UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA QUÍMICA

LARISSA ALVILINO FERREIRA DA SILVA

MODELAGEM PREDITIVA DA ESTABILIDADE DE FORMULAÇÕES COSMÉTICAS MULTICOMPONENTES: UM ESTUDO COMPARATIVO DE MODELOS DE APRENDIZADO DE MÁQUINA

UBERLÂNDIA - MG
2026

LARISSA ALVILINO FERREIRA DA SILVA

MODELAGEM PREDITIVA DA ESTABILIDADE DE FORMULAÇÕES COSMÉTICAS MULTICOMPONENTES: UM ESTUDO COMPARATIVO DE MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia Química da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de bacharel em Engenharia Química.

Orientadora: Prof.ª Dr.ª Sarah Arvelos Altino

UBERLÂNDIA - MG

2026

AGRADECIMENTOS

Agradeço, antes de tudo, a Deus, por ter sido a força que me sustentou ao longo desta caminhada, e por ter me guiado até aqui.

À minha família, em especial aos meus pais e à minha irmã, minha eterna gratidão. Vocês foram o alicerce da minha trajetória, e a paciência e o apoio incondicional de vocês foram essenciais para que eu pudesse concluir esta etapa tão importante da minha formação.

Ao Victor, meu grande amor, que a engenharia química colocou em meu caminho ainda no primeiro estágio, agradeço por ser meu maior incentivador e minha maior fonte de inspiração. Obrigada por acreditar em mim, mesmo quando eu duvidava de mim mesma.

Agradeço às minhas queridas amigas Laura e Ana Beatriz, que me acompanham desde a infância e que sempre acreditaram em mim, muitas vezes mais do que eu mesma, minha profunda gratidão pelo apoio constante. Às amizades construídas ao longo da graduação, Isabela e Maria Gabriela, que sabem tão bem o valor de cada conquista alcançada nessa trajetória. Obrigada a todas vocês, por caminharem ao meu lado, tornando essa jornada mais leve e mais feliz.

Por fim, agradeço à Universidade Federal de Uberlândia e ao corpo docente, pelos anos de aprendizado e crescimento que tornaram possível a conclusão desta importante etapa da minha formação acadêmica. Agradeço, em especial, à minha orientadora, Sarah, pela dedicação, paciência e conhecimento compartilhados ao longo do desenvolvimento deste trabalho, sem os quais este resultado não seria possível. Agradeço também à banca presente, pela disponibilidade e pelas contribuições a este trabalho.

Houve muitos momentos, ao longo dessa trajetória, em que desejei apenas que tudo passasse logo, que o cansaço e as dificuldades chegassem ao fim o quanto antes. Hoje, porém, percebo que estou exatamente onde sempre quis estar, cercada pelas melhores pessoas que eu poderia ter ao meu lado. A todos vocês, minha mais sincera gratidão.

RESUMO

A estabilidade físico-química é um atributo crítico no desenvolvimento de formulações cosméticas multicomponentes, e sua avaliação depende, tradicionalmente, de ensaios empíricos demorados. Neste contexto, o presente trabalho teve como objetivo comparar o desempenho de modelos de aprendizado de máquina baseados em árvores, como os de Árvore de Decisão, Floresta Aleatória e XGBoost, na classificação da estabilidade de formulações de *shampoo*, utilizando as frações mássicas dos ingredientes como variáveis preditoras. Como base de dados, empregou-se o conjunto experimental de 812 formulações disponibilizado na literatura, do qual se selecionou a estabilidade como variável-alvo binária (estável ou não-estável). A metodologia envolveu a preparação e a análise exploratória dos dados, seguida de divisão estratificada em treino e teste, validação cruzada e otimização de hiperparâmetros por busca em grade (*GridSearchCV*). A análise exploratória evidenciou elevada esparsidade, forte assimetria e ausência de separabilidade linear entre as classes, o que justificou o emprego de modelos não lineares. O desempenho foi avaliado por meio das métricas F1-*score*, acurácia, precisão e *recall*, além da análise dos erros por matrizes de confusão. Observou-se uma progressão consistente do desempenho com o aumento da complexidade dos modelos: a Árvore de Decisão, usada como referência, obteve F1-*score* de 0,66 e acurácia de 68%, enquanto os modelos de conjunto *Random Forest* e XGBoost elevaram essas métricas de forma equivalente, com F1-*score* de teste de 0,71 e acurácia de aproximadamente 77%. Diante da equivalência estatística entre os dois modelos no conjunto de teste, adotou-se como critério de seleção o desempenho médio na validação cruzada, no qual o *Random Forest* apresentou o melhor resultado (F1-*score* de 0,70), sendo selecionado como modelo final. A análise de importância das variáveis indicou os espessantes e os polímeros condicionantes como os ingredientes mais influentes sobre a estabilidade, em consonância com o caráter combinatório do sistema. Os resultados ilustram o potencial de modelos preditivos baseados em árvores como ferramenta de apoio ao desenvolvimento de formulações cosméticas mais ágil e orientado por dados.

Palavras-chave: modelagem preditiva; árvores de decisão; formulações cosméticas; estabilidade; aprendizado de máquina.

ABSTRACT

Physicochemical stability is a critical attribute in the development of multicomponent cosmetic formulations, and its assessment traditionally relies on time-consuming empirical tests. In this context, the present study aimed to compare the performance of tree-based machine learning models, such as Decision Tree, Random Forest and XGBoost, in the classification of shampoo formulation stability, using ingredient mass fractions as predictor variables. As a database, the experimental dataset of 812 formulations available in the literature was employed, from which stability was selected as the binary target variable (stable or unstable). The methodology involved data preparation and exploratory data analysis, followed by stratified splitting into train and test sets, cross-validation and hyperparameter optimization through grid search (GridSearchCV). The exploratory analysis revealed high sparsity, strong asymmetry and the absence of linear separability between classes, which justified the use of non-linear models. Performance was evaluated using F1-score, accuracy, precision and recall metrics, as well as error analysis through confusion matrices. A consistent progression in performance was observed with increasing model complexity: the Decision Tree, used as a reference, achieved an F1-score of 0.66 and accuracy of 68%, while the ensemble models, Random Forest and XGBoost, improved these metrics to an equivalent extent, reaching a test F1-score of 0.71 and accuracy of approximately 77%. Given the statistical equivalence between the two ensemble models on the test set, the mean cross-validation performance was adopted as the selection criterion, in which Random Forest achieved the best result (F1-score of 0.70), being selected as the final model. Feature importance analysis indicated thickeners and conditioning polymers as the most influential ingredients on stability, consistent with the combinatorial nature of the system. The results illustrate the potential of tree-based predictive models as a tool to support more agile and data-driven cosmetic formulation development.

Keywords: predictive modeling; decision trees; cosmetic formulations; stability; machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Atualmente, a indústria de cosméticos e cuidados pessoais exerce papel central na economia do Brasil e no mercado global de beleza, o que realça a relevância de trabalhos dedicados à formulação e estabilidade de produtos multicomponentes. Dados divulgados pelo Panorama do Setor de Beleza e Cuidados Pessoais indicam que, em 2024, o Brasil consolidou-se como o 3º maior mercado consumidor mundial de produtos de higiene, perfumaria e cosméticos, e o 4º mercado no ranking global de países que mais lançam produtos anualmente, atrás apenas dos Estados Unidos, China e Índia, segundo a Associação Brasileira da Indústria de Higiene Pessoal, Perfumaria e Cosméticos (ABIHPEC, 2025).

Esse cenário de mercado global e nacional reforça a importância de abordagens eficientes para o desenvolvimento de formulações cosméticas, especialmente considerando a complexidade de formulações multicomponentes. Tal complexidade está diretamente relacionada ao desafio de garantir a estabilidade físico-química dos produtos ao longo do tempo. As emulsões, amplamente empregadas em cosméticos, são sistemas termodinamicamente instáveis que devem demonstrar resiliência e consistência adequada, mantendo a integridade da sua superfície interfásica mesmo após serem expostas a tensões consecutivas de fatores como temperatura, agitação e aceleração da gravidade (BARZOTTO et al., 2009).

O estudo da estabilidade, tipicamente realizado por meio de análises aceleradas de curto e longo prazo, fornece informações fundamentais sobre o comportamento do produto e sua vida útil. O desafio é agravado pela necessidade de avaliar diferentes tipos de estabilidade (física e química) e suas origens, que podem ser intrínsecas (inerentes à composição) ou extrínsecas (fatores ambientais ou de processamento), sendo a instabilidade a principal causa de falhas e descarte de lotes (BARZOTTO et al., 2009).

Nesse contexto, a aplicação de técnicas de aprendizado de máquina para modelagem preditiva da estabilidade de formulações cosméticas pode trazer ganhos significativos, como redução de custos, aceleração de desenvolvimento e menor dependência de longos testes empíricos.

O conjunto de dados que serve de base a este trabalho foi disponibilizado por Chitre et al. (2024), no estudo *Accelerating formulation design* via *machine learning*, que aplicou uma abordagem experimental de alto rendimento (*high-throughput*) para gerar sistematicamente formulações de *shampoo*. O propósito dos autores foi construir uma base ampla e padronizada, capaz de subsidiar o uso de aprendizado de máquina na aceleração do desenvolvimento de produtos cosméticos. O *dataset* reúne 812 formulações e registra, para cada amostra, diferentes

propriedades físico-químicas, como reologia, turbidez, pH e estabilidade, o que o torna adequado à modelagem preditiva de múltiplos atributos de interesse.

Dentre essas propriedades, o presente trabalho concentra-se na estabilidade das formulações, adotada como atributo-alvo. Dessa forma, a base de Chitre et al. (2024) é aqui utilizada como estudo de caso de escopo acadêmico, coerente com a proposta de um trabalho de conclusão de curso, embora o conjunto de dados comporte análises mais abrangentes, opta-se por um recorte específico, cujo detalhamento é apresentado a seguir.

2. FUNDAMENTOS TEÓRICOS

2.1.Aprendizado de máquina supervisionado

De modo geral, por meio do reconhecimento de padrões e da aprendizagem a partir de uma base de dados (FACELI et al., 2011), o aprendizado de máquina permite que algoritmos e modelos estatísticos melhorem progressivamente o desempenho de uma tarefa, tornando-se aptos a realizar previsões.

Usualmente, um conjunto de exemplos é dividido em dois subconjuntos disjuntos, o conjunto de treinamento que é usado para o aprendizado do conceito e o conjunto de teste usado para medir o grau de efetividade do conceito aprendido. O treinamento é baseado na comparação entre o resultado obtido do modelo e a variável previamente classificada, processo esse que é repetido até obter-se um erro mínimo. Já o conjunto de teste é composto por dados que não foram apresentados durante o treinamento, o que permite verificar a capacidade de generalização do modelo. Na fase de avaliação, o modelo é comparado com os dados de teste e os resultados são gerados (SATHYA; ABRAHAM, 2013) e para que seja estatisticamente válida, é necessário que os dois subconjuntos sejam disjuntos, garantindo que o desempenho observado no teste reflita o comportamento real do modelo diante de dados desconhecidos.

