069		-1.283	
	Mate		0		82.06		11.995		14.464		209.199		0.674		-0.963	
	Brilho		0		73.64		11.941		13.421		180.12		0.724		-0.484	
	Amarelo Vivo		0		81.76		0.758		4.971		24.706		9.269		98.59	
	Laranja Claro		0		74.22		2.291		7.679		58.974		4.458		22.419	
	Ocre		0		95.85		8.343		13.42		180.084		2.208		5.183	
	Amarelo Alaranjado		0		77.61		0.982		6.02		36.237		8.002		71.154	

Figura A.1: Resultados das Estatísticas dos Dados - Parte 1

S	Column	D	Min	D	Max	D	Mean	D	Std. deviation	D	Variance	D	Skewness	D	Kurtosis	Histogram
	Laranja Escuro	0	57.93	0.974	4.954	24.542	6.764	50.518								
	Vermelho Carmin	0	59.74	1.113	5.459	29.803	6.175	41.289								
	Vermelho	0	79.73	5.067	11.02	121.436	3.214	11.188								
	Bordeaux	0	74.95	1.232	6.097	37.172	6.611	51.229								
	Castanho	0	57.98	0.68	3.826	14.638	8.002	79.688								
	Azul	0	67.86	4.988	12.115	146.779	2.638	6.014								
	Verde	0	68.7	1.581	5.994	35.932	5.545	35.622								
	Verde Claro	0	50.68	0.087	1.42	2.015	20.812	518.514								
	Branco	0	98.44	43.445	28.799	829.363	-0.065	-1.397								
	Preto	0	74.25	4.012	8.14	66.259	3.116	11.277								

Figura A.2: Resultados das Estatísticas dos Dados - Parte 2



## Apêndice B

# Complemento aos Resultados e Discussão

No presente apêndice, é apresentada uma extensão do capítulo 4, onde os resultados e discussões são detalhados na seção principal deste trabalho. Com o objetivo de proporcionar uma visão mais completa e transparente das análises realizadas, este complemento fornece tabelas auxiliares que abrangem os resultados específicos do conjunto de treino da primeira abordagem avaliada e a análise da modelação por grupo.

### B.1 Conjunto de treino dos modelos da 1<sup>a</sup> Abordagem

A tabela B.1 apresenta os resultados das métricas de desempenho dos modelos no conjunto de treino, que são importantes para a formação do *ensemble* de modelos analisado na segunda abordagem que se desenvolveu.

A salientar que os resultados de desempenho no conjunto de treino são maioritariamente melhores do que os resultados no conjunto de teste. Isso ocorre porque o modelo é minuciosamente ajustado e otimizado com base nos dados de treino, adaptando-se aos padrões específicos desse conjunto de dados. Consequentemente, essa adaptação excessiva pode levar a uma queda no desempenho quando o modelo é confrontado com dados não vistos no conjunto de teste, onde a generalização das relações entre as variáveis se torna um fator crucial.

Tabela B.1: Métricas de desempenho dos modelos no Conjunto de Treino

Target	Conjunto de Treino									
	LR		MLP		ANND		RF		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.16	0.91	0.01	0.37	0	0.03	0.01	0.18	0.02	0.35
Brilho	0.17	1.68	0	0	0.09	0.13	0.01	0.09	0.01	0.15
Amarelo Vivo	1.33	4.01	0.81	4.07	0.41	2.45	0.13	0.78	0.34	1.49
Laranja Claro	2.52	5.33	1.51	4.54	0.23	0.71	0.28	1.04	0.6	1.78
Ocre	6.71	10.23	4.08	7.05	0.95	1.54	0.7	1.59	1.53	3
Amarelo Alaranjado	1.48	4.02	0.52	2.8	0.29	1.82	0.11	0.72	0.27	1.33
Laranja Escuro	1.41	3.8	0.61	2.71	0.43	2.56	0.14	0.67	0.38	1.31
Vermelho Carmim	1.77	4.43	1.12	4.69	0.65	3.72	0.16	0.65	0.43	1.36
Vermelho	3.99	7.23	3.3	6.78	1.43	2.78	0.51	1.33	1.11	2.4
Bordeaux	1.97	5.25	1.12	5.29	0.48	2.31	0.19	0.9	0.43	1.63
Castanho	1.07	3.2	0.62	2.86	0.59	3.46	0.13	0.58	0.33	1.16
Azul	4.73	7.86	1.51	3.87	0.47	1.18	0.31	1.05	0.68	2.01
Verde	2.03	4.51	1.32	5.07	0.62	2.34	0.16	0.68	0.4	1.39
Verde Claro	0.15	1.1	0.08	1.26	0.08	1.25	0.03	0.39	0.06	0.67
Branco	7.13	9.31	4.78	7.45	1.09	2.04	1.08	2.25	2.25	4.05
Preto	2.83	5.22	1.53	3.79	0.45	0.88	0.28	0.92	0.71	1.82
Total	39.47	78.09	22.92	62.59	8.26	29.2	4.2	13.83	9.53	25.91

Portanto, essa discrepância entre os conjuntos de treino e teste destaca a importância de avaliar o desempenho em dados não utilizados. A capacidade do modelo de generalizar para novos dados é um indicador crítico da sua robustez. Portanto, a validação em conjuntos de teste independentes desempenha um papel fundamental na garantia de que o modelo não apenas se ajusta bem aos dados de treino, mas também é capaz de fornecer previsões precisas e confiáveis em contextos diversos.

