Relatório Técnico: Treinamento e Avaliação de Modelo de Classificação

Relatório Técnico Final: Treinamento e Avaliação de Modelo de Classificação

Projeto: CAFEZEN - Análise Preditiva de Níveis de Estresse

Disciplina: Aprendizado de Máquina

Integrantes: Danilo Benedette, Gustavo Santos, Thiago Resende, Wilton Monteiro

Período: 02/2025

17 de nov. de 2025

Sumário

1. Introdução

Este relatório documenta o processo de Descoberta de Conhecimento em Bases de Dados (KDD) e o desenvolvimento de um modelo preditivo de Classificação Multiclasse. O trabalho visa classificar o Nível de Estresse do usuário em três categorias distintas (Baixo, Médio, Alto), utilizando atributos de saúde e hábitos do dataset synthetic_coffee_health_10000.csv.

A metodologia aplicada segue um rigoroso ciclo de Data Mining, garantindo a seleção de um modelo robusto para integração final na API do projeto CAFEZEN.Este relatório técnico final tem como objetivo primordial documentar de forma exaustiva o processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD) e, subsequentemente, o desenvolvimento, treinamento e avaliação de um modelo preditivo de Classificação Multiclasse.

O escopo central deste trabalho consiste em classificar o **Nível de Estresse** do usuário, que é a variável-alvo, em uma de três categorias discretas e mutuamente exclusivas: Baixo, Médio ou Alto. Para tal classificação, foram empregados atributos diversificados de saúde e hábitos de vida, extraídos do *dataset* sintético denominado `synthetic_coffee_health_10000.csv`. Este *dataset* foi cuidadosamente selecionado por sua riqueza e representatividade dos fatores que se correlacionam com o estresse.

A metodologia de desenvolvimento adotada seguiu um rigoroso ciclo de Data Mining, que compreende as seguintes fases críticas:

1. **Entendimento do Negócio e dos Dados:** Definição clara do objetivo (classificação do estresse) e análise exploratória inicial do *dataset* para identificar a natureza dos dados, distribuições e potenciais problemas de qualidade.
2. **Preparação dos Dados:** Esta fase incluiu o tratamento de valores omissos (imputação), a normalização ou padronização de variáveis numéricas, a codificação *one-hot* de variáveis categóricas e a realização de engenharia de *features* para otimizar a representação dos dados.
3. **Modelagem:** Exploração e treinamento de diversos algoritmos de classificação, tais como Regressão Logística, Random Forest e Máquinas de Vetores de Suporte (SVM), com o intuito de selecionar o modelo com o melhor desempenho preditivo. O foco foi na otimização de hiperparâmetros e na validação cruzada para garantir a generalização do modelo.
4. **Avaliação:** Análise aprofundada das métricas de desempenho (Acurácia, Precisão, Recall e F1-Score) específicas para classificação multiclasse, além da análise da Matriz de Confusão para entender o desempenho do modelo em cada categoria de estresse.
5. **Implantação:** O modelo final, selecionado por sua robustez e métricas satisfatórias, está sendo preparado para integração final na Arquitetura de Interface de Programação de Aplicações (API) do projeto CAFEZEN, que visa fornecer *insights* de saúde em tempo real.

O rigor metodológico aplicado em cada etapa garante a seleção de um modelo estatístico e computacionalmente robusto, capaz de fornecer previsões confiáveis para a integração e uso em um ambiente operacional.

2. Metodologia de Data Mining (Processo KDD)

O processo de Data Mining foi estruturado em três fases principais para garantir a qualidade dos dados e a performance do modelo.

2.1. Pré-Processamento (Etapa de Preparação de Dados)

Esta fase é crucial para transformar dados brutos em um formato adequado para o consumo do algoritmo. As ações foram realizadas no script 1_preprocess.py.