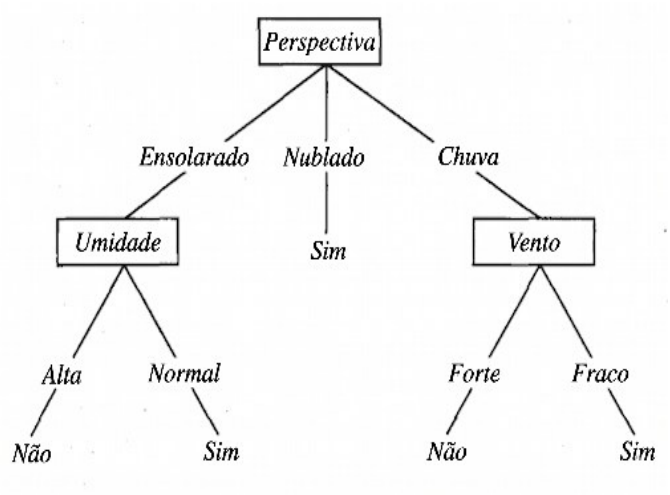
Quando as classes do problema não se distribuem de forma equilibrada, essa divisão pode ser conduzida de forma estratificada. A estratificação dispõe os dados de maneira que cada partição reproduza, em pequena escala, a proporção do conjunto original (TAN; STEINBACH; KUMAR, 2009).

2.2.Árvores de Decisão

Uma mesma base de dados pode ser analisada sob a perspectiva de diferentes modelos de aprendizado de máquina. Entre as décadas de 1960 e 1980, período em que diversos algoritmos foram desenvolvidos, surgiram também as Árvores de Decisão (LOH, 2014), que representam uma série de escolhas de acordo com os dados analisados. Este modelo de aprendizado de máquina supervisionado representa regras de decisão a partir da construção de uma estrutura em forma de árvore. Por meio dela, é possível lidar com dados categóricos e numéricos, além de problemas de classificação e regressão, sendo também bastante útil para interpretação de resultados (MITCHELL, 1997). Em situações que exigem a análise e entendimento de processos, constitui um recurso valioso para a tomada de decisão.

A Figura 1 é um exemplo de Árvore de Decisão para o conceito de *PlayTennis*, ou seja, uma partida de tênis. Nela, uma instância é classificada ao percorrer a árvore a partir do nó raiz até o nó-folha apropriado, que retorna a classificação a ele associada (neste caso, sim ou não). A interpretação do resultado é exemplificada pela classificação de uma manhã de sábado como adequada ou não para se jogar tênis (MITCHELL, 1997).

Figura 1. Representação da árvore de decisões para o conceito de uma partida de tênis.



Fonte: Mitchell, 1997. Adaptado pela autora (2026).

2.3.Métodos de conjunto: *Random Forest* e XGBoost

Entre os métodos de conjunto (*ensemble*) derivados das Árvores de Decisão, um dos mais conhecidos é o *Random Forest* (RF, Floresta Aleatória), utilizado no desenvolvimento de modelos de predição. Introduzido por Breiman em 2001 (Breiman, 2001) é um método do tipo *bagging* (do inglês, *Bootstrap Aggregating*), capaz de gerar múltiplas amostras do conjunto de dados original, por meio de reamostragem com reposição (CHEN et al., 2020). Ao construir várias Árvores de Decisão em paralelo, cada uma delas é treinada com uma subamostra aleatória dos dados e dos atributos (*features*), e a predição final é feita por votação majoritária entre os modelos individuais. Com isso, a abordagem reduz a variância do modelo e diminui a chance de sobreajuste em comparação com a árvore individual (BREIMAN, 2001).

O XGBoost, por sua vez, é um método do tipo *boosting*, capaz de realizar a otimização das métricas de forma sequencial (BRUCE, 2017). Nessa abordagem, cada árvore é construída para corrigir os erros residuais da anterior, empregando o gradiente descendente para minimizar a função de perda (CHEN; GUESTRIN, 2016).

3. OBJETIVOS

3.1. Geral

Comparar o desempenho de modelos de aprendizado de máquina baseados em árvores de decisão - como Árvore de Decisão, Floresta Aleatória e XGBoost - na classificação da estabilidade de formulações cosméticas multicomponentes, tendo como variáveis preditoras as frações mássicas dos ingredientes e utilizando o conjunto de dados experimental de formulações de *shampoo* disponibilizado por Chitre et al. (2024).

3.2.Específicos

a) Preparar e organizar o conjunto de dados experimental, selecionando as frações mássicas dos ingredientes como variáveis preditoras e definindo a variável-alvo binária de estabilidade (*Stability_Test*), com verificação de valores faltantes e da coerência composicional das formulações;

b) Caracterizar o conjunto de dados por meio de análise exploratória, tais como estatísticas descritivas, matrizes de correlação (Pearson, Spearman e Kendall) e gráficos de distribuição (*pair plots* e *violin plots*), evidenciando a esparsidade, a assimetria e a natureza não linear das relações entre as variáveis;

c) Definir os conjuntos de treino e teste por divisão estratificada com validação cruzada (*k-fold*), preservando a proporção entre formulações estáveis e instáveis e tratando o desbalanceamento entre as classes;

d) Treinar e otimizar os hiperparâmetros dos três modelos baseados em árvores (Árvore de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*) e XGBoost (*Extreme Gradient Boosting*)) por meio de busca em grade (*GridSearchCV*);

e) Avaliar e comparar o desempenho preditivo e a capacidade de generalização dos modelos, empregando as métricas F1-*score*, acurácia, precisão e *recall*, além da análise dos erros (falsos positivos e falsos negativos) por meio de matrizes de confusão;

f) Interpretar a relevância físico-química das variáveis mais influentes na estabilidade, relacionando os ingredientes de maior importância aos mecanismos característicos de sistemas surfactantes.

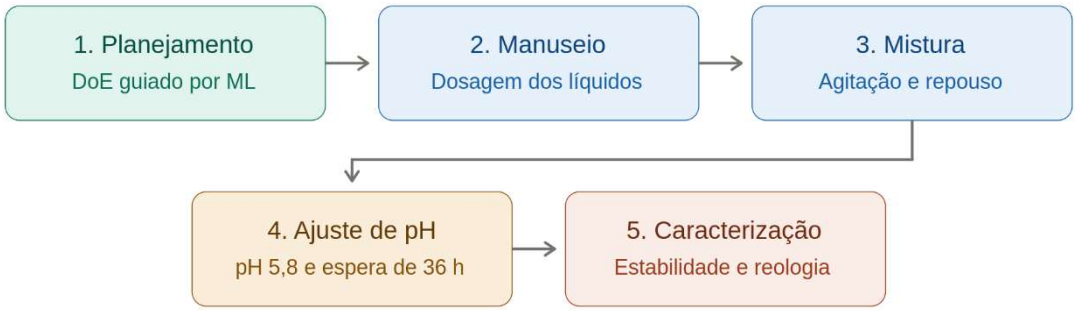
4. METODOLOGIA

Todo o desenvolvimento computacional deste trabalho foi realizado no ambiente Google Colab, utilizando a linguagem Python e bibliotecas amplamente empregadas em ciência de dados. A manipulação e a preparação dos dados foram conduzidas com as bibliotecas *pandas* e *NumPy* (MCKINNEY, 2010; HARRIS et al., 2020), a visualização e a análise exploratória, com *Matplotlib* e *Seaborn* (HUNTER, 2007; WASKOM, 2021), e a construção, o treinamento e a avaliação dos modelos, com *Scikit-learn* e *XGBoost* (PEDREGOSA et al., 2011; CHEN; GUESTRIN, 2016). A escolha do Colab justifica-se pelo acesso facilitado a recursos computacionais, pela reprodutibilidade e pela integração com armazenamento em nuvem.

4.1.Origem e preparação do conjunto de dados

O conjunto de dados utilizado foi disponibilizado por Chitre et al. (2024) e reúne formulações de *shampoo* descritas pelas frações mássicas (% m/m) de seus ingredientes. Os dados foram gerados por um fluxo de trabalho experimental semiautomatizado, sistematizado na Figura 2, que compreende cinco etapas: o planejamento das formulações por um projeto de experimentos (*Design of Experiments*, DoE) guiado por aprendizado de máquina, o manuseio e a dosagem automatizados dos líquidos, a mistura, o ajuste de pH, e a caracterização das amostras. Ao longo desse processo, cada formulação foi avaliada quanto à estabilidade de fases, à turbidez e à reologia, o que resultou em uma base padronizada e adequada à modelagem preditiva.

Figura 2. Fluxograma do *workflow* experimental de geração do *dataset.*



Fonte: Adaptado de Chitre et al. (2024).

Neste trabalho de conclusão de curso, a propriedade de interesse é a estabilidade das formulações. Uma formulação foi definida como estável quando se apresentou como uma fase única e homogênea, sem separação de fases ou floculação, caso contrário, foi classificada como instável. A Figura 3 ilustra exemplos de amostras rotuladas como estáveis (*stable*) e instáveis (*unstable*). Cabe destacar que os autores utilizaram essas imagens para treinar um modelo de visão computacional dedicado à classificação da estabilidade a partir de fotografias das amostras, tarefa distinta da abordada nesta pesquisa. No presente trabalho, a estabilidade é modelada a partir da composição das formulações, isto é, das frações mássicas dos ingredientes, e não de suas imagens.

Figura 3. Exemplos de formulações classificadas como estáveis (*Stable*) e instáveis (*Unstable*). As marcações em vermelho correspondem à região identificada pelo detector de objetos empregado pelos autores.

Fonte: Chitre et al. (2024).

A partir da base original, foram selecionadas as variáveis correspondentes às frações mássicas dos ingredientes, adotadas como variáveis preditoras, e a variável *Stability_Test*, de natureza binária, adotada como alvo da classificação, indicando se a formulação permaneceu estável. No presente trabalho de conclusão de curso, foram conduzidas operações de inspeção do conjunto, como a verificação de valores faltantes ou inconsistentes, a confirmação da coerência composicional das formulações (soma das frações mássicas) e a análise da distribuição das classes.

4.2.Análise exploratória dos dados

Antes da etapa de modelagem, o conjunto de dados foi caracterizado por meio de análise exploratória, com o objetivo de compreender a distribuição das variáveis e as relações entre elas. Foram calculadas estatísticas descritivas (média, desvio padrão, mínimo, mediana e máximo) das frações mássicas dos ingredientes e analisada a distribuição da variável-alvo. As relações entre variáveis foram investigadas por meio de gráficos de dispersão par a par (*pair*

plots) e de gráficos de violino (*violin plots*) (HINTZE; NELSON, 1998; WASKOM, 2021), que combinam informações de distribuição, mediana e dispersão segmentadas por classe. Adicionalmente, foram calculadas as matrizes de correlação de Pearson, Spearman e Kendall (FACELI et al., 2011) entre as concentrações dos ingredientes e a variável-alvo, de modo a avaliar tanto associações lineares quanto não paramétricas e a verificar a eventual presença de multicolinearidade entre as variáveis preditoras. Essa caracterização orientou a escolha dos modelos empregados nas etapas seguintes.

4.3.Divisão dos conjuntos de treino e teste

O conjunto de dados foi dividido em 80% para treinamento e 20% para teste, por meio de divisão estratificada, de forma a preservar, em ambas as partições, a proporção original entre formulações estáveis e instáveis. Durante o treinamento, empregou-se validação cruzada *k-fold*, com $k = 5$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), para avaliar a robustez dos modelos sob diferentes particionamentos dos dados. O desbalanceamento natural entre as classes foi considerado tanto na condução do treinamento quanto na escolha das métricas de avaliação.

4.4.Modelos de classificação e otimização de hiperparâmetros

Foram avaliados três modelos supervisionados baseados em árvores de decisão. A Árvore de Decisão foi adotada como modelo de referência (*baseline*) (MITCHELL, 1997; LOH, 2014; BREIMAN et al., 1984), por sua simplicidade e interpretabilidade. Em seguida, foram avaliados dois métodos de conjunto (*ensemble*): o *Random Forest* (BREIMAN, 2001), baseado na estratégia de *bagging*, e o XGBoost (CHEN; GUESTRIN, 2016), baseado na estratégia de *boosting*, ambos escolhidos por sua capacidade de capturar interações não lineares entre os ingredientes e de lidar com dados de alta dimensionalidade.

