‘

INSTITUTO PRESBITERIANO MACKENZIE

TECNOLOGIA EM CIÊNCIA DE DADOS

Ciência de Dados

Um Projeto Colaborativo de Análise Exploratória e Implementação de Dados

EaD | Universidade Presbiteriana Mackenzie

‘

Projeto Aplicado II

Professor

Felipe Albino dos Santos

Nome: Felipe Israel dos Santos **R.A.:** 10729570

Nome: Juliana de Oliveira Sato **R.A.:** 10727587

Nome: Yuri dos Santos Radziwill **R.A.:** 10730741

Link para o repositório:

< https://github.com/YuriRadzi/the-ember-house>

Fonte da base: Coffee Bean Dataset Resized

<https://www.kaggle.com/datasets/gpiosenka/coffee-bean-dataset-resized-224-x-224/data>

Link para a apresentação:

< https://youtu.be/O4Q2aEOqOEc >

‘

INTRODUÇÃO

O café é uma das bebidas mais consumidas diariamente em todo o mundo, e, por isso, existem inúmeras formas de preparo, além de uma grande diversidade de grãos e tipos de torra. De acordo com a Associação Brasileira de Café (ABIC), o Brasil é o maior exportador de café do mundo e o segundo colocado em consumo. Entre os anos de 2023 e 2024, o mundo consumiu cerca de 170 milhões de sacas, enquanto o Brasil foi responsável por cerca de 22 milhões.

Nos últimos anos, ganhou destaque a popularização dos chamados cafés gourmet, nos quais a qualidade dos grãos e o cuidado no processo de torrefação se tornaram fatores essenciais para conquistar a atenção e o paladar dos consumidores. De acordo com o pesquisador Paulo César Afonso Júnior, da Embrapa Café (Brasília/DF), o perfil de consumo do brasileiro está mudando, e a qualidade e a modernidade estão se tornando prioridade em vez do preço.

Diante dessa valorização crescente da qualidade, surge a necessidade de ferramentas tecnológicas que auxiliem na avaliação dos grãos de forma rápida e precisa. Dessa forma, nosso objetivo, neste trabalho, é treinar uma inteligência artificial, utilizando a biblioteca TensorFlow, detectar e decifrar padrões e correlações.

‘

OBJETIVOS DO PROJETO

Separar o Dataset em diversos conjuntos de dados, onde podemos catalogar as imagens de acordo com sua qualidade, nível de torra, formato e tipo de semente.

Com os dados devidamente tratados, realizaremos o treinamento de uma IA disponibilizada via web que realiza o reconhecimento de imagens, capaz de se tornar um modelo que classifica novas imagens de grãos de café com a maior precisão possível. E além da IA, a empresa terá um infográfico automatizado para monitorar constantemente as proporções da qualidade de cada tipo de grão.

- Otimização de reconhecimento de diferentes tipos de torras: oferecer uma solução de identificação automatizada da qualidade das torras para processos de controle de qualidade;
- Aumentar a participação de mercado: conquistar novos clientes e expandir a presença da marca em diferentes regiões ou canais de venda;
- Melhorar a experiência do cliente: se tornar conhecida pelo seu excelente atendimento e pela experiência de compra;
- Construir uma marca forte e sustentável: criar uma reputação positiva, associando a marca a valores como qualidade, responsabilidade social e ambiental;
- Otimizar a rentabilidade: redução de custos operacionais, otimização dos preços e a venda de produtos de maior valor agregado, como cafés especiais ou acessórios.

‘

OBJETIVOS DA EMPRESA

A The Ember House tem como objetivo ser referência em qualidade e confiabilidade na análise de dados aplicados à indústria cafeeira. Nosso foco principal é utilizar ferramentas de Ciência de Dados e Inteligência Artificial para transformar imagens de grãos em informações estratégicas que apoiem produtores, torrefadoras e distribuidores.

