

Processo Seletivo	00044/2025 - Bolsista Pesquisador - Projetos e Pesquisa – Residência em IA	Etapa	Estudo de caso
Entidade	Centro Universitário SENAI/SC – Campus Florianópolis	Data	07/11/2025

Dados a serem preenchidos pelo Candidato(a):

Nome Completo	Dayane da Silva Xavier Schweitzer		
E-mail	falecomigodayane@gmail.com	CPF	04769847939

1. Contextualização do problema

A empresa agrícola, localizada em Santa Catarina, recebe grandes volumes de maçãs de diferentes produtores e precisa avaliar rapidamente a qualidade dos frutos para orientar decisões de classificação, pagamento e destinação do lote. Inicialmente, esse processo era baseado em imagens de câmeras, o que limita a avaliação a características externas do fruto. Buscando identificar também variações internas e defeitos não visíveis, a empresa investiu em espectrofotômetros que medem a refletância do fruto no intervalo de 780 a 1080 nm. Entretanto, os equipamentos adquiridos possuem granularidades diferentes de pontos de leitura (por exemplo, 96, 97, 100, 106 e 115 pontos), o que dificulta a comparação direta dos dados e o uso de um modelo único de classificação.

2. Problema a ser solucionado

Como padronizar e aproveitar leituras de espectrofotômetros com granularidades diferentes, de forma a construir um modelo automatizado capaz de classificar a qualidade das maçãs em cinco classes (1 = saudável; 2–5 = diferentes tipos de variação indesejada), usando um único pipeline de IA?

3. Objetivo Geral

Desenvolver uma prova de conceito que demonstre a viabilidade de utilizar dados espectrais, provenientes de equipamentos com granularidades distintas, para classificar automaticamente a qualidade das maçãs em cinco classes, por meio de um modelo de inteligência artificial.

4. Objetivos específicos

- Padronizar os espectros de diferentes equipamentos em uma grade comum de comprimentos de onda no intervalo 780–1080 nm.

- Integrar os arquivos Classe_1.csv a Classe_5.csv em um dataset unificado, com rótulos de classe.
- Construir um pipeline de pré-processamento, normalização e modelagem adequado a dados espectrais.
- Treinar e avaliar um modelo de IA para classificação multiclasse (cinco classes de qualidade).
- Gerar métricas que demonstrem a viabilidade técnica da abordagem.
- Documentar desafios encontrados e sugerir evoluções futuras para uma solução em produção.

5. Premissas

- As cinco classes fornecidas representam cenários relevantes para o negócio, sendo a Classe 1 associada a frutos saudáveis e as classes 2–5 a variações indesejadas.
- As leituras de cada arquivo .csv são amostras independentes, em que cada linha representa um espectro de uma maçã ou região do fruto.
- As colunas dos arquivos são amostras igualmente espaçadas no intervalo de 780 a 1080 nm, permitindo o uso de interpolação linear para reamostragem.
- Para a PoC, considera-se suficiente trabalhar com um modelo de classificação supervisionada (ex.: Random Forest) treinado com os dados históricos fornecidos.

6. Limitações

- A PoC utiliza apenas os dados fornecidos, sem controle sobre o processo de coleta (posicionamento do sensor, iluminação, lote, safra etc.).
- O modelo baseline é treinado em um único split treino/teste;
- O estudo foca em classificação por espectroscopia, sem combinar imagens RGB ou outros sensores.

7. Riscos

- Risco de overfitting caso o modelo aprenda padrões específicos da base atual e generalize pouco para novos lotes.
- Possíveis mudanças futuras na calibração dos espectrofotômetros podem exigir nova etapa de ajuste ou re-treinamento.
- Interpretação incorreta das classes (por exemplo, se algum defeito for reclassificado pela empresa) pode impactar a utilidade do modelo.
- Caso os dados futuros apresentem granularidades ainda mais diversas, pode ser necessário revisar a estratégia de interpolação e padronização.