Os hiperparâmetros de cada modelo foram otimizados por meio de busca em grade (*GridSearchCV*) associada à validação cruzada de cinco partições, adotando-se como critério de seleção o desempenho médio na validação cruzada. Para a Árvore de Decisão, foram ajustados parâmetros como a profundidade máxima (*max_depth*), o número mínimo de amostras para divisão de um nó (*min_samples_split*), o número mínimo de amostras por folha (*min_samples_leaf*), o critério de divisão (*criterion*) e a estratégia de particionamento (*splitter*). Para os modelos de conjunto, foram ajustados, entre outros, o número de árvores (*n_estimators*)

e a profundidade máxima. A otimização foi conduzida em relação ao F1-*score* e à acurácia, permitindo avaliar o impacto da métrica de otimização sobre o desempenho dos modelos.

4.5.Métricas de avaliação e análise de erros

Após a otimização, os modelos foram avaliados no conjunto de teste, composto por dados não utilizados durante o treinamento, de modo a verificar a capacidade de generalização. O desempenho foi quantificado pelas métricas: acurácia, precisão, sensibilidade (*recall*) e F1-*score*, e os acertos e erros de classificação foram organizados em matrizes de confusão (FACELI et al., 2011). A análise dos erros concentrou-se nos falsos positivos e falsos negativos, discutindo-se suas implicações práticas para o desenvolvimento de formulações cosméticas. Detalhes sobre estas métricas e sobre a construção da matriz de confusão serão apresentados em conjunto com a seção de resultados.

4.6.Interpretação e relevância físico-química

Por fim, realizou-se a análise da importância relativa das variáveis (*feature importance*) dos modelos (BREIMAN, 2001), com o objetivo de identificar os ingredientes mais influentes na classificação da estabilidade. Os componentes de maior relevância foram discutidos à luz dos mecanismos físico-químicos característicos de sistemas surfactantes, relacionando os resultados da modelagem ao comportamento esperado das formulações.

5. RESULTADOS E DISCUSSÃO

5.1.Análise Exploratória De Dados

O *dataset* utilizado neste trabalho é proveniente do estudo de Chitre et al. (2024) composto por 812 formulações, obtidas a partir de um total de 822 amostras preparadas pelos autores, 10 foram excluídas por eles devido a dados de reologia ausentes, turbidez muito alta e por falhas no ajuste de pH. O *dataset* reúne dezoito ingredientes comerciais, provenientes do portfólio da BASF, distribuídos entre doze surfactantes aniônicos, não-iônicos, anfotéricos, catiônicos, quatro polímeros condicionantes e dois espessantes, conforme documentado no repositório público disponibilizado por Lapkin et al. (2024). No entanto, cada formulação é preparada com apenas quatro desses ingredientes: uma mistura binária de surfactantes, um polímero condicionante e um espessante em base aquosa. As concentrações de cada ingrediente, expressas em percentual mássico (% m/m), constituem as variáveis preditoras do modelo desenvolvido neste trabalho de conclusão de curso, sendo que os ingredientes não utilizados em cada formulação assumem valor zero (CHITRE et al., 2024).

Uma formulação é classificada como estável quando apresenta fase única homogênea, sem separação de fases ou floculações (CHITRE et al., 2024). Os autores avaliaram as formulações em dois momentos, primeiro após mistura (antes do ajuste de pH) e 36 horas após o ajuste de pH. Apenas as formulações estáveis nas duas inspeções foram registradas como estáveis no *dataset*.

As nomenclaturas das substâncias contidas no artigo referem-se aos nomes comerciais dos ingredientes, registrados pela BASF, e se definem de forma classificatória, como consta na Tabela 1. Adicionalmente, cada ingrediente é identificado por sua respectiva nomenclatura INCI (*International Nomenclature of Cosmetic Ingredients*), que constitui o padrão internacional para a designação de matérias-primas em formulações cosméticas, permitindo a padronização e comparação entre diferentes estudos.

Tabela 1. Família de Ingredientes no *dataset.*

Nome Comercial	**Nome INCI**	**Tipo Químico**	**Aplicação Principal**
Texapon® SB 3 KC	*Disodium Laureth Sulfosuccinate*	Surfactante aniônico	Limpeza principal
Plantapon® ACG 50	*Disodium Cocoyl Glutamate*	Surfactantes aniônico suaves	Limpeza suave
Plantapon® LC 7	*Laureth-7 Citrate*		
Plantacare® 818	*Coco-Glucoside*	Surfactantes não-iônico	Suavidade, baixa irritação
Plantacare® 2000	*Decyl-Glucoside*		
Dehyton® MC	*Sodium Cocoamphoacetate*	Surfactante anfotérico	Espuma, suavidade
Dehyton® PK 45	*Cocamidopropyl Betaine*		
Dehyton® ML	*Sodium Lauroamphoacetate*		
Dehyton® AB 30	*Coco-Betaine*		
Plantapon® Amino SCG-L	*Sodium Cocoyl Glutamate*	Surfactantes aniônicos (aminoácidos)	Limpeza suave
Plantapon® Amino KG-L	*Potassium Cocoyl Glycinate*		
Dehyquart® A-CA	*Cetrimonium Chloride*	Surfactante Catiônico	Condicionamento
Dehyquart® CC6	*Polyquaternium-6*	Polímero Condicionante	Condicionamento, viscosidade
Dehyquart® CC7 Benz	*Polyquaternium-7**		
Luviquat® Excellence	*Polyquaternium-16*		
Salcare® Super 7	*Polyquaternium-7**		
Arlypon® F	*Laureth-2*	Espessante	Viscosidade
Arlypon® TT	*PEG/PPG-120/10 Trimethylolpropane Trioleate (and) Laureth-2*		

**Nota: O Dehyquart® CC7 Benz e o Salcare® Super 7 compartilham a mesma nomenclatura INCI (Polyquaternium-7), porém são produtos distintos: o Dehyquart® CC7 Benz é composto por aproximadamente 54% de DADMAC e 46% de acrilamida (PM ≈ 500.000 g/mol), enquanto o Salcare® Super 7 apresenta composição invertida, com 75% de acrilamida e 25% de DADMAC (PM entre 100.000 e 200.000 g/mol) (LAPKIN et al., 2024).*

Fonte: Elaborado pela autora com dados de Lapkin et al. (2024).

Conforme apresentado na Tabela 1, os prefixos dos nomes comerciais indicam a classe química e a aplicação principal de cada ingrediente. Em relação aos sufixos numéricos e alfanuméricos, estes indicam informações complementares como o teor de matéria ativa. Por exemplo, Plantapon® ACG 50 contém aproximadamente 50% de matéria ativa, o tipo de cadeia

carbônica ou especificações de contra-íon. Informações adicionais, incluindo teor de matéria ativa, conteúdo de água e aditivos presentes em cada ingrediente, estão disponíveis no repositório público do estudo (CHITRE et al., 2024).

Na análise exploratória dos dados deste trabalho de conclusão de curso, uma das funções utilizadas foi a *DataFrame.describe()* da Python Pandas, no qual gera um resumo estatístico das colunas numéricas do *dataset*, auxiliando na compreensão da distribuição e consistência dos dados. A Tabela 2 apresenta essas estatísticas descritivas das concentrações dos ingredientes.

Tabela 2. Estatísticas descritivas das concentrações dos ingredientes do *dataset* (% m/m).

Ingrediente	**Média**	**Desvio Padrão**	**Mínimo**	**Mediana**	**Máximo**
Texapon SB 3 KC	1,09	2,90	0,00	0,00	16,24
Plantapon ACG 50	1,56	3,80	0,00	0,00	13,68
Plantapon LC 7	1,50	3,64	0,00	0,00	14,94
Plantacare 818	1,84	3,98	0,00	0,00	14,45
Plantacare 2000	1,85	4,04	0,00	0,00	13,99
Dehyton MC	1,64	3,84	0,00	0,00	13,74
Dehyton PK 45	1,84	4,09	0,00	0,00	13,30
Dehyton ML	1,54	3,62	0,00	0,00	14,39
Dehyton AB 30	2,04	4,29	0,00	0,00	13,52
Plantapon Amino SCG-L	1,74	4,00	0,00	0,00	13,44
Plantapon Amino KG-L	1,44	3,63	0,00	0,00	13,71
Dehyquart A-CA	2,31	4,49	0,00	0,00	13,36
Luviquat Excellence	0,71	1,01	0,00	0,00	3,08
Dehyquart CC6	0,51	1,04	0,00	0,00	4,86
Dehyquart CC7 Benz	0,42	0,91	0,00	0,00	4,02
Salcare Super 7	0,46	0,90	0,00	0,00	3,39
Arlypon F	1,32	1,71	0,00	0,00	5,27
Arlypon TT	1,63	1,69	0,00	1,50	5,64

Fonte: Elaborado pela autora com dados de Chitre et al. (2024).

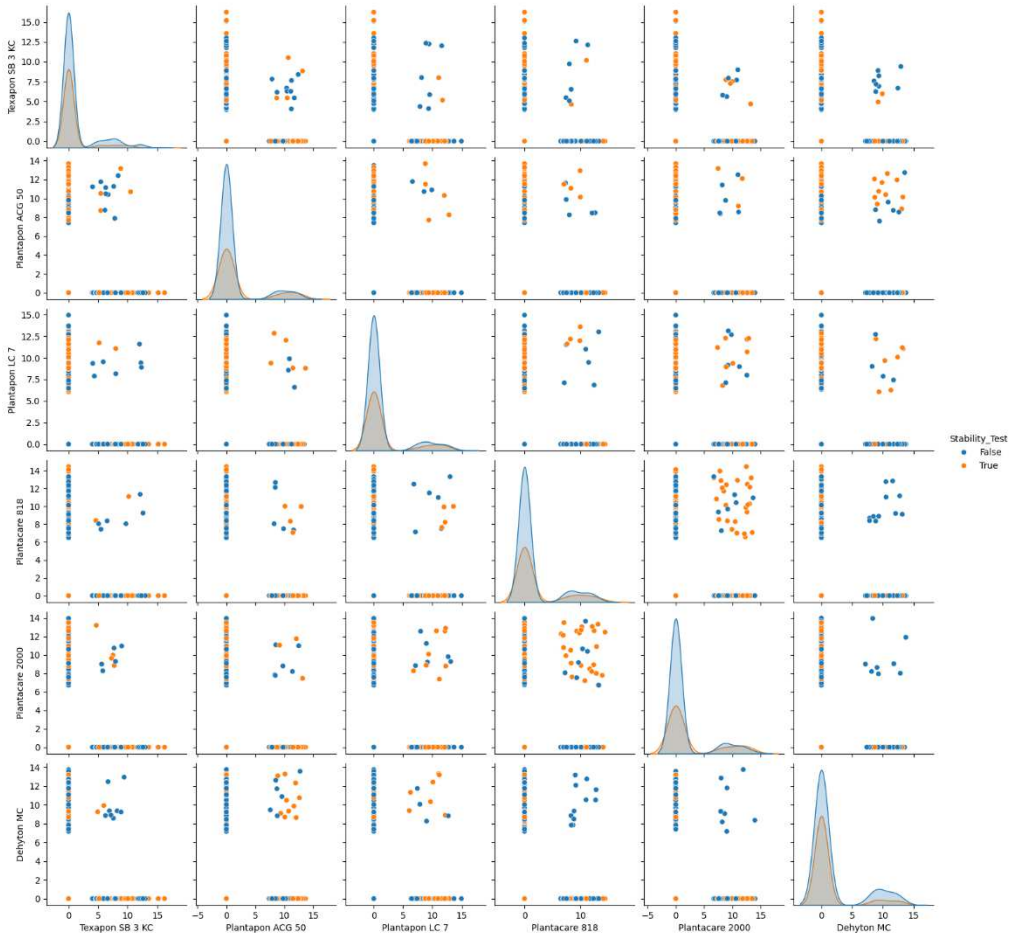
A partir dos dados observados na Tabela 2, foi possível fazer algumas considerações a respeito das concentrações dos ingredientes. Em relação aos valores de mínimo e mediana,

percebe-se que a maioria está com valores nulos, com exceção do Arlypon® TT, que tem 1,50 como mediana. Isso ocorre porque, como mencionado anteriormente, cada formulação usa apenas dois surfactantes dos doze disponíveis, os demais ingredientes não possuem concentração na formulação (CHITRE et al., 2024). Observa-se também as baixas médias, que variam apenas entre 0,42% (Dehyquart CC7 Benz) e 2,31% (Dehyquart A-CA). Em contrapartida, têm-se desvios padrões maiores que as médias, variando entre 0,90 e 4,49, o que indica uma distribuição assimétrica e alta variabilidade do conjunto de dados, caracterizando uma distribuição de cauda longa (HAIR et al., 2009). Além disso, os valores máximos atingem em média 15% m/m, o que caracteriza a presença de valores extremos no *dataset*.