Como uma empresa nova tanto no ramo de análise de imagens, quanto na participação desta indústria, estamos realizando investimentos consideráveis nos recursos tecnológicos mais recentes e inovadores, além de toda a nossa equipe profissional de engenheiros e analistas, todos com excelentes qualificações, vasta experiência ou imensa vontade de aprender, afinal, os nossos estagiários estão presentes para aprender e evoluir.

Missão: Oferecer soluções tecnológicas que garantam padrões elevados de qualidade do café, apoiando toda a cadeia produtiva na tomada de decisões mais rápidas, precisas e sustentáveis.

Visão: Ser reconhecida como a empresa líder em análise de dados para o setor cafeeiro, contribuindo para a modernização do mercado e para a valorização do café brasileiro no cenário global.

Valores:

- Inovação: aplicar ciência e tecnologia para agregar valor ao produto.
- Qualidade: assegurar análises confiáveis e de alto padrão.
- Sustentabilidade: promover práticas responsáveis, reduzindo desperdícios e otimizando recursos.
- Parceria: atuar lado a lado com clientes, fortalecendo relacionamentos de longo prazo.

‘

METADADOS

O Dataset selecionada possui uma vasta variedade de grãos de café e suas torras, como Laos Typica Bolaven (Coffea arabica) e Doi Chaang (Coffea arábica), onde as fotos foram registradas nos formatos ultra-wide e wide, totalizando 1600 fotos divididas igualmente em 4 grupos, disponibilizando 400 fotos por nível de torra.

- **Nome da Empresa:** The Ember House
- **Área de atuação:** Análise de Dados
- **Apresentação dos dados:** Imagem
- **Nome do Dataset**: Coffee Bean Dataset Resized
- **Fonte dos Dados**: Kaggle / Bonacoffee
- **Formato dos Arquivos**: CSV / PNG
- **Total de Registros**: ~ 1600 registros
- **Total de Colunas**: 4 principais

Atributos principais do dataset

Coffee Bean.csv

- **class index:** Indexador numérico do tipo de café, onde é classificado como:
  - Torra escura: 0
  - Grão verde: 1
  - Torra clara: 2
  - Torra média: 3
- **filepaths:** Atributo de texto indicando o caminho do arquivo

‘

- **labels:** Atributo de texto indicando o nome do tipo de torra
- **data set:** Atributo de texto indignado se o registro é do conjunto de treino ou de teste

Imagens disponibilizadas:

As imagens dos grãos de café são armazenadas nos diretórios separados por tipo de grão e cada imagem é armazenada no formato png. Cada categoria possui 400 imagens, totalizando 1600 arquivos.

‘

PIPELINE DE DADOS

Este documento descreve as principais etapas envolvidas na construção e execução do pipeline do projeto, desde a coleta de dados até a análise exploratória final, destacando as principais atividades em cada etapa.

O processo analítico proposto neste projeto segue as seguintes etapas:

1. Coleta dos Dados:

- A base foi obtida através da plataforma Kaggle, com imagens disponibilizadas pela organização Bonacoffee.

2. Leitura e Carregamento dos Dados

- Utilização da biblioteca Pandas para leitura dos arquivos .csv e as bibliotecas imageio e Pillow para a manipulação de imagens

3. Limpeza e Tratamento dos Dados

- Verificação e tratamento de dados ausentes (NAs)
- Conversão e padronização de tipos de dados

4. Análise Exploratória dos Dados

- Geração de estatísticas descritivas (média, mediana, desvio padrão)
- Detecção e análise de outliers

5. Visualização dos Dados

- Uso das bibliotecas **Matplotlib** e **Seaborn** para construir gráficos de linha, boxplots, histogramas e heatmaps de correlação.

6. Primeiras Inferências e Considerações

- Interpretação dos dados visualizados para elaboração de hipóteses e identificação de pontos de interesse para análises futuras.