8. Arquitetura proposta

A arquitetura proposta organiza a solução em um **pipeline modular**, implementado em Python, com os seguintes componentes principais:

1. Camada de ingestão de dados (`src/data/load_data.py`)

- Leitura dos arquivos Classe_1.csv a Classe_5.csv na pasta data/.
- Cada arquivo é interpretado como leituras espectrais com granularidade própria (número de colunas variável).
- As linhas representam amostras; as colunas, pontos de leitura no intervalo 780–1080 nm.

2. Camada de padronização espectral

- Para cada arquivo, é construído um eixo de comprimentos de onda específico, de acordo com o número de pontos do equipamento.
- Define-se um eixo de referência comum com **100 pontos** igualmente espaçados entre 780 e 1080 nm.
- Cada espectro é reamostrado via **interpolação linear** para esse eixo comum, gerando features wl_000 a wl_099.
- Os dados das cinco classes são concatenados em um único DataFrame, com a coluna class_id indicando a classe (1–5).

3. Camada de pré-processamento (`src/utils/preprocessing.py`)

- Separação entre **features (X)** e **alvo (y)**, considerando apenas as colunas wl_*** como entrada do modelo e class_id como rótulo.
- Aplicação de **normalização min-max por coluna**, trazendo todas as features para a faixa [0, 1].

4. Camada de modelagem (`src/models/baseline.py`)

- Divisão dos dados em treino e teste (80%/20%), com estratificação das classes.
- Treinamento de um modelo **RandomForestClassifier** como baseline de classificação multiclasse.
- Cálculo das métricas de avaliação (acurácia, precision, recall, F1-score por classe).

5. Camada de orquestração (`src/main.py`)

- Coordena a execução das etapas anteriores em sequência:
 - Carregamento + interpolação;
 - Pré-processamento e normalização;
 - Treinamento e avaliação do modelo;
- Exibe no console os resultados da PoC (acurácia global e relatório de classificação).

9. Diagrama de Classe:

[Arquivos Classe_1..5]

 ingestão

[Padronização 780–1080nm + Interpolação p/ 100 pontos]



[Dataset unificado com w1_000..w1_099 + class_id]



[Normalização Min-Max]



[RandomForest Classifier]



[Métricas (acurácia, precision, recall, F1) e insights]

10. Resultados, desafios e trabalhos futuros

A PoC atingiu acurácia de 96,22% no conjunto de teste, com F1-score médio (macro) em torno de 0,97. As classes 2, 4 e 5 foram classificadas com precisão e recall iguais a 1,0, indicando alta separabilidade. As classes 1 (fruto saudável) e 3 também apresentaram desempenho elevado ($F1 > 0,91$), embora com maior confusão relativa entre si, o que é esperado caso compartilhem características espectrais mais próximas. Esses resultados demonstram que, mesmo com granularidades diferentes nos espectrofotômetros, a estratégia de padronização por interpolação, combinada a um modelo de classificação supervisionada, é capaz de discriminar com alta confiança entre maçãs saudáveis e frutos com diferentes variações indesejadas.

Desafios encontrados

- Tratamento da granularidade heterogênea dos espectrofotômetros, exigindo a definição de uma grade comum de comprimentos de onda e a escolha de uma técnica de reamostragem.
- Decisão sobre o número de pontos-alvo (100) e o uso de interpolação linear, balanceando simplicidade e fidelidade dos espectros.
- Ausência, nesta fase, de metadados adicionais (lote, produtor, condições de coleta), que poderiam auxiliar na análise de possíveis fontes de variação.

Ideias para desenvolvimentos futuros

- Avaliar outros modelos de IA (SVM, redes neurais, gradient boosting) e comparar desempenho e custo computacional com o Random Forest.
- Realizar validação cruzada e testes com novos lotes de dados para avaliar a robustez do modelo em cenários reais de produção.