A alta esparsidade e a distribuição assimétrica observadas nos dados merecem atenção na etapa de modelagem. Em situações dessa natureza, algoritmos de aprendizado de máquina tradicionais tendem a favorecer a classe majoritária, neste caso, as formulações instáveis, classificando-a corretamente com maior frequência em detrimento da classe minoritária, o que pode comprometer o desempenho preditivo do modelo (CASTRO; BRAGA, 2011).

A Figura 4 apresenta um conjunto de gráficos do tipo *pair plot* (gráfico de pares), método usado para visualização da relação entre múltiplas variáveis em um determinado *dataset*. Sua grade de diagramas de dispersão ajuda a identificar como as informações interagem entre si, evidenciando correlações, padrões e tendências entre os dados (WILKE, 2019). A diagonal principal indica a distribuição individual de cada variável, já os dados que se encontram fora deste eixo apontam a relação entre cada par de variáveis. A coloração serve como um indicador de classe, no qual azul indica instabilidade (*False*) e a cor laranja indica estabilidade (*True*).

Figura 4. *Pair Plot* das frações mássicas dos ingredientes selecionados por estabilidade.



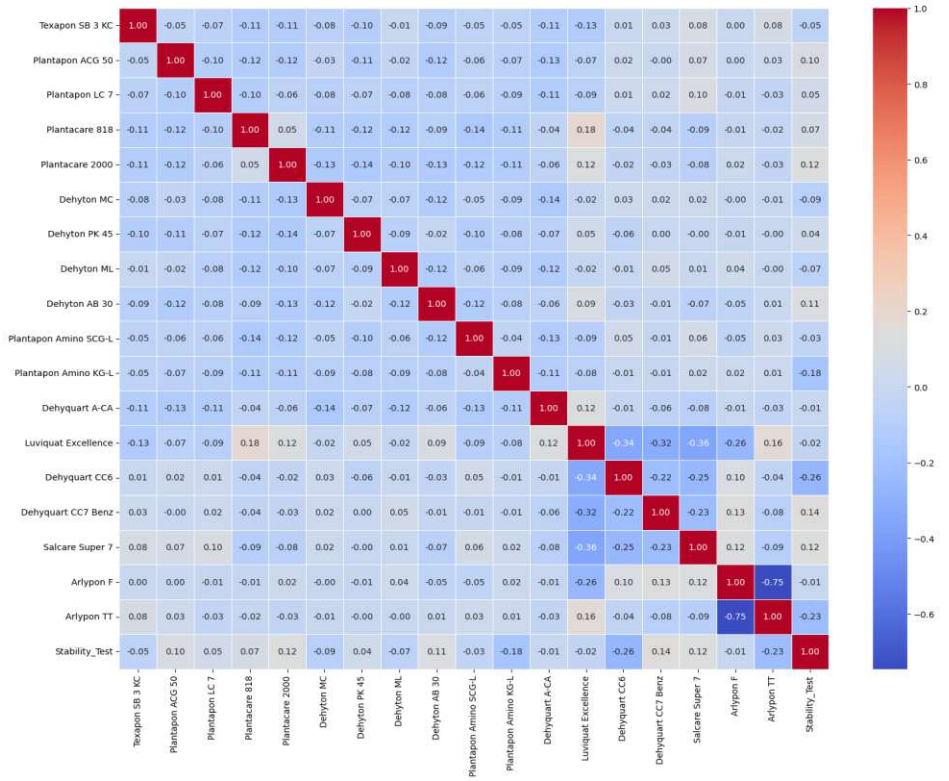
Fonte: Elaborado pela autora (2026).

É possível observar pela Figura 4 a complexa e assimétrica distribuição que os dados compõem. Não é possível identificar visualmente nenhum padrão linear, além de pontos das duas classes se sobrepondo na maioria dos pares, não há fronteiras evidentes de decisão entre estável e instável.

Essa análise implica diretamente na modelagem do processo e justificativa do trabalho proposto. Utilizar modelos lineares ou fenomenológicos seriam insuficientes para um grupo de dados que interage de forma tão complexa entre si. O uso de modelos de aprendizado de máquina torna possível a captura das relações não-lineares e de alta dimensionalidade sem que hipóteses simplificadoras sejam utilizadas no sistema analisado (WASKOM, 2021).

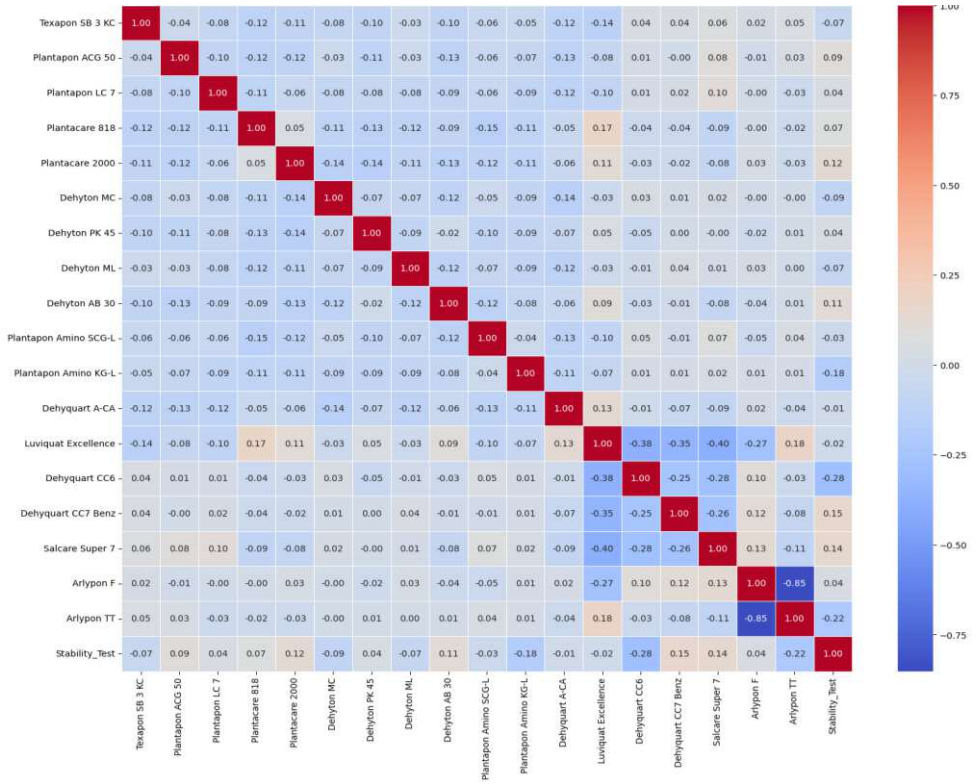
Como complementação da análise, as Figuras 5, 6 e 7 apresentam os mapas de calor das matrizes de correlação de Pearson, Spearman e Kendall, respectivamente, calculadas entre as concentrações dos ingredientes e a variável-alvo de estabilidade (*Stability_Test*).

Figura 5. Matriz de Correlação de Pearson.



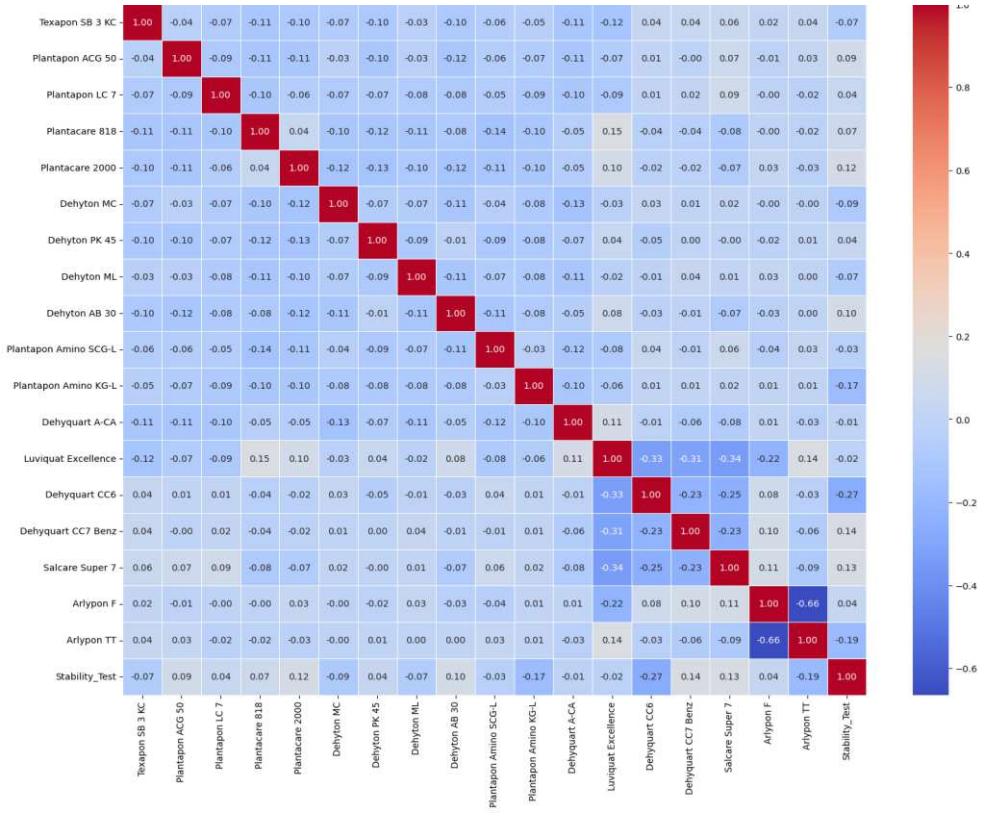
Fonte: Elaborado pela autora (2026).

Figura 6. Matriz de Correlação de Spearman.



Fonte: Elaborado pela autora (2026).

Figura 7. Matriz de Correlação de Kendall.



Fonte: Elaborado pela autora (2026).

Cada uma das métricas é responsável por um tipo de associação que mudam juntas, fornecendo informações sobre a existência de uma possível relação entre elas (BRIDEAU, 2025). A matriz de correlação de Pearson mede a correlação linear entre variáveis contínuas, é sensível a outliers e exige normalidade dos dados. Por outro lado, as correlações de Spearman e Kendall são medidas não-paramétricas baseadas no ranking das observações, adequadas para distribuições assimétricas e dados com presença de empates. A principal diferença entre Spearman e Kendall reside no método de cálculo, enquanto Spearman se baseia nas diferenças entre os rankings, Kendall é calculado a partir da proporção de pares concordantes e discordantes, sendo mais conservador e robusto à presença de empates (CONOVER, 1999). A Tabela 3 evidencia esses resultados e comportamentos.