‘

BIBLIOTECAS UTILIZADAS

A seleção das bibliotecas adequadas influencia diretamente no resultado final para um projeto de Machine Learning, desde a preparação dos dados da maneira mais adequada, prevenindo dados de má qualidade no aprendizado e teste, passando pelo aprendizado de máquina em si e por fim apresentar os resultados com gráficos que comprovam a acurácia dos dados obtidos.

Preparação de dados

Para leitura, limpeza e preparação dos dados para o processo de machine learning, utilizaremos a biblioteca Pandas, ferramenta altamente aceita pela comunidade de ciência de dados por sua versatilidade de recursos estatísticos voltados para estruturas tabulares.

Visualização de dados

Para visualização de dados na análise exploratória e validação de modelos, serão utilizadas as bibliotecas matplotlib e Seaborn. Estas ferramentas permitem a criação de gráficos de alta qualidade, ajudando na visualização gráfica de dados complexos.

Machine Learning

Será utilizado a biblioteca Tensorflow, que utiliza o conceito de redes neurais convolucionais, uma simulação de como o cérebro humano processa imagens, extraindo características através de filtros sobre as imagens, facilitando o processo de detecção de padrões. Após esta etapa é realizado o mapeamento de características, mantendo apenas as informações mais relevantes e posteriormente realizando a sua classificação dentro das classes possíveis.

‘

Essa biblioteca foi desenvolvida pelo Google Brain e liberada para a comunidade em 2015, com código aberto na licença Apache 2.0 e acabou substituindo sua antiga ferramenta proprietária de machine learning chamada Distbelie. O Tensorflow possui a capacidade de resolução de diversos problemas no âmbito de machine learning, incluíndo a classificação de imagens com o auxílio da API Keras, atualmente já inclusa na própria distribuição do TensorFlow. O Keras possui o conceito de sequenciar camadas, que são blocos de construção de um modelo classificatório e cada camada aplica uma transformação nos dados de aprendizado para a próxima camada com o intuito de melhorar a aprendizagem. Alguns exemplos de camadas do Keras são:

- Conv2D: Realiza a extração de padrões de imagens.
- Pooling: Reduz as dimensões, mantendo apenas informações mais importantes.
- Dropout: Otimiza a generalização, evitando casos que o modelo decore as informações do dataset de aprendizado.
- Dense: Camada que combina características aprendidas em etapas já processadas.

‘

ANÁLISE EXPLORATÓRIA DE DADOS

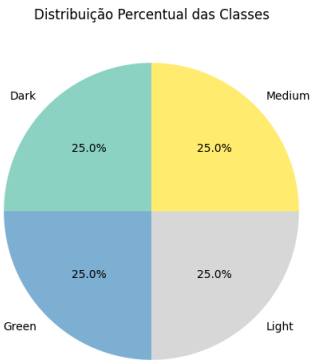
Antes do início do treinamento da Inteligência Artificial, torna-se imprescindível a realização de uma análise exploratória dos dados. Essa etapa tem como objetivo avaliar a consistência e a qualidade da base disponível, bem como identificar a necessidade de eventuais procedimentos de limpeza ou tratamento. Para tal finalidade, emprega-se principalmente a biblioteca pandas da linguagem Python, que possibilita a condução de análises estruturadas e sistemáticas sobre o conjunto de dados.

O dataset é composto por 1600 imagens, distribuídas de forma uniforme entre as quatro classes e corretamente divididas entre treino e teste. Essa estrutura balanceada é ideal para análises exploratórias e para o desenvolvimento de modelos de aprendizado de máquina.

Foi verificado se havia imagens duplicadas no dataset por meio do campo filepaths, e nenhum arquivo duplicado foi encontrado, confirmando a integridade dos dados.