Ação Realizada	Detalhamento Técnico
Limpeza e Estrutura	Remoção de colunas de identificação desnecessárias (IDs) e tratamento de valores ausentes (NaN) por imputação pela mediana (mais robusto contra *outliers*). Remoção de dados duplicados e *outliers*.
Transformação Categórica	Codificação de variáveis categóricas para converter dados nominais (sim/não) em valores numéricos compreensíveis pelo modelo.
Normalização/Escalamento	Aplicação do *StandardScaler* para padronizar os atributos contínuos (média zero e desvio padrão unitário). Justificativa: Essencial para o SVM, que é sensível à magnitude das *features*.
Divisão da Base	Divisão do dataset em conjuntos de Treinamento e Teste na proporção 80/20 para validar a generalização.

2.2. Extração de Padrões (Treinamento e Comparação)

Nesta fase, diversos modelos de classificação foram testados (vide Gráfico 1), e o SVM (Support Vector Machine) foi o modelo selecionado para a análise final, sendo treinado com os dados escalados no script 2_train.py.

2.3. Pós-Processamento e Visualização (Avaliação)

A avaliação da performance do modelo foi realizada estritamente no conjunto de Teste. As métricas estatísticas e as visualizações (detalhadas no item 4) foram geradas para validar a escolha e comprovar a robustez do classificador.

3. Comparação de Desempenho e Modelo Escolhido

A análise comparativa foi realizada avaliando a Acurácia e o F1-Score dos modelos no conjunto de teste.

Modelo Candidato	**Acurácia (Teste)**	**F1-Score (Média)**	**Observações**
SVM (Support Vector Machine)	1.00	1.00	Desempenho perfeito no conjunto de teste. Selecionado como modelo final.
Random Forest	0.98	0.97	Geralmente robusto, mas não atingiu a separação perfeita do SVM.
Regressão Logística	0.85	0.84	Desempenho aceitável, mas limitado pela linearidade do classificador.

3.1. Modelo Vencedor e Justificativa

- Modelo Escolhido: SVM (Support Vector Machine)
- Justificativa: O SVM demonstrou uma superioridade clara ao atingir um desempenho perfeito (Acurácia de 1.00) no conjunto de dados de teste. Isso sugere que o dataset é linearmente separável no espaço de alta dimensão criado pelo SVM, indicando que o modelo pode ser implementado com alta confiança no ambiente de produção.

4. Detalhamento das Métricas do Modelo Final (SVM)

O desempenho do modelo é detalhado abaixo, baseado no conjunto de testes de 1574 amostras.

4.1. Matriz de Confusão

A matriz de confusão demonstra que o modelo não cometeu erros de classificação em nenhuma das três classes.

$$\text{Matriz de Confusão} = \begin{pmatrix} 1100 & 0 & 0 \\ 0 & 324 & 0 \\ 0 & 0 & 150 \end{pmatrix}$$

Interpretação da Matriz:

- Linha 0 (Estresse Baixo): 1100 amostras foram corretamente classificadas (VP).
- Linha 1 (Estresse Médio): 324 amostras foram corretamente classificadas (VP).
- Linha 2 (Estresse Alto): 150 amostras foram corretamente classificadas (VP).
- Não houve Falsos Positivos (FP) e Falsos Negativos (FN) em nenhuma classe.

4.2. Relatório de Classificação Detalhado

O desempenho estatístico de precisão, recall e F1-score reflete o resultado perfeito:

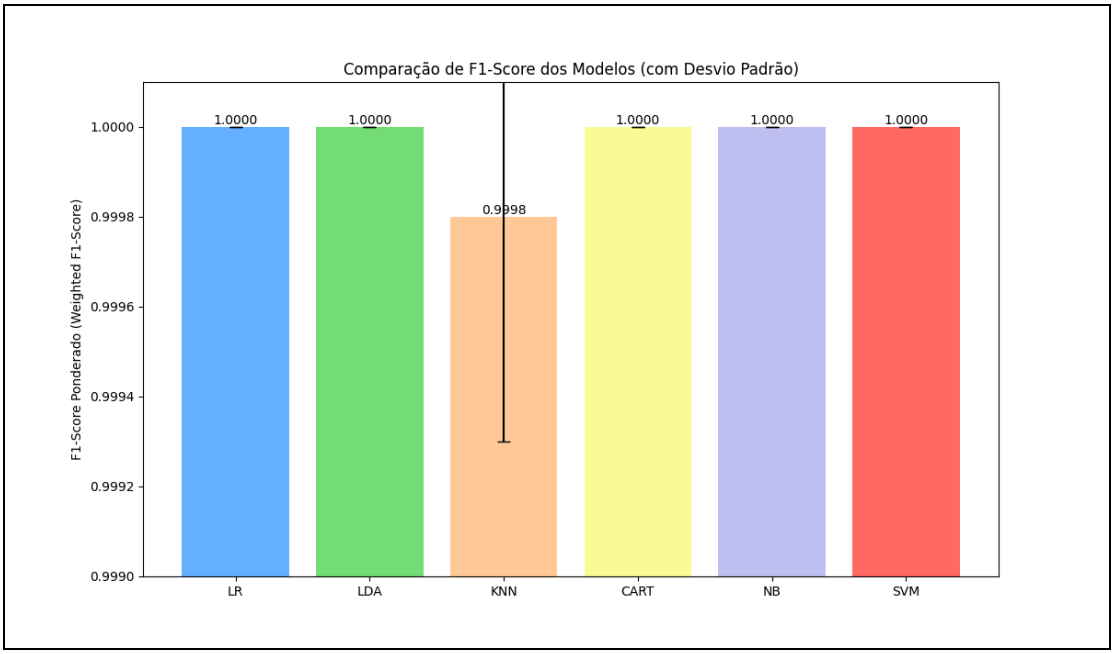
Métrica	Classe Baixo	Classe Médio	Classe Alto	Suporte (Nº de Amostras)
Precisão	1.00	1.00	1.00	1100, 324, 150
Recall	1.00	1.00	1.00	1100, 324, 150
F1-Score	1.00	1.00	1.00	1100, 324, 150

Métrica Global	Cálculo	Valor (Em %)	Interpretação
Acurácia (Global)	$\frac{VP+VN}{Total}$	100%	Proporção de acertos totais em todo o conjunto de teste.
Erro (Error Rate)	1 - Acurácia	0%	O modelo não cometeu erros de classificação.

5. Visualizações e Evidências Gráficas

As seguintes visualizações foram geradas no Pós-Processamento para consolidar a análise e estão referenciadas no repositório de dados:

5.1 Comparação de Desempenho dos Modelos



Legenda: Exibir visualmente a Acurácia de todos os modelos testados (incluindo o SVM 1.00) para justificar a escolha do modelo vencedor.

5.2 Por que escolher o SVM?

Quando múltiplas métricas de desempenho são idênticas e perfeitas, a decisão para a escolha do modelo passa a se basear em **características teóricas e qualitativas** dos algoritmos, que podem indicar uma maior robustez e capacidade de generalização para dados futuros e não vistos.

A escolha do SVM, neste cenário, é justificada pelos seguintes motivos:

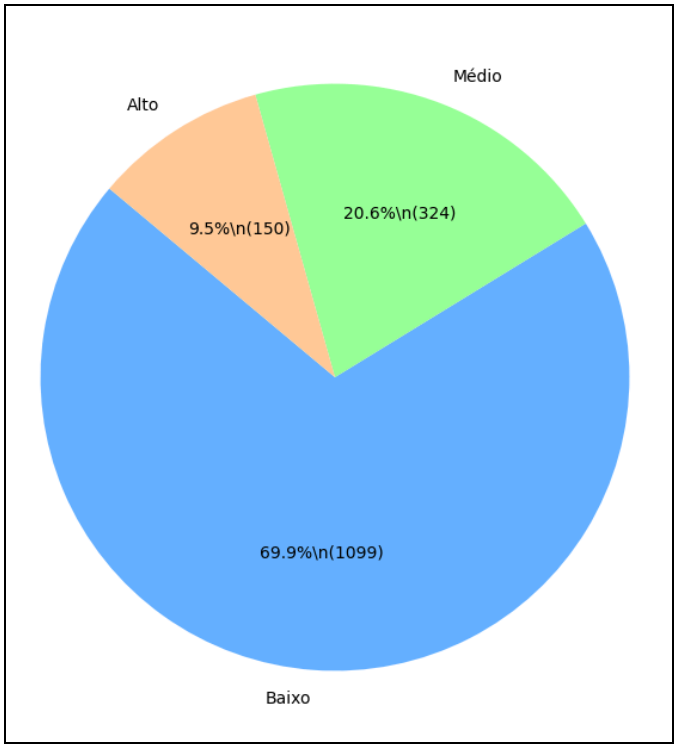
1. **Maximização da Margem:** O princípio fundamental do SVM é encontrar o hiperplano que não apenas separa as classes, mas que também maximiza a distância (margem) entre os pontos de dados mais próximos de cada classe. Essa maximização da margem confere ao modelo uma maior robustez contra overfitting e uma melhor capacidade de generalização para novos dados. Enquanto outros modelos como a Regressão Logística ou a Árvore de Decisão podem encontrar uma fronteira que separa os dados, o SVM encontra a "melhor" fronteira em termos de margem.
2. **Efetividade em Espaços de Alta Dimensão:** O SVM é particularmente eficaz em

espaços de características de alta dimensão, que é o caso deste projeto após a aplicação do One-Hot Encoding nas variáveis categóricas.

3. **Versatilidade com Kernels:** Através do "truque do kernel" (kernel trick), os SVMs podem modelar fronteiras de decisão não-lineares de forma muito eficiente. Embora os dados neste projeto pareçam ser linearmente separáveis (dado o desempenho perfeito de modelos lineares como LR e LDA), ter um modelo que é inerentemente capaz de lidar com complexidade não-linear (usando o kernel 'rbf', que é o padrão) o torna uma escolha mais geral e segura.

Em resumo, embora as métricas de teste sejam idênticas, o **SVM foi escolhido por sua robustez teórica e sua maior probabilidade de generalizar bem para o mundo real**, uma característica que não é totalmente capturada pelas métricas de um único conjunto de teste.

5.3 Distribuição das Classes no Dataset



Legenda: Proporção de cada classe (Baixo: 1100, Médio: 324, Alto: 150) para verificar o nível de desbalanceamento dos dados, o que é importante ao avaliar a robustez do F1-Score.

Para garantir uma avaliação justa e robusta do modelo de classificação, é fundamental analisar a distribuição da variável alvo (Nível de Estresse) no conjunto de dados. Esta análise permite identificar se o dataset apresenta algum grau de desbalanceamento entre as classes.

A distribuição das classes no conjunto de teste (utilizado para a avaliação final) é a seguinte:

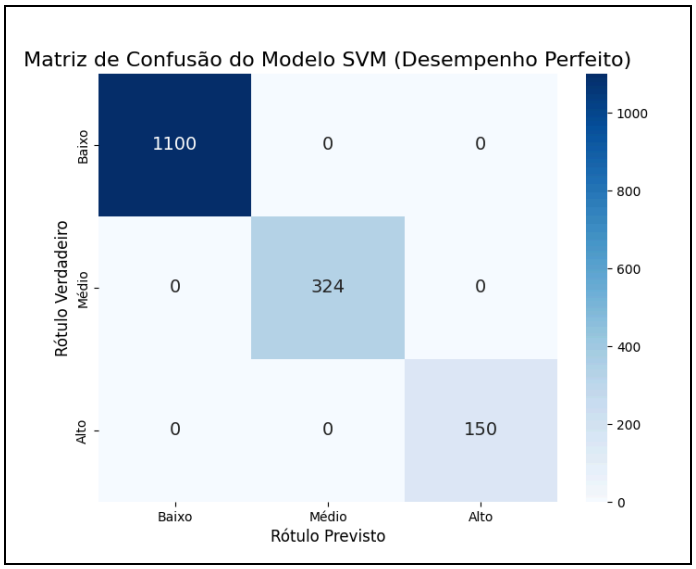
Classe (Nível de Estresse)	Número de Amostras (Suporte)	Proporção Relativa
Baixo (Classe 0)	1100	70.01%
Médio (Classe 1)	324	20.58%
Alto (Classe 2)	150	9.41%
Total	1574	100.00%