Tabela 3. Comparação dos coeficientes de correlação de Pearson, Spearman e Kendall para os pares de variáveis mais relevantes.

Par de ingredientes	**Pearson**	**Spearman**	**Kendall**	**Interpretação**
Arlypon F × Arlypon TT	-0,75	-0,85	-0,66	Forte negativa
Luviquat Excellence × Dehyquart CC6	-0,34	-0,38	-0,33	Moderada negativa
Luviquat Excellence × Salcare Super 7	-0,36	-0,40	-0,34	Moderada negativa
Plantacare 2000 × *Stability_Test*	+0,12	+0,12	+0,12	Fraca positiva
Plantapon Amino KG-L × *Stability_Test*	-0,18	-0,18	-0,17	Fraca negativa

Fonte: Elaborado pela autora (2026).

Ao comparar os valores obtidos de correlação, pode-se ter três possíveis casos. As correlações positivas indicam que as duas variáveis crescem juntas, ou seja, quando uma aumenta a outra tende a aumentar também. Caso sejam negativas, significa que elas possuem comportamento oposto, dessa forma, quando uma aumenta, a outra tende a diminuir. Já um valor nulo indica a ausência de associação entre as variáveis (HAIR et al., 2009). Para fins de interpretação da magnitude dos coeficientes de correlação, adotou-se o padrão de Cohen (1992), segundo o qual coeficientes entre 0,10 e 0,29 representam uma associação fraca (pequena), coeficientes entre 0,30 e 0,49 representam uma associação moderada (média) e coeficientes iguais ou superiores a 0,50 representam uma associação forte (grande). Embora esses pontos de corte tenham sido originalmente propostos para o coeficiente de Pearson, sua aplicação como referência de magnitude é amplamente utilizada na literatura para interpretação de coeficientes de correlação em geral, incluindo os de Spearman e Kendall.

Em relação aos tensoativos apresentados na Tabela 3, suas correlações com a variável-alvo mostram-se de baixa magnitude, variando entre -0,18 (Plantapon Amino KG-L) e +0,12 (Plantacare 2000). Isso é um grande indicativo da ausência de multicolinearidade entre as variáveis preditoras (HAIR et al., 2009), o que apenas confirma a diversidade estatística presente no *dataset*, como já mencionado anteriormente neste trabalho.

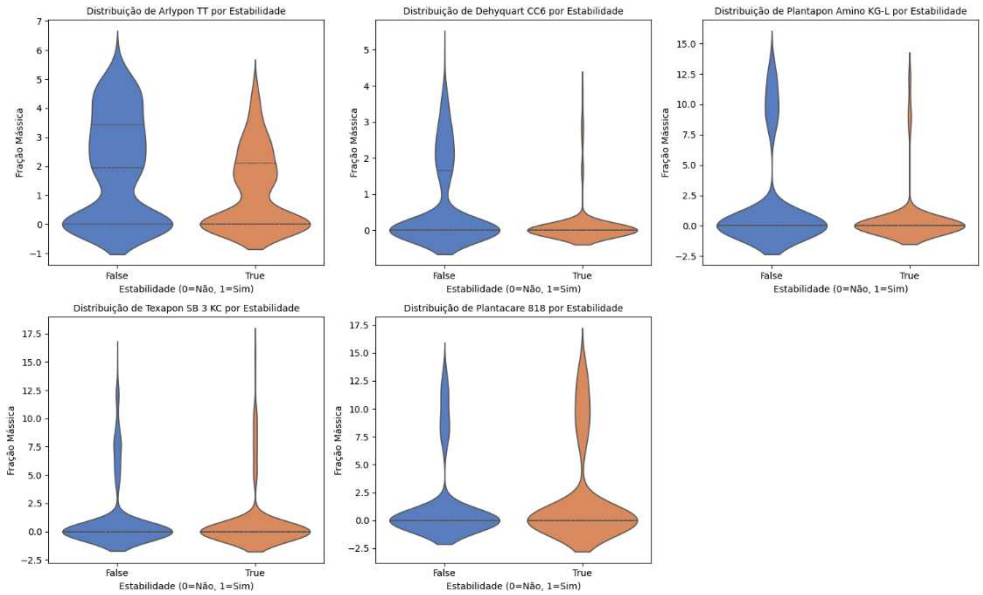
Comparando os valores detalhados pelas três métricas, é notável que a correlação mais forte se dá entre os espessantes Arlypon F × Arlypon TT, com valores de -0,75 em Pearson,

-0,85 em Spearman e -0,66 em Kendall. Essa forte correlação negativa pode ser justificada pelo fato de cada formulação usar apenas um espessante por vez, tornando-os mutuamente excludentes (CHITRE et al., 2024). Nota-se também correlações moderadas entre os polímeros condicionantes Luviquat Excellence e Dehyquart CC6, e entre Luviquat Excellence e Salcare Super 7, variando entre -0,33 e -0,40, justificado pelo menos motivo combinatório anterior, cada formulação utiliza apenas um polímero condicionante por vez.

Acerca da variável-alvo *Stability_Test*, responsável por indicar se a formulação é estável ou instável, nenhum ingrediente isolado apresentou uma alta correlação. Os valores máximos obtidos foram +0,12 do composto Plantacare 2000 e -0,18 do Plantapon Amino KG-L. Esses resultados podem ser justificados pela determinação da estabilidade, que ocorre pela interação conjunta dos ingredientes, e não por um componente isolado, o que justifica o uso de *Machine Learning* neste trabalho, capaz de capturar relações não-lineares e interações entre variáveis (BREIMAN, 2001).

Outra forma de analisar o *dataset* de estudo é através de *violin plots*, que são apresentados na Figura 8. Este tipo de gráfico combina diagramas de blocos com estimativa de densidade de probabilidade (KDE), permitindo visualizar de forma simultânea os parâmetros de distribuição, mediana e dispersão (WILKE, 2019). Essa representação apresenta as frações mássicas de cinco ingredientes selecionados, que foram segmentados pela variável de estabilidade, no qual *False* indica instabilidade e *True* estabilidade da formulação.

Figura 8. *Violin Plots* das frações mássicas dos ingredientes por estabilidade.



Fonte: Elaborado pela autora (2026).

Ao analisar de forma comparativa, é possível observar um padrão geral nos cinco ingredientes, condizente com a natureza combinatória do *dataset*, como visto anteriormente.

Há distribuições assimétricas e esparsas em ambas as classes, suas medianas encontram-se quase nulas e possuem longos extremos superiores, vistos como caudas. Além disso, os ingredientes Texapon® SB 3 KC e Plantacare® 818, isoladamente, não possuem diferença entre classes, evidenciado pelos violinos com morfologia praticamente idêntica entre estável e instável.

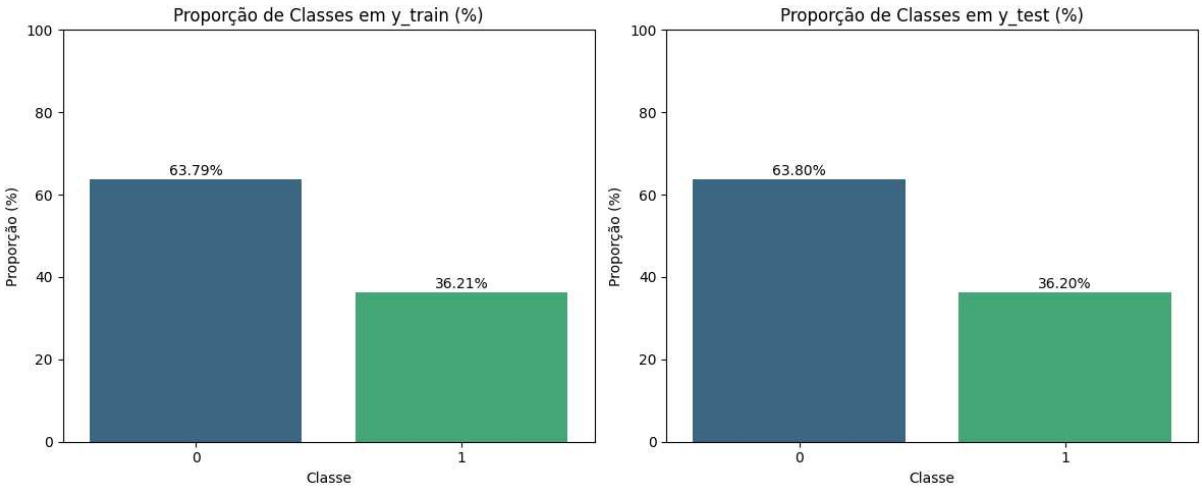
O composto Arlypon® TT mostra-se como uma exceção a essa análise. O violino da classe instável (*False*) apresenta maior volume entre 1 e 4% m/mm, enquanto a classe estável (*True*) mostra distribuição mais concentrada em entre 0 e 2% m/m. Esses parâmetros indicam uma influência sutil do espessante sobre a estabilidade, mas que acaba sendo insuficiente para separar as classes de forma univariada.

Tais resultados reforçam a conclusão obtida pela análise de correlação, no qual a estabilidade das formulações é determinada pela interação conjunta dos ingredientes, e não por qualquer componente isolado (BREIMAN, 2001).

5.2.Aplicação do Modelo de Árvores de Decisão

Neste trabalho, a divisão foi feita de forma estratificada, ou seja, mantendo a proporção original das classes (estável e instável) em ambos os conjuntos. A Figura 9 apresenta a proporção das classes obtida após essa divisão.

Figura 9. Proporção de classes nos conjuntos de treino (*y_train*) e teste (*y_test*) após divisão estratificada do *dataset*.



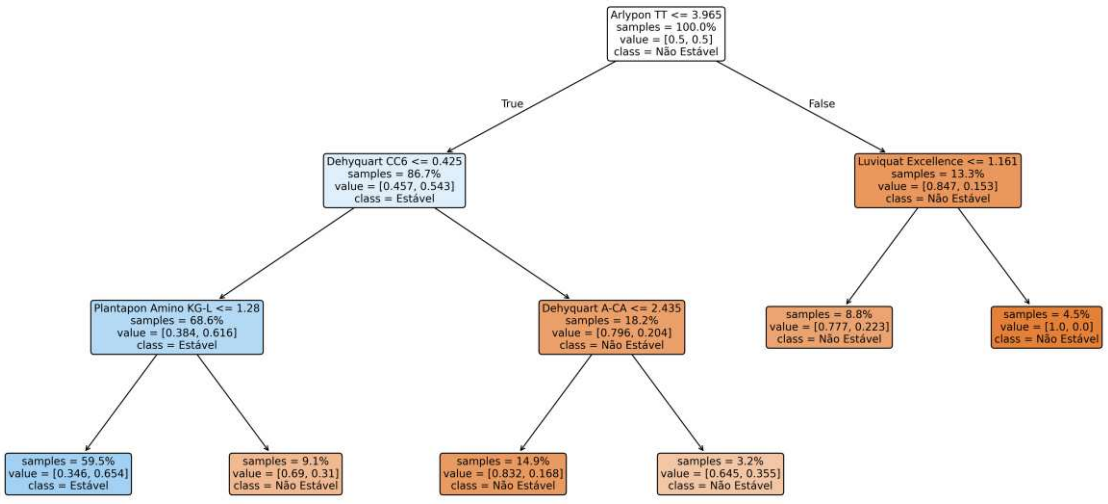
Fonte: Elaborado pela autora (2026).