## B.2 Complemento à Comparação da Modelação entre *Clusters*

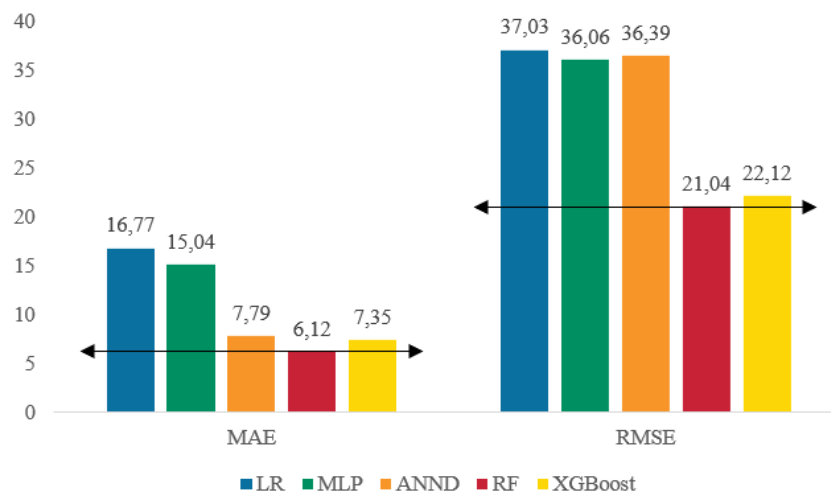
Além da análise por agrupamento discutida na subsecção 4.5, será conduzida uma análise individualizada para cada grupo de dados, juntamente com a identificação dos modelos mais apropriados para os subgrupos e suas características específicas.

A tabela B.2 apresenta os resultados das métricas de desempenho dos modelos no *Cluster A*.

Analisando as métricas globais da tabela B.2, fica evidente a presença de variações significativas nos valores de MAE e RMSE entre os diferentes modelos. Os valores mais baixos de MAE e RMSE foram obtidos pelo modelo RF, registrando 6.12 e 21.04, respectivamente, como ilustrado na figura B.1. Esse nível de consistência sugere que o modelo RF tem a habilidade de captar as complexas relações existentes entre os parâmetros de cor e as concentrações de tinta, consolidando-se como uma escolha sólida para previsões nesse contexto.

Tabela B.2: Métricas de desempenho dos modelos no Conjunto de Teste do *Cluster A*

Target	Cluster A - Conjunto de teste									
	LR		MLP		ANND		RF		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.03	0.74	0.27	1.01	0.05	1.18	0.04	0.84	0.04	0.73
Brilho	0.03	0.72	0.08	1.64	0.06	1.17	0.04	0.72	0.04	0.73
Amarelo Vivo	1.27	4.28	0.66	3.17	1.26	5.66	0.27	1.92	0.32	1.99
Laranja Claro	1.38	3.18	1.23	2.71	1.23	3.56	0.45	1.67	0.59	1.80
Ocre	3.07	4.76	2.29	3.88	0.02	0.06	1.28	2.67	1.42	2.71
Amarelo Alaranjado	2.28	5.10	0.95	3.48	2.31	10.02	0.31	1.89	0.40	2.01
Laranja Escuro	0.22	0.75	0.55	1.08	0.12	0.61	0.12	0.58	0.15	0.61
Vermelho Carmim	0.18	0.68	0.16	0.70	0.14	0.89	0.13	0.64	0.15	0.64
Vermelho	0.84	1.73	1.41	2.92	0.02	0.04	0.45	1.31	0.54	1.37
Bordeaux	0.12	0.55	0.09	0.55	0.10	0.56	0.05	0.39	0.06	0.41
Castanho	0.11	0.47	0.73	1.29	0.10	0.57	0.07	0.37	0.07	0.37
Azul	0.80	1.86	0.61	1.87	0.74	2.68	0.26	1.22	0.33	1.27
Verde	0.88	2.76	0.53	2.46	0.96	3.94	0.19	0.96	0.25	1.14
Verde Claro	0.24	1.23	0.16	1.08	0.15	1.25	0.06	0.60	0.07	0.64
Branco	4.70	6.70	4.80	6.76	0.15	1.91	2.11	4.06	2.49	4.41
Preto	0.64	1.52	0.52	1.46	0.38	2.29	0.29	1.20	0.41	1.30
Total	16.77	37.03	15.04	36.06	7.79	36.39	6.12	21.04	7.35	22.12