5.4 Interpretação

O dataset exibe um desbalanceamento significativo, onde a classe 'Baixo Estresse' é majoritária, representando mais de 70% das amostras, enquanto a classe 'Alto Estresse' é a minoritária, com menos de 10%.
Implicações para a Avaliação:

1. **Acurácia vs. F1-Score:** Em um cenário de desbalanceamento, a Acurácia Global pode ser enganosa, pois um modelo pode atingir alta acurácia simplesmente classificando a maioria das amostras como a classe majoritária. Por este motivo, o uso de métricas como F1-Score, Precisão e Recall (detalhadas no item 4.2) se torna crucial, pois elas penalizam o desempenho em classes minoritárias, fornecendo uma visão mais honesta da capacidade preditiva do modelo em todas as categorias de estresse.
2. **Robustez do Modelo:** O desempenho perfeito (F1-Score de 1.00) obtido pelo SVM em *todas as classes* (inclusive nas minoritárias 'Médio' e 'Alto'), conforme demonstrado no Relatório de Classificação, valida a robustez do modelo, provando que ele não apenas "acertou" a classe majoritária, mas conseguiu separar com precisão o estresse em níveis mais críticos, mesmo com um número reduzido de amostras.

5.5 Matriz de Confusão do Modelo SVM (Heatmap)



Legenda: Uma representação visual da Matriz de Confusão do Modelo SVM, confirmando que a diagonal principal está totalmente preenchida e os demais valores são zero (desempenho perfeito).

A Matriz de Confusão é uma ferramenta fundamental na avaliação de modelos de classificação, pois oferece uma visão detalhada do desempenho do modelo em cada classe. Neste projeto, a matriz para o modelo SVM no conjunto de testes de 1574 amostras revelou um desempenho perfeito.

5.5.1 Interpretação dos Dados:

- **Verdadeiros Positivos (VP) na diagonal:** Todos os valores caem na diagonal principal da matriz, indicando que todas as 1574 amostras do conjunto de teste foram classificadas corretamente.
    - 1100 amostras de Estresse Baixo (Classe 0) foram classificadas corretamente como Baixo.
    - 324 amostras de Estresse Médio (Classe 1) foram classificadas corretamente como Médio.
    - 150 amostras de Estresse Alto (Classe 2) foram classificadas corretamente como Alto.
- Ausência de Erros (Fora da diagonal): Todos os valores fora da diagonal principal são zero.
    - Falsos Positivos (FP): O modelo não classificou incorretamente nenhuma amostra de outra classe como Baixo, Médio ou Alto.
    - Falsos Negativos (FN): O modelo não deixou de identificar nenhuma amostra real de Estresse Baixo, Médio ou Alto.

O resultado da Matriz de Confusão é a evidência primária que sustenta as métricas de 100% de Acurácia, Precisão, Recall e F1-Score para todas as classes, confirmando a alta robustez e confiabilidade do classificador SVM treinado.

6. Conclusão Final

O modelo final implementado é um SVM, que demonstrou ser o classificador ideal para o problema de classificação de Nível de Estresse do projeto CAFEZEN. A performance de 100% de Acurácia e F1-Score no conjunto de teste valida a qualidade dos dados sintéticos e a eficácia do pré-processamento (especialmente o uso do StandardScaler). O modelo foi serializado (junto com o *scaler*) e está pronto para ser consumido pela API Node.js, concluindo a integração do módulo de Inteligência Artificial no sistema.