Observa-se que o conjunto de treino tem cerca de 63,79% de formulações instáveis (classe 0) e 36,21% de formulações estáveis (classe 1). No conjunto de teste, essa distribuição é praticamente a mesma, com 63,80% e 36,20%, respectivamente. A semelhança entre as proporções nos dois conjuntos indica que a divisão foi realmente estratificada, ou seja, adequada. Ao manter a distribuição original das classes em ambas as partições, ajuda-se o modelo a ser treinado e avaliado em condições que representam o *dataset* completo.

Ao analisar a Figura 9, é possível identificar nos resultados um desbalanceamento de classes, no qual a classe instável (0) representa aproximadamente o dobro da classe estável (1). Esse desequilíbrio é uma característica intrínseca do problema, uma vez que, do total de 812 formulações do *dataset*, apenas 294 foram classificadas como estáveis. Em situações assim, modelos treinados sem qualquer tratamento tendem a favorecer a classe majoritária, o que reduz a capacidade de previsão sobre a classe minoritária (CASTRO; BRAGA, 2011).

Aprofundando a análise, o conjunto de dados pode ser examinado sob a perspectiva dos modelos de aprendizado de máquina apresentados anteriormente. Considera-se, inicialmente, a Árvore de Decisão, cujo modelo otimizado está representado na Figura 10.

Figura 10. Árvore de decisão treinada para classificação da estabilidade.



Fonte: Elaborado pela autora (2026).

O modelo final foi obtido com ponderação de classes (*class_weight = 'balanced'*), estratégia que atribui pesos inversamente proporcionais à frequência de cada classe,

compensando o desbalanceamento discutido na seção anterior (CASTRO; BRAGA, 2011). Por essa razão, os valores exibidos no campo *value* correspondem a proporções ponderadas, e não a contagens brutas: o nó raiz apresenta distribuição [0,5; 0,5], ainda que o conjunto original contenha cerca de 64% de formulações instáveis.

A busca em grade selecionou uma árvore de profundidade máxima igual a 3, com no mínimo 20 amostras por nó folha. Trata-se, portanto, de um modelo deliberadamente compacto, cuja estrutura apresentada na Figura 10 corresponde à árvore completa, e não a um recorte. Essa simplicidade decorre do próprio processo de otimização, que preteriu configurações mais profundas, e traz como contrapartida direta a interpretabilidade, uma das principais vantagens atribuídas a esse modelo (MITCHELL, 1997).

O nó raiz utiliza o espessante Arlypon® TT como critério de partição inicial, com limiar em 3,965% m/m, o que o identifica como o atributo mais informativo do conjunto. Quando a concentração excede esse valor (ramo *False*), obtém-se um subgrupo que concentra 13,3% das amostras e no qual 84,7% são instáveis, indicando que teores elevados de espessante estão associados à desestabilização das formulações. No ramo oposto (Arlypon® TT ≤ 3,965% m/m), que reúne 86,7% das amostras, a distribuição entre as classes é praticamente equilibrada (45,7% contra 54,3%), o que demanda divisões adicionais.

No segundo nível, ambos os ramos recorrem a polímeros condicionantes: o Dehyquart® CC6, com limiar de 0,425% m/m, no ramo majoritário, e o Luviquat® Excellence, com limiar de 1,161% m/m, no ramo minoritário. Apenas no terceiro nível surgem os tensoativos, representados pelo Plantapon® Amino KG-L e pelo Dehyquart® A-CA.

Das seis folhas resultantes, apenas uma classifica as amostras como estáveis, ainda que ela concentre a maior parte do conjunto: 59,5% das amostras, com 65,4% de formulações estáveis. Essa folha corresponde a formulações que combinam baixos teores de espessante, de Dehyquart® CC6 e de Plantapon® Amino KG-L. Nas demais folhas predomina a classe instável, com destaque para a folha mais pura do modelo (4,5% das amostras), na qual a totalidade das formulações é instável, resultado de concentrações simultaneamente elevadas de Arlypon® TT e Luviquat® Excellence.

A hierarquia da árvore reforça o achado da análise de permutação (parte final do trabalho de conclusão de curso), no qual o espessante ocupa o nó raiz, os polímeros condicionantes definem o segundo nível e os tensoativos aparecem somente no terceiro, o que sugere um papel secundário destes últimos na determinação da estabilidade.

Cabe registrar uma limitação quanto à leitura dos limiares. A estratégia de particionamento selecionada (*splitter = 'random'*) sorteia candidatos aleatórios de corte para

cada atributo, em vez de buscar exaustivamente a divisão ótima. Os valores numéricos apresentados devem, portanto, ser interpretados como ordens de grandeza, e não como pontos de corte físico-químicos precisos.

Como complementação da análise, é possível representar os resultados de classificação obtidos e o comportamento de cada classe através de matrizes de confusão. A Figura 11 exemplifica o conceito. Esta ferramenta de avaliação organiza os resultados da classificação em quatro categorias, permitindo identificar não apenas a taxa de acertos global do modelo, mas também a natureza e distribuição dos erros cometidos (MARIANO, 2021), tais como:

- Verdadeiro Positivo (VP): número de valores positivos classificados corretamente;
- Verdadeiro Negativo (VN): número de valores negativos classificados corretamente;
- Falso Positivo (FP): número de valores negativos classificados como positivos;
- Falsos Negativo (FN): número de valores positivos classificados como negativos.

Figura 11. Matriz de confusão.

Matriz de Confusão		**Classe predita**	
		Positiva	**Negativa**
Classe original	**Positiva**	**VP**	**FN**
	Negativa	**FP**	**VN**

Fonte: adaptado de Ferrari; De Castro Silva (2017).

É importante destacar também as métricas utilizadas como critério de avaliação dos modelos analisados, sendo elas acurácia, precisão, sensibilidade (recall) e F1-*score*, calculadas conforme indicadas na Tabela 4.

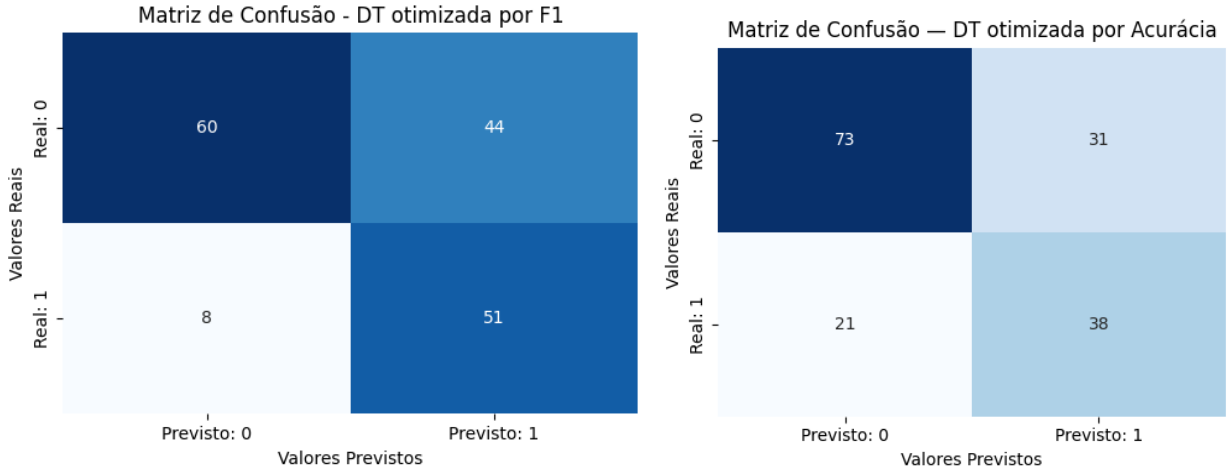
Inicialmente, foram geradas as matrizes de confusão para o modelo de Árvore de Decisão, avaliado no conjunto de teste sob duas configurações: uma para o modelo otimizado por F1-score e outro para o modelo otimizado por acurácia, conforme a Figura 12. As métricas correspondentes estão sintetizadas na Tabela 5.

Tabela 4. Critérios de avaliação dos modelos.

Critério	**Fórmula**
Acurácia	$\frac{VP + VN}{VP + VN + FP + FN}$
Precisão	$\frac{VP}{VP + FP}$
Sensibilidade (*Recall*)	$recall = \frac{VP}{VP + FN}$
F1-s*core*	$\frac{2 \times (Precisão \times Recall)}{Precisão + Recall}$

Fonte: Adaptado de Pádua (2018).

Figura 12. Matrizes de confusão aplicadas aos dados do conjunto de teste utilizando o modelo de Árvore de Decisão com parâmetros otimizados utilizando F1 (a) e acurácia (b). "Não Estável" = 0/ "Estável" = 1.



(a) (b)

Fonte: Elaborado pela autora (2026).

Tabela 5. Desempenho da Árvore de Decisão no conjunto de teste, segundo a métrica adotada na otimização.

Métrica		**Otimizada por F1**	**Otimizada por Acurácia**
Acurácia global		0,68	0,68
Classe Estável (1)	Precisão	0,54	0,55
	Sensibilidade	0,86	0,64
	F1-*score*	0,66	0,59
Classe Não Estável (0)	Precisão	0,88	0,78
	Sensibilidade	0,58	0,70
	F1-*score*	0,70	0,74
Erros	Falsos positivos	44	31
	Falsos negativos	8	21

Fonte: Elaborado pela autora (2026).

No modelo otimizado por F1-*score*, foram registrados 60 verdadeiros negativos (VN) e 51 verdadeiros positivos (VP), totalizando 111 classificações corretas em 163 amostras, o que corresponde a uma acurácia global de 68,1%. Entre os erros, destacam-se 44 falsos positivos (FP), ou seja, formulações instáveis classificadas como estáveis, e apenas 8 falsos negativos (FN), correspondentes a formulações estáveis classificadas como instáveis. A taxa de FP foi de 42,3% em relação ao total de amostras instáveis, enquanto a de FN limitou-se a 13,6% das amostras estáveis.

O modelo otimizado por acurácia apresentou 73 VN e 38 VP, totalizando igualmente 111 acertos e a mesma acurácia global de 68,1%. Os FP foram reduzidos para 31 (29,8% das instáveis), ao passo que os FN aumentaram para 21 (35,6% das estáveis).

A comparação entre as duas matrizes revela o impacto direto da métrica de otimização sobre o perfil de erros. Ambos os modelos cometeram exatamente o mesmo número de erros (52) e alcançaram acurácia idêntica, diferindo apenas quanto à sua distribuição: o modelo otimizado por acurácia converteu 13 falsos positivos em 13 falsos negativos. Esse resultado evidencia, de forma concreta, a limitação da acurácia como métrica isolada em conjuntos desbalanceados, ela foi incapaz de distinguir dois modelos com comportamentos substancialmente distintos (CASTRO; BRAGA, 2011).

A diferença, no entanto, é expressiva quando se observa a classe minoritária. A otimização por F1-*score* elevou a sensibilidade sobre as formulações estáveis de 0,64 para 0,86, resultando em F1-*score* superior para essa classe (0,66 contra 0,59). Em outras palavras, a otimização por acurácia não produziu ganho algum em acurácia, mas comprometeu a capacidade do modelo de identificar formulações estáveis, favorecendo a classe majoritária, um comportamento esperado em conjuntos desbalanceados.

Do ponto de vista prático, considerando o modelo como ferramenta de triagem prévia à etapa experimental, os FN representam o erro mais crítico: correspondem a formulações viáveis descartadas antes de serem testadas, constituindo uma perda irrecuperável e não detectável. Os FP, por sua vez, implicam ensaios improdutivos, cujo custo se limita ao próprio experimento, uma vez que a instabilidade é identificada em laboratório. O modelo otimizado por F1-*score* privilegia justamente essa prioridade, preservando 86% das formulações estáveis.