Figura B.1: Comparação dos resultados do desempenho dos modelos no *Cluster A*

Na análise das previsões para cada base de tinta, conforme demonstrado, destacam-se alguns padrões distintivos. Notavelmente, ao examinar as variáveis *Laranja Escuro* e *Vermelho Carmim*, observam-se valores relativamente baixos de MAE e RMSE. Esses resultados sugerem que os modelos tiveram um desempenho mais consistente nesse cenário específico. Por outro lado, as variáveis *Branco* e *Ocre* apresentam-se como desafios consideráveis, evidenciados pelos valores elevados de MAE e RMSE registrados em múltiplos modelos. No entanto, é intrigante destacar o desempenho do modelo ANND, particularmente em relação a esses *targets* desafiadores. A capacidade do ANND de lidar eficazmente com categorias que apresentam dificuldades merece atenção especial, uma vez que evidencia a sua adaptação a cenários complexos e a sua habilidade em capturar nuances nos dados.

É relevante enfatizar que uma diferença substancial nos valores de MAE e RMSE já se manifesta ao comparar os modelos LR e MLP com as abordagens mencionadas anteriormente. Estes resultados apontam para uma redução significativa, que muitas vezes chega a metade dos erros anteriores. Esses benefícios podem ser atribuídos à estratégia de segmentação em conjuntos, a qual demonstra ser altamente benéfica nesse contexto.

Os resultados das métricas de desempenho dos modelos no *Cluster B* podem ser observados na tabela B.3.

Tabela B.3: Métricas de desempenho dos modelos no Conjunto de Teste do *Cluster B*

<i>Target</i>	<i>Cluster B</i> - Conjunto de teste									
	LR		MLP		ANND		RF		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brilho	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amarelo Vivo	1.36	3.86	1.65	4.14	1.53	4.92	0.39	2.37	0.49	2.73
Laranja Claro	1.44	3.52	1.83	2.84	1.35	4.00	0.41	1.46	0.54	1.61
Ocre	2.91	4.60	2.00	3.56	0.13	2.10	1.26	2.72	1.49	2.91
Amarelo Alaranjado	1.46	3.98	1.22	3.55	1.57	5.60	0.31	1.92	0.42	1.95
Laranja Escuro	0.34	1.23	0.30	1.23	0.27	1.53	0.11	0.81	0.16	0.85
Vermelho Carmim	0.22	0.74	0.16	0.75	0.15	0.88	0.14	0.64	0.17	0.67
Vermelho	0.90	1.91	1.11	2.08	0.03	0.04	0.46	1.48	0.55	1.53
Bordeaux	0.17	0.81	0.13	0.81	0.17	1.31	0.10	0.76	0.12	0.77
Castanho	0.10	0.44	0.08	0.44	0.05	0.21	0.04	0.37	0.08	0.39
Azul	0.59	1.50	0.90	1.86	0.37	0.94	0.22	0.94	0.27	1.02
Verde	0.62	1.41	0.90	1.68	0.46	1.71	0.16	0.77	0.18	0.80
Verde Claro	0.11	0.72	0.11	0.72	0.05	0.68	0.03	0.26	0.04	0.38
Branco	4.47	6.36	4.22	6.11	0.70	9.34	2.20	4.47	2.46	4.56
Preto	0.55	1.17	0.69	1.22	0.52	1.14	0.26	0.93	0.33	0.97
Total	15.25	32.25	15.27	30.99	7.35	34.40	6.09	19.89	7.29	21.12

Assim como ocorreu no *Cluster A*, a análise dos resultados no *Cluster B*, quando examinados por atributo de tinta, revela uma interessante variação no desempenho

dos modelos. Os *targets Extender Mate* e *Extender Brilho* demonstram erros de previsão extremamente baixos, sugerindo que os modelos foram capazes de se adaptar perfeitamente a esses *targets*. Além disso foram identificadas dificuldades similares às observadas no *Cluster A*, especialmente nas bases *Branco* e *Ocre*, nas quais o modelo ANND também apresentou um desempenho superior.

Assim, destaca-se o modelo RF como particularmente exemplar no contexto do *Cluster B*, conforme evidenciado na figura B.2, uma vez que na avaliação global apresentou valores de MAE e RMSE menores. Isso ressalta a sua proficiência em compreender e interpretar as complexas relações entre os parâmetros de cor e as concentrações de tinta. Esse feito desempenha um papel significativo na geração de previsões excepcionais, as quais, por sua vez, fundamentam ações e decisões cruciais no âmbito da afinação de tintas.

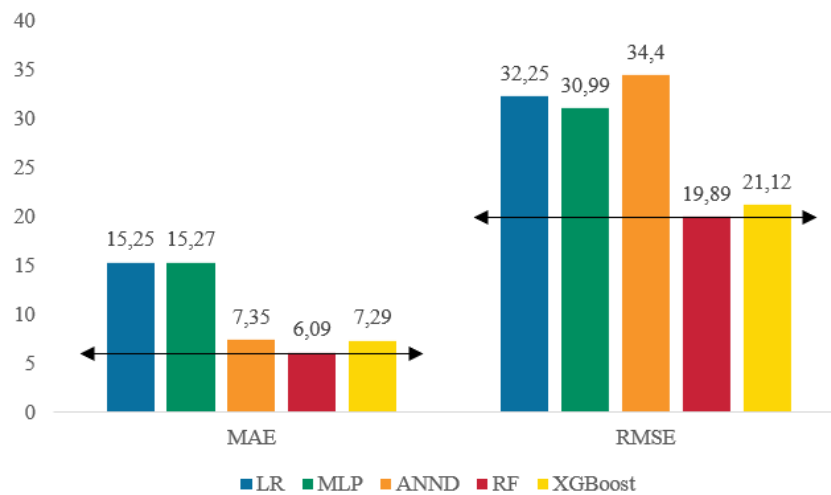


Figura B.2: Comparação dos resultados do desempenho dos modelos no *Cluster B*

Na tabela B.4, é possível visualizar os resultados das métricas de desempenho dos modelos no contexto do *Cluster C*.

A análise dos resultados dos modelos aplicados ao *Cluster C*, conforme apresentados na tabela B.4, proporciona *insights* valiosos sobre o desempenho dos modelos de previsão e suas capacidades em lidar com a complexidade das diversas características presentes nesse conjunto de dados.

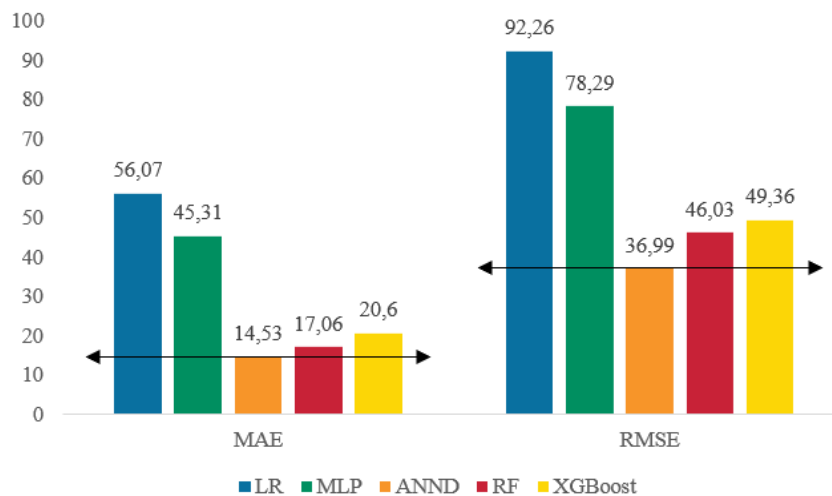
No contexto do *Cluster C*, identifica-se um cenário mais desafiador em relação aos *targets* de previsão, especialmente nas categorias *Ocre*, *Vermelho*, *Bordeaux*, *Castanho*, *Azul*, *Verde*, *Branco* e *Preto*. Nestes atributos, os valores de MAE e RMSE apresentam uma clara elevação em comparação com os registos anteriores, que podem estar relacionados com as características mencionadas na subsecção 3.4.2, onde se detetou uma dispersão expressiva e assimétrica nos dados deste grupo.

Ao analisar essas métricas, o modelo ANND emerge como o melhor modelo globalmente, demonstrando uma capacidade perceptível em fazer previsões em diversos

Tabela B.4: Métricas de desempenho dos modelos no Conjunto de Teste do *Cluster C*

<i>Target</i>	<i>Cluster C - Conjunto de teste</i>									
	LR		MLP		ANND		RF		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.00	0.00	0.00	0.00	0.08	0.48	0.02	0.53	0.04	0.62
Brilho	0.00	0.00	0.00	0.00	0.38	1.33	0.03	0.57	0.04	0.62
Amarelo Vivo	1.22	3.58	0.71	2.49	0.37	2.19	0.32	1.80	0.48	2.07
Laranja Claro	3.39	6.31	2.98	6.03	0.82	2.51	0.96	3.15	1.25	3.56
Ocre	9.20	11.98	6.72	9.35	1.80	3.30	2.40	4.77	2.82	5.05
Amarelo Alaranjado	0.54	2.29	0.40	2.27	0.26	2.36	0.26	1.94	0.34	1.99
Laranja Escuro	3.26	6.21	2.15	4.47	0.79	2.41	0.76	2.54	1.02	2.98
Vermelho Carmim	3.22	6.43	2.71	5.94	0.55	1.94	0.85	2.65	1.06	2.82
Vermelho	7.41	10.64	7.52	10.12	1.42	2.61	2.73	5.39	3.15	5.57
Bordeaux	4.01	7.81	5.54	9.53	0.93	2.52	1.25	4.02	1.51	4.18
Castanho	2.28	4.58	3.37	5.30	1.24	3.37	0.83	2.55	0.98	2.66
Azul	6.41	8.84	2.86	4.77	1.81	3.81	1.36	3.48	1.59	3.67
Verde	3.40	6.04	1.54	3.35	0.34	0.91	0.64	2.36	0.75	2.58
Verde Claro	0.09	1.21	0.08	1.21	0.03	1.21	0.06	1.26	0.09	1.28
Branco	6.93	9.15	6.08	8.30	2.57	3.96	3.23	5.56	3.81	6.07
Preto	4.71	7.17	2.65	5.16	1.12	2.06	1.35	3.45	1.66	3.62
Total	56.07	92.26	45.31	78.29	14.53	36.99	17.06	46.03	20.60	49.36

atributos. O ANND conseguiu manter consistentemente valores de MAE e RMSE mais baixos (14.53 e 36.99, respectivamente), como evidenciado na figura B.3. Esse desempenho permite que o modelo ANND supere os demais em uma ampla gama de *targets* dentro do grupo. Isso pode ser atribuído à natureza intrincada da sua arquitetura de rede neuronal, no caso uma MLP, que tem a capacidade de aprender representações latentes bastantes complexas, que também tinha sido *target* de destaque nos agrupamentos anteriores.

Figura B.3: Comparação dos resultados do desempenho dos modelos no *Cluster C*

Já a tabela B.5 permite analisar os resultados das métricas de desempenho dos modelos no contexto do *Cluster D*.

Tabela B.5: Métricas de desempenho dos modelos no Conjunto de Teste do *Cluster D*

<i>Target</i>	<i>Cluster D - Conjunto de teste</i>									
	LR		MLP		ANND		RF		XGBoost	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mate	0.04	0.76	0.05	0.76	0.06	1.39	0.03	0.78	0.05	0.76
Brilho	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Amarelo Vivo	1.62	4.46	2.47	5.47	1.00	4.90	0.56	2.95	0.83	3.23
Laranja Claro	4.18	7.64	3.88	7.41	2.75	5.48	1.19	3.86	1.41	4.26
Ocre	10.25	13.57	7.54	10.79	0.30	2.07	2.42	5.45	2.89	5.87
Amarelo Alaranjado	0.74	2.58	0.74	2.85	0.38	2.34	0.39	2.48	0.50	2.38
Laranja Escuro	1.85	4.26	1.82	4.34	1.31	4.09	0.64	2.43	0.91	2.77
Vermelho Carmim	3.49	6.50	3.70	6.64	2.19	7.15	0.74	2.26	1.05	2.99
Vermelho	6.57	9.38	6.12	8.74	0.44	1.99	2.11	4.56	2.70	5.11
Bordeaux	3.78	7.59	4.80	8.54	2.66	8.03	0.79	2.76	1.09	3.04
Castanho	2.15	5.16	1.82	5.16	1.31	4.80	0.58	2.26	0.81	2.43
Azul	8.41	11.17	3.82	6.83	5.29	8.77	1.55	4.09	1.80	4.19
Verde	3.55	6.32	2.85	6.14	2.78	6.84	0.86	2.61	1.09	2.77
Verde Claro	0.24	1.71	0.25	1.73	0.11	1.53	0.20	1.78	0.20	1.70
Branco	6.01	8.22	6.04	8.19	0.28	4.15	2.52	4.76	3.12	5.06
Preto	4.89	7.15	4.22	5.96	4.33	7.21	1.10	2.78	1.46	2.99
Total	57.79	96.48	50.12	89.55	25.19	70.74	15.68	45.79	19.92	49.56

Dirigindo o foco para a última linha da tabela, surge uma variação no desempenho agregado dos modelos no *Cluster D*, expressa por meio das métricas MAE e RMSE. A análise dos resultados nesse contexto ratifica a superioridade do modelo RF em relação aos restantes modelos, conforme evidenciado pela figura B.4.

A variabilidade no desempenho entre os atributos sugere a possibilidade de que os modelos enfrentem dificuldades na compreensão de padrões específicos. As categorias como *Ocre*, *Vermelho* e *Branco* demonstram um grau substancial de complexidade, refletido nos valores superiores de MAE e RMSE. Esses resultados sugerem que os modelos encontraram desafios ao tentar prever com precisão esses *targets* mais complexos, onde, mais uma vez, o ANND apresentou melhor desempenho.

Assim como observado no *Cluster C*, é notável que os valores de MAE e RMSE no presente contexto são significativamente elevados, apresentando também uma dispersão expressiva e assimétrica nos dados, tendo sido mencionado na subsecção 3.4.2.

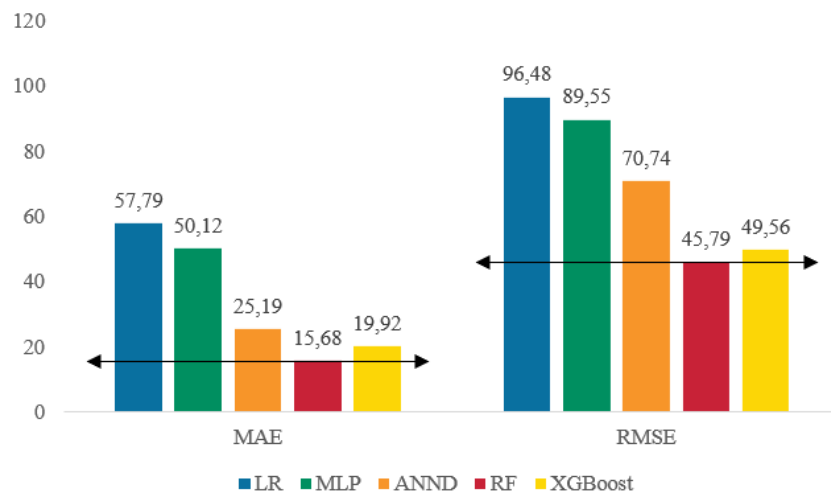


Figura B.4: Comparação dos resultados do desempenho dos modelos no *Cluster D*

## Apêndice C

# Complemento às Considerações Finais

Nesta secção, são apresentados os relatórios dos testes desenvolvidos nas cores cinza, azul e amarelo. Estes relatórios compreendem diversos elementos visuais e métricas para avaliar a diferença de cor entre a amostra e a referência.

Os gráficos incluídos nos relatórios exibem representações visuais das amostras em comparação com as referências, permitindo uma avaliação clara das discrepâncias de cor. Além disso, os gráficos de  $\Delta E$  são apresentados, destacando numericamente a diferença de cor. Essas métricas são fundamentais para compreender o quanto as cores divergem e a magnitude das variações percebidas.

Os relatórios também englobam a questão da estabilidade das cores em diferentes cenários de observação e iluminação, apresentando gráficos que demonstram o grau de metamerismo. Esta análise é fundamental para avaliar a consistência das cores.

Pelos Lotes de Laboratório das coordenadas L, a e b, manifesta-se a diferença de cor no plano  $Oab$  e, separadamente no eixo de luminosidade, ajudando a identificar quais componentes de cor são os mais afetados.

Por último, mas não menos importante, os relatórios apresentam as variações em cada coordenada de cor entre a referência e a amostra em vários sistemas de cor.

No geral, os relatórios desta secção fornecem uma visão abrangente e detalhada das análises visuais e métricas, sendo essenciais para avaliar a precisão das formulações de tintas e garantir que atendam às especificações de cor desejadas.

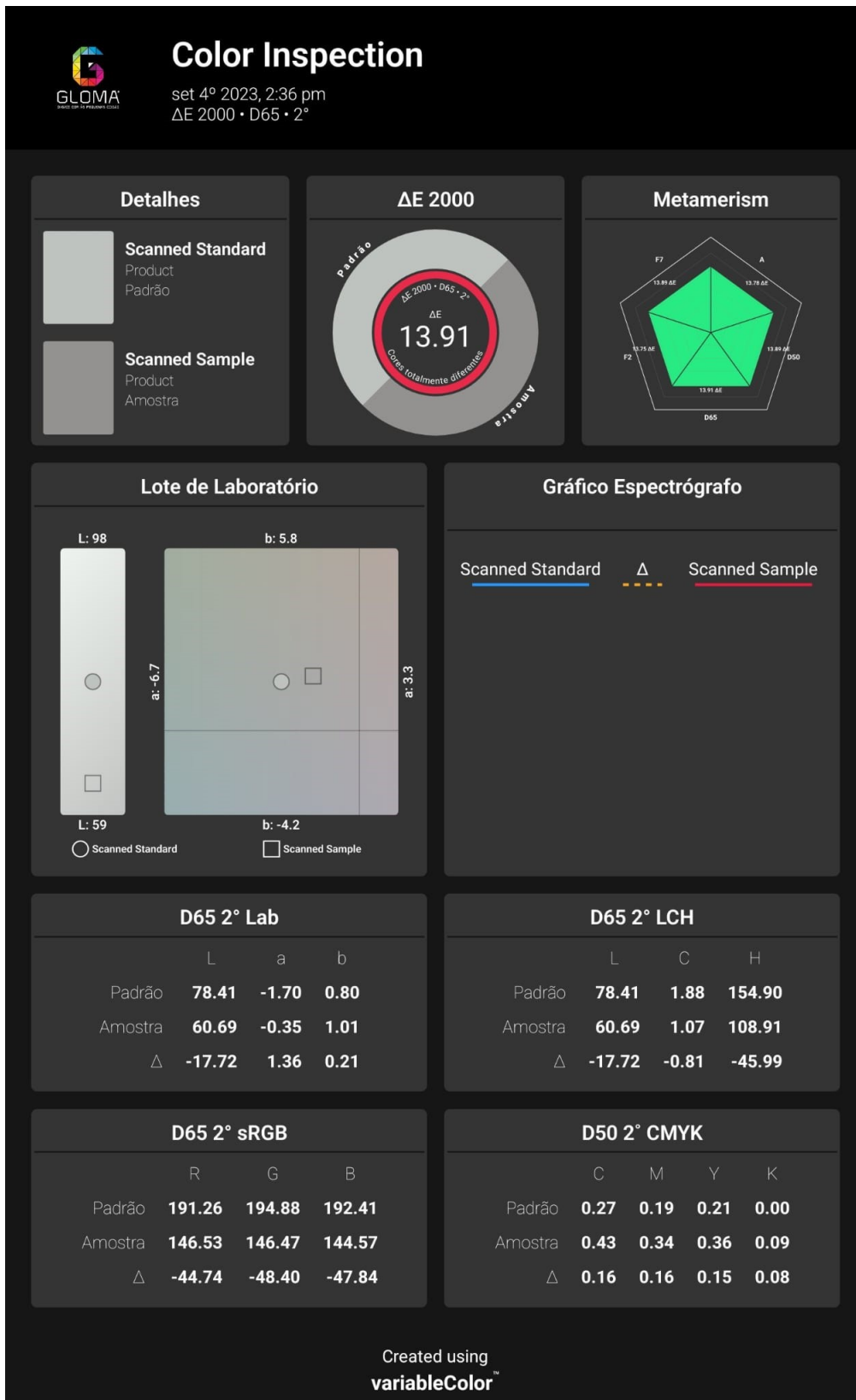


Figura C.1: Relatório do teste Cinza

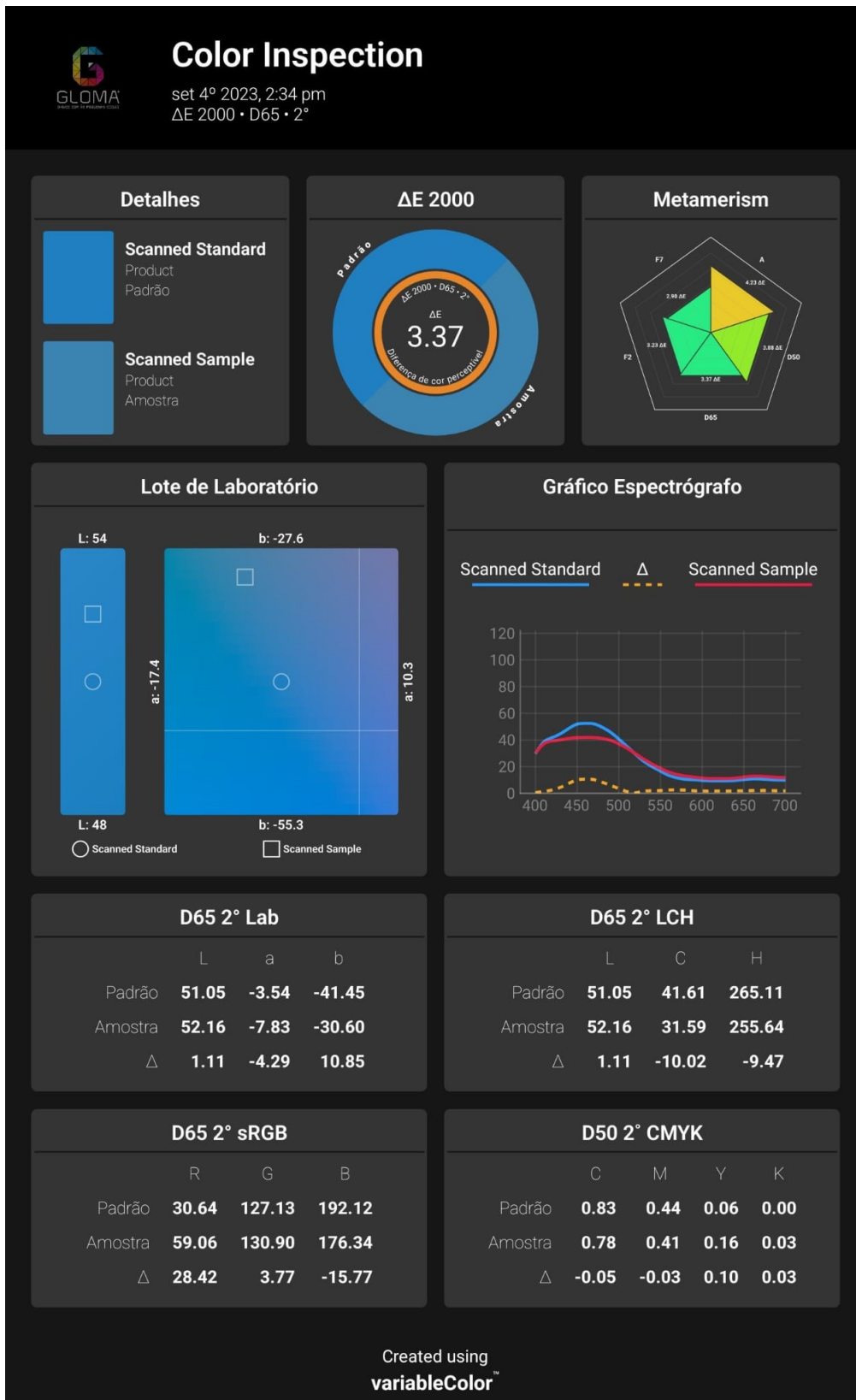


Figura C.2: Relatório do teste Azul

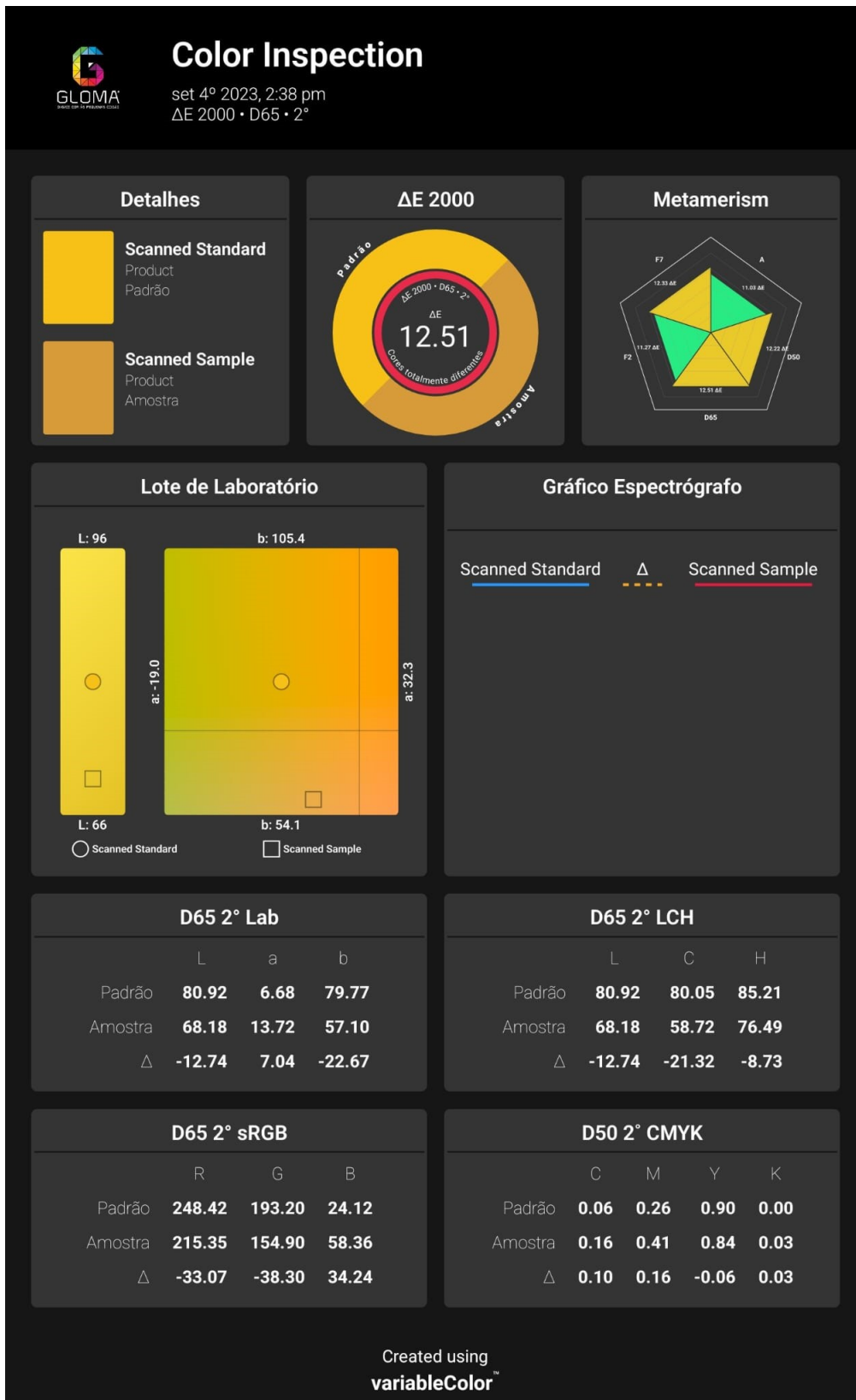


Figura C.3: Relatório do teste Amarelo