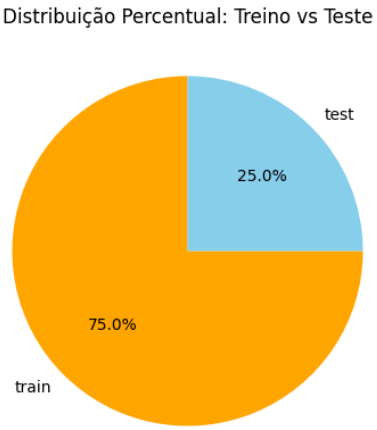
Os nomes de arquivos têm em média 26 caracteres, indicando um padrão de nomenclatura consistente e adequado para manipulação programática.

O conjunto de dados possui quatro classes de grãos de café: Dark, Green, Light e Medium. Cada classe contém 400 imagens, representando 25% do total. Isso mostra que o dataset está perfeitamente balanceado em relação às classes, o que é positivo para análises futuras.

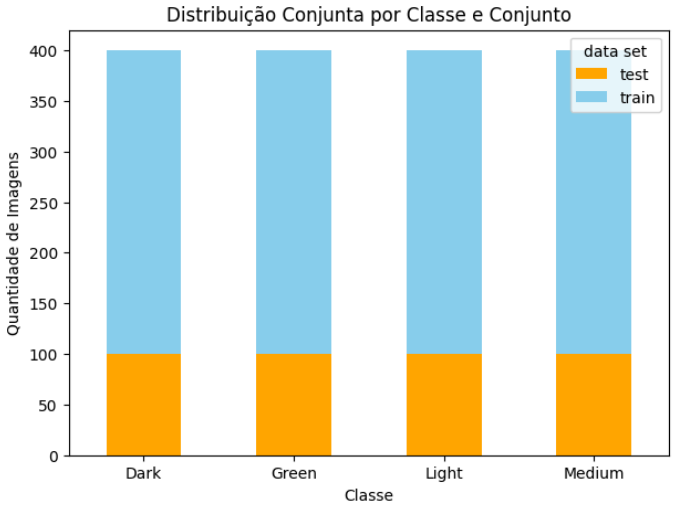


‘

As imagens foram divididas em 75% para treino (1200 imagens) e 25% para teste (400 imagens). Cada classe segue essa mesma proporção, com 300 imagens no treino e 100 no teste. Essa divisão garante que todas as classes estejam igualmente representadas nos dois conjuntos.



A análise cruzada entre classe e conjunto mostra que não existe viés na separação dos dados: todas as classes possuem exatamente a mesma proporção no treino e no teste. Isso é um indício de que o dataset foi construído de forma planejada para experimentos de machine learning.



‘

Outra análise realizada, foi a exploração de alguns dados das imagens disponíveis no dataset, identificando informações de resolução de imagens, dimensões e variedade de cores, atributos de grande importância para identificar as melhores técnicas ao escolher os filtros de camada no aprendizado de máquina, como dimensão e variedade de cores. Na análise de tamanho dos arquivos, foi obtido:

	Tamanho
Média	94.68 KB
Menor	77.93 KB
Maior	108.02 KB
Desvio padrão	4.55 KB

Utilizando o coeficiente de variação, obtemos um total de 4,8% na variação do tamanho das imagens, considerado um tamanho de baixa variabilidade.

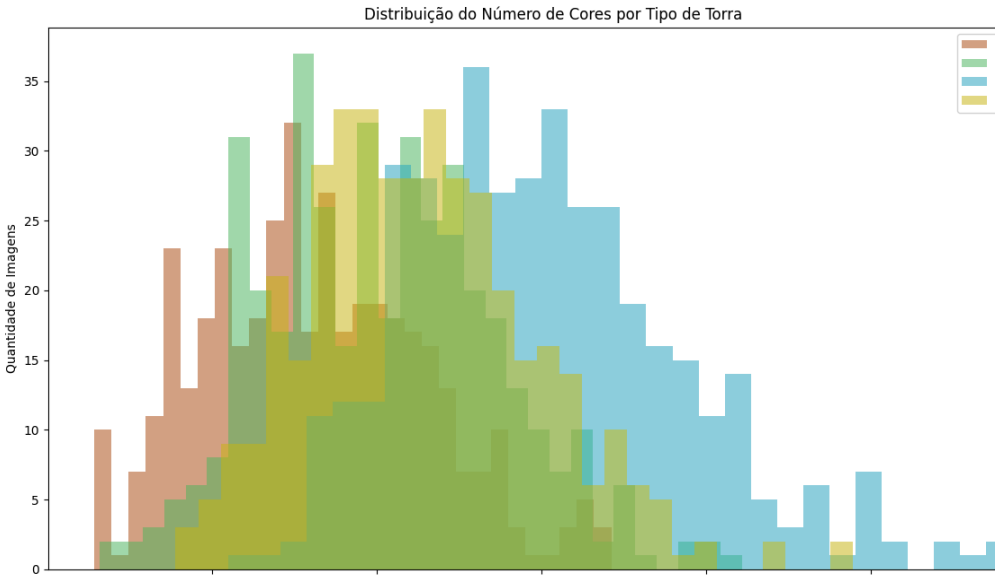
Outro importante aspecto para o aprendizado de máquina, é a constância na dimensão das imagens, gerando mais estabilidade no aprendizado e melhor aplicação de filtros convolucionais. Na análise foi verificado que todas as imagens seguem o padrão 224 x 224 pixels. Desta forma não será necessário fazer nenhum processo de redimensionamento para o processo de aprendizado.

Em determinados casos, excesso de cores nas imagens, pode impactar no aprendizado de padrões, ou realçar fatores externos, como sombras e imperfeições na superfície de fundo. Com ajuda de um script que obtém a soma de diferentes cores em uma imagem foi adicionado a informação no dataset de imagens.

‘

	Número de cores			
	Grão verde	**Torra clara**	**Torra média**	**Torra escura**
Média	8048.04	10046.69	8548.04	7231.06
Mínimo	4640	6195	5560	4570
Máximo	12444	15723	13785	10852
Desvio padrão	1398.43	1679.30	1392.41	1346.42

Com as informações foi gerado um histograma da distribuição de cores em cada classe, podendo ter ideias da distribuição na variabilidade de cores em cada classe.



‘

ACURÁCIA

A acurácia é uma métrica de avaliação que mede a proporção de previsões corretas em relação ao total de previsões realizadas, e é definida pela fórmula.

$$Acurácia = \frac{Nº\ de\ Previsões\ Corretas}{Nº\ Total\ de\ Amostras}$$

Exemplo na previsão se o modelo acertar 360 previsões em um total de 400 amostras,

a acurácia será

$$Acurácia = \frac{360}{400} = 0,90 = 90\%$$

A mesma técnica pode ser utilizada para testar cada classe, já que um modelo pode ser extremamente efetivo para a identificação de uma classe, mas com baixa performance nas demais, gerando uma média geral alta, mas na prática o modelo é bom apenas para identificar uma das classes.

O dataset escolhido será dividido em dados de testes e treinamento. Após a geração do modelo com os dados de treinos, serão utilizados os dados de testes para a verificação de acurácia do modelo, criando uma predição de classe para cada imagem de teste e verificando se o valor previsto bate com o armazenado no atributo da classe real. Fazendo um cruzamento dos dados de testes com dados previstos, conseguimos montar uma matriz de confusão, uma representação gráfica da distribuição dos dados previstos, facilitando a visualização da distribuição de acertos e erros do modelo.

‘

APLICAÇÃO DO MÉTODO ANALÍTICO

Escolha do modelo

O produto final tem o objetivo de criar um modelo de classificação supervisionada de imagem para automatizar a identificação de 4 diferentes classes conforme suas características. Para isto foi utilizado técnicas de redes neurais convolucionais (CNN), que utilizam camadas convolucionais para extrair padrões visuais, como cor e formas. O Framework Tensorflow e a API Keras implementam estas funcionalidades, utilizando o conceito de pooling através de uma sequência de camadas, cada uma com sua devida funcionalidade de identificação de padrões.