Diante desses resultados, adotou-se o F1-*score* como métrica de otimização para os modelos de conjunto, apresentados nas seções seguintes. Ainda assim, ambas as configurações da Árvore de Decisão mantiveram taxas de erro consideráveis, o que reforça a necessidade de modelos com maior capacidade preditiva.

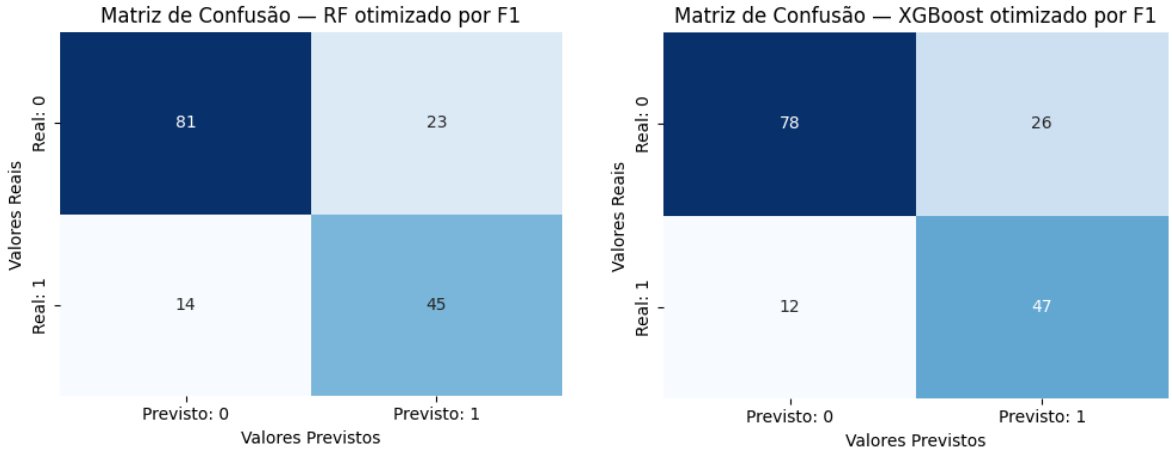
5.3. Aplicação dos Modelos do tipo Ensemble (RF e XGBoost)

Diante das limitações observadas na árvore individual, foram avaliados dois métodos de conjunto (*ensemble*): o *Random Forest*, baseado na estratégia de *bagging*, e o XGBoost, baseado na estratégia de *boosting*. Conforme justificado na seção anterior, ambos foram otimizados exclusivamente pelo F1-*score*, métrica mais adequada ao conjunto desbalanceado em estudo. Para assegurar a comparabilidade entre os três modelos, aplicou-se a mesma ponderação de classes em todos eles: no *Random Forest*, por meio do parâmetro *class_weight='balanced'*, no XGBoost por meio do parâmetro *scale_pos_weight*, ajustado em 1,762, valor correspondente à razão entre as frequências das classes no conjunto de treinamento e equivalente à ponderação adotada nos demais modelos.

A busca em grade selecionou, para o *Random Forest*, uma floresta de 200 árvores sem limitação de profundidade, com no mínimo cinco amostras por nó folha. Para o XGBoost, foram selecionadas 200 árvores de profundidade máxima igual a 7, com taxa de aprendizado de 0,01 e subamostragem de 0,7, uma configuração marcadamente regularizada, na mesma direção conservadora observada nos modelos anteriores. As matrizes de confusão obtidas para ambos

os modelos são apresentadas na Figura 13, e a Tabela 6 sintetiza o desempenho dos três modelos sobre o conjunto de teste.

Figura 13. Matrizes de confusão obtidas no conjunto de teste: (a) *Random Forest*; (b) XGBoost.



(a) (b)

Elaborado pela autora (2026).

Ambos os modelos de conjunto superaram de forma consistente o modelo de referência. A acurácia elevou-se de 68,1% para 77,3% (*Random Forest*) e 76,7% (XGBoost), enquanto o F1-*score* sobre a classe estável passou de 0,662 para 0,709 e 0,712, respectivamente. Esse ganho decorre, sobretudo, da precisão: a árvore individual apresentava precisão de apenas 0,54 sobre a classe estável, ao passo que os modelos de conjunto alcançaram 0,66 e 0,64.

A natureza dessa melhoria fica evidente ao se observar o número de formulações classificadas como estáveis. Conforme a Figura 13, os modelos de conjunto atribuíram essa classe a 68 (*Random Forest,* 23+45) e 73 (XGBoost, 26+47) das 163 amostras, aproximando-se substancialmente das 59 formulações efetivamente estáveis, ao passo que a árvore individual havia classificado 95 (44+51) amostras como estáveis, comportamento excessivamente permissivo. Em termos de erros, os falsos positivos caíram de 44 para 23 e 26, uma redução de cerca de 48% e 41%, respectivamente, ao custo de uma queda moderada na sensibilidade (de 0,86 para 0,76 e 0,80).

Esse resultado é coerente com o comportamento esperado dos métodos de conjunto. Ao agregar múltiplas árvores treinadas sobre reamostragens distintas dos dados e das variáveis, o *Random Forest* reduz a variância do modelo e diminui a chance de sobreajuste em comparação com a árvore individual (BREIMAN, 2001). O XGBoost, por sua vez, alcança efeito análogo por meio da construção sequencial de árvores voltadas à correção dos erros residuais das

anteriores (CHEN; GUESTRIN, 2016). Em ambos os casos, a agregação atenuou as decisões específicas e pouco generalizáveis que a árvore individual havia incorporado.

Tabela 6. Desempenho dos modelos sobre o conjunto de teste.

Métrica	**Árvore de Decisão**	***Random Forest***	**XGBoost**
F1-*score* médio (validação cruzada)	0,644	0,701	0,690
F1-*score* (classe estável, teste)	0,662	0,709	0,712
Acurácia	0,681	0,773	0,767
Precisão (classe estável)	0,54	0,66	0,64
Sensibilidade (classe estável)	0,86	0,76	0,80
Falsos positivos	44	23	26
Falsos negativos	8	14	12
Formulações previstas como estáveis	95	68	73

Fonte: Elaborado pela autora (2026).

Entre os dois modelos de conjunto, contudo, o desempenho mostrou-se praticamente equivalente. O XGBoost obteve F1-*score* marginalmente superior no conjunto de teste (0,712 contra 0,709), diferença de três milésimos que corresponde, na prática, a uma única amostra entre 163. O *Random Forest*, por outro lado, apresentou o maior F1-*score* médio na validação cruzada (0,701 contra 0,690), acurácia superior (0,773 contra 0,767) e uma classificação correta a mais. As diferenças concentram-se no perfil de erros: o XGBoost identificou duas formulações estáveis a mais (sensibilidade de 0,80 contra 0,76), ao custo de três falsos positivos adicionais.

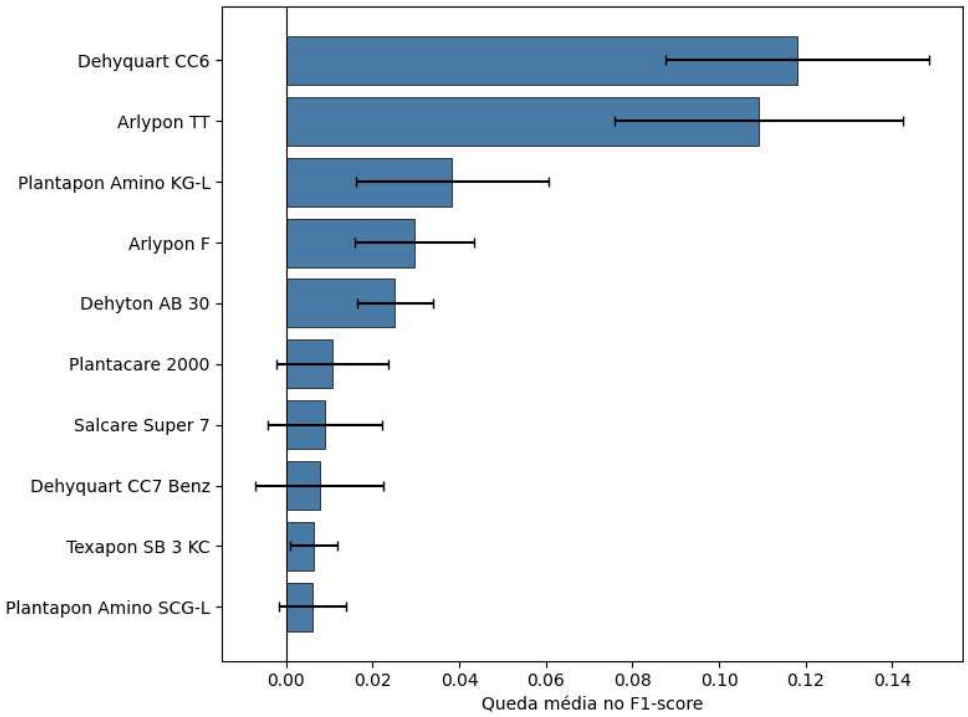
Diferenças dessa magnitude, restritas a poucas amostras em um conjunto de teste de 163 elementos, não permitem afirmar a superioridade de um modelo sobre o outro. Adotou-se, portanto, o *Random Forest* como modelo final, em conformidade com o critério de seleção estabelecido na metodologia, o desempenho médio na validação cruzada, no qual esse modelo apresentou o melhor resultado. Essa escolha apoia-se, ainda, na maior confiabilidade dessa estimativa, obtida como média sobre cinco partições de 649 amostras, em contraposição a uma medida única sobre 163, bem como na notável estabilidade do *Random Forest* entre as etapas de validação e teste (0,701 e 0,709), indicativa de boa capacidade de generalização.

Do ponto de vista prático, o ganho pode ser dimensionado pela ótica da triagem experimental. Das 163 formulações candidatas, 59 são estáveis. Na ausência de um modelo preditivo, todas as 163 precisariam ser ensaiadas para identificá-las. Adotando-se o *Random Forest* como filtro prévio, seriam levadas ao ensaio apenas as 68 formulações indicadas como estáveis, entre as quais 45 efetivamente o são. Reduz-se, assim, o esforço experimental a menos da metade do total, preservando-se 45 das 59 formulações estáveis. Em termos de rendimento, enquanto o ensaio das formulações candidatas sem triagem produziria cerca de uma formulação estável a cada três, o modelo eleva essa proporção a duas em cada três.

5.4.Interpretabilidade da Floresta Aleatória

A Figura 14 apresenta a importância relativa das variáveis, obtida por permutação sobre o conjunto de teste. Dois componentes destacam-se: o polímero condicionante Dehyquart® CC6 (0,118) e o espessante Arlypon® TT (0,109), com importância cerca de três vezes superior à do terceiro colocado. Seguem-se, com contribuição modesta, o Plantapon® Amino KG-L (0,038), o espessante Arlypon® F (0,030) e o Dehyton® AB 30 (0,025). Para os demais ingredientes, o desvio padrão é da ordem da própria média, de modo que suas contribuições não se distinguem do ruído.

Figura 14. Importância relativa das variáveis para o modelo *Random Forest*, obtida por permutação sobre o conjunto de teste e expressa como a queda média no F1-*score* após o embaralhamento de cada ingrediente. As barras representam a média e o desvio padrão de 30 repetições. São apresentados os dez ingredientes mais influentes.



Fonte: Elaborado pela autora (2026).

A distribuição por classe é informativa: as duas primeiras posições correspondem a um polímero condicionante e a um espessante, e ambos os espessantes figuram entre os quatro primeiros, ao passo que o mais influente dentre os doze tensoativos ocupa apenas a terceira posição. Os agentes estruturantes mostram-se, portanto, determinantes para a estabilidade, cabendo à escolha individual do tensoativo papel secundário.

Esse comportamento é coerente com a físico-química do sistema. A associação entre os polímeros catiônicos e os tensoativos aniônicos e anfotéricos leva à formação de complexos polieletrólito-tensoativo que, conforme a composição, permanecem solúveis ou sofrem separação de fases associativa, um fenômeno conhecido como coacervação (KAKIZAWA; MIYAKE, 2019). A estabilidade corresponde à condição em que o complexo se mantém em fase única, e é o polímero catiônico que governa esse equilíbrio. Reforça essa leitura o fato de o Dehyquart® CC6 (Polyquaternium-6) ser o polímero de maior densidade de carga do conjunto, por tratar-se de homopolímero, enquanto os demais polímeros, copolímeros de carga diluída (LAPKIN et al., 2024), situam-se entre os menos influentes.

Os resultados convergem com a análise da árvore de decisão: ambas posicionam os agentes estruturantes (espessante e polímero condicionante) acima dos tensoativos em importância para a estabilidade. Trata-se de medidas de naturezas distintas, seleção de divisões e degradação do desempenho preditivo, de modo que a concordância entre elas reforça esse achado.

6. CONCLUSÃO

De forma geral, analisando todos os modelos, é possível notar uma evolução consistente no desempenho à medida que os modelos se tornam mais complexos. A Árvore de Decisão, como *baseline*, obteve F1-*score* de 0,6623 e acurácia de 68% quando otimizada por F1-*score*. Os modelos de conjunto elevaram essas métricas de forma equivalente: o *Random Forest* atingiu F1-*score* de 0,709 e acurácia de 77,3%, enquanto o XGBoost obteve F1-*score* de 0,712 e acurácia de 76,7%. Diante dessa equivalência no conjunto de teste, o *Random Forest* foi adotado como modelo final por apresentar o melhor F1-*score* médio na validação cruzada (0,701 contra 0,690 do XGBoost), estimativa considerada mais confiável por resultar da média sobre cinco partições.

Essa progressão também é visível nas matrizes de confusão, onde ocorre redução expressiva de falsos positivos frente à Árvore de Decisão. A Árvore de Decisão registrou 44 FP e 8 FN, o *Random Forest* reduziu os falsos positivos para 23, ao custo de um aumento nos falsos negativos para 14, e o XGBoost apresentou um resultado próximo, com 26 FP e 12 FN. Os dois modelos de conjunto atenuaram de forma semelhante o comportamento excessivamente permissivo observado na árvore individual.

A análise exploratória dos dados evidenciou a elevada esparsidade e forte assimetria nas distribuições, características decorrentes da natureza combinatória do *dataset*, os quais por si só já indicavam que modelos lineares seriam inadequados para o problema. A evolução observada entre os modelos, tanto nos parâmetros de desempenho quanto na diminuição de erros nas matrizes de confusão, confirma que o aumento da complexidade dos algoritmos *ensemble* melhora a capacidade preditiva em problemas de classificação com dados desbalanceados e de alta dimensionalidade.

Do ponto de vista industrial, empregar modelos preditivos como o *Random Forest* no desenvolvimento de formulações cosméticas surge como uma alternativa promissora para reduzir custos, acelerar o ciclo de desenvolvimento e diminuir a dependência de testes empíricos de estabilidade, que por si só demandam muito tempo. Assim, será possível contribuir para processos de formulações mais ágeis e orientados por dados.

7. REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE INDÚSTRIAS DO SETOR DE HIGIENE PESSOAL, PERFUMARIA E COSMÉTICOS (ABIHPEC). **Panorama do setor de higiene pessoal, perfumaria e cosméticos – resultados 2025**. Disponível em: https://abihpec.org.br/site2019/wp-content/uploads/2025/04/Panorama-do-Setor-de-Beleza-e-Cuidados-Pessoais_12.11.25_Port.pdf. Acesso em: 2 dez. 2025.

AMAZON WEB SERVICES (AWS). **Machine Learning Developer Guide**. Disponível em: https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/machinelearning-dg.pdf. Acesso em: 30 jun. 2026.

BARZOTTO, I. L. M. et al. Estabilidade de emulsões frente a diferentes técnicas de homogeneização e resfriamento: **Visão Acadêmica**, [S.l.], v. 10, n. 2, p. 36–42, 2009. DOI: https://doi.org/10.5380/acd.v10i2.21333. Disponível em: https://revistas.ufpr.br/academica/article/view/21333. Acesso em: 2 dez. 2025.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. Boca Raton: Chapman and Hall/CRC, 1984. DOI: https://doi.org/10.1201/9781315139470. Disponível em: https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-olshen-charles-stone. Acesso em: 30 jun. 2026.

BREIMAN, L. **Machine Learning:** Random forests, v. 45, n. 1, p. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324. Acesso em: 25 maio 2026.

BRIDEAU, L. **Correlation: Pearson, Spearman, and Kendall's Tau**. Charlottesville: University of Virginia Library, 2025. Disponível em: https://library.virginia.edu/data/articles/correlation-pearson-spearman-and-kendalls-tau. Acesso em: 1 jul. 2026.

BRUCE, A.; BRUCE, P.; GEDEK, P. **Practical statistics for data scientists**: 50+ essential concepts using R and Python. Sebastopol: O'Reilly Media, 2017.

CASTRO, C. L. de; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle & Automação Sociedade Brasileira de Automática**, v. 22, n. 5, p. 441–466, out. 2011. DOI: https://doi.org/10.1590/S0103-17592011000500002. Disponível em: https://www.scielo.br/j/ca/a/pXMZjzHJcJtkLVYLLDHHTxw/?lang=pt. Acesso em: 14 jul. 2026.

CHEN, L. et al. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. **Information Sciences**, v. 509, p. 150–163, 2020. DOI: https://doi.org/10.1016/j.ins.2019.09.005. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0020025519308503?via%3Dihub. Acesso em: 30 jun. 2026.

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. **Proceedings** [...]. New York: ACM, 2016. DOI: https://doi.org/10.1145/2939672.2939785. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939785 Acesso em 30 jun. 2026.

CHITRE, A. et al. Accelerating formulation design via machine learning: generating a high-throughput shampoo formulations dataset. **Nature,** v. 11, 1234, 2024. DOI: https://doi.org/10.1038/s41597-024-03573-w. Disponível em: https://www.nature.com/articles/s41597-024-03573-w. Acesso em: 25 maio 2026.

COHEN, J. Statistical power analysis. **Current Directions in Psychological Science**, v. 1, n. 3, p. 98–101, 1992. DOI: https://doi.org/10.1111/1467-8721.ep10768783. Disponível em: https://journals.sagepub.com/doi/10.1111/1467-8721.ep10768783. Acesso em: 14 de julho 2026.

CONOVER, W. J. **Practical nonparametric statistics**. 3. ed. New York: John Wiley & Sons, 1999.

FACELI, K. et al. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

FERRARI, D. G.; DE CASTRO SILVA, L. N. **Introdução à mineração de dados**. [S.l.]: Saraiva Educação, 2017.

HAIR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, v. 585, n. 7825, p. 357-362, 2020. DOI: https://doi.org/10.1038/s41586-020-2649-2. Disponível em: https://www.nature.com/articles/s41586-020-2649-2. Acesso em 30 jun. 2026.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2. ed. New York: Springer, 2009. ISBN 978-0-387-84857-0. DOI: https://doi.org/10.1007/978-0-387-84858-7. Disponível em: https://link.springer.com/book/10.1007/978-0-387-84858-7. Acesso em 30 jun. 2026.

HINTZE, J. L.; NELSON, R. D. Violin plots: a box plot-density trace synergism. **The American Statistician**, v. 52, n. 2, p. 181-184, 1998. DOI: https://doi.org/10.1080/00031305.1998.10480559. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559. Acesso em 30 jun. 2026.

HUNTER, J. D. Matplotlib: a 2D graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90-95, 2007. DOI: https://doi.org/10.1109/MCSE.2007.55. Disponível em: https://ieeexplore.ieee.org/document/4160265. Acesso em 30 jun. 2026.

KAKIZAWA, Y.; MIYAKE, M. Creation of new functions by combination of surfactant and polymer: complex coacervation with oppositely charged polymer and surfactant for shampoo and body wash. **Journal of Oleo Science**, v. 68, n. 6, p. 525–539, 2019. DOI: https://doi.org/10.5650/jos.ess19081. Disponível em: https://www.jstage.jst.go.jp/article/jos/68/6/68_ess19081/_article. Acesso em 30 jun. 2026.

LAPKIN, A.; CHITRE, A.; QUERIMIT, R.; RIHM, S.; KARAN, D.; ZHU, B.; WANG, L.; WANG, K.; HIPPALGAONKAR, K. **BASF formulation ingredient structures**. Figshare, 2024. DOI: https://doi.org/10.6084/m9.figshare.25451929. Disponível em:

https://springernature.figshare.com/articles/figure/Structures_of_Formulations_Ingredients/25451929. Acesso em: 12 jul. 2026.

LOH, W.-Y. Fifty years of classification and regression trees. **International Statistical Review**, v. 82, n. 3, p. 329–348, 2014. DOI: https://doi.org/10.1111/insr.12016. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/insr.12016?__cf_chl_rt_tk=ONfxkM7oMV0J.6.O3nWXaQe0PDPiCaPM9nxN7kzvR_M-1784409130-1.0.1.1-o9l7LqFeQHJ5GufelG.DAfFt0eKbl5CW2nfiOo7TmpA. Acesso em: 12 jul. 2026.

McKINNEY, W. Data structures for statistical computing in Python. In: PYTHON IN SCIENCE CONFERENCE, 9., 2010, Austin. **Proceedings** [...]. Austin: SciPy, 2010. p. 56-61. DOI: https://doi.org/10.25080/Majora-92bf1922-00a. Disponível em: https://proceedings.scipy.org/articles/Majora-92bf1922-00a. Acesso em: 12 jul. 2026.

MARIANO, D. Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e F-score. **Revista Brasileira de Bioinformática e Biologia Computacional**, v. 1, p. 15, 2021. Disponível em: https://bioinfo.com.br/metricas-de-avaliacao-em-machine-learning-acuracia-sensibilidade-precisao-especificidade-e-f-score/. Acesso em: 30 jul. 2026.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997.

ONODA, M.; EBECKEN, N. F. F. Implementação em Java de um algoritmo de árvore de decisão acoplado a um SGBD relacional. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 16., 2001, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: SBBD, 2001. p. 55-64.

PÁDUA, M. Machine learning: métricas de avaliação — acurácia, precisão e recall. **Medium**, 2018. Disponível em: https://medium.com/@mateuspdua/machine-learning-m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-e-recall-d44c72307959. Acesso em: 2 jul. 2026.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011. Disponível em: https://www.jmlr.org/papers/v12/pedregosa11a.html. Acesso em: 15 jul. 2026.

QUINLAN, J. R. **C4.5:** programs for machine learning. [S.l.]: Elsevier, 2014.

SATHYA, R.; ABRAHAM, A. Comparison of supervised and unsupervised learning algorithms for pattern classification. **International Journal of Advanced Research in Artificial Intelligence**, v. 2, n. 2, p. 34–38, 2013. DOI: https://doi.org/10.14569/IJARAI.2013.020206. Disponível em: https://thesai.org/Publications/ViewPaper?Volume=2&Issue=2&Code=IJARAI&SerialNo=6. Acesso em: 12 jul. 2026.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

WASKOM, M. Seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021. Disponível em: https://joss.theoj.org/papers/10.21105/joss.03021. Acesso em: 25 maio 2026.

WILKE, C. O. **Fundamentals of data visualization**: a primer on making informative and compelling figures. Sebastopol: O'Reilly Media, 2019. Disponível em: https://clauswilke.com/dataviz/. Acesso em: 25 maio 2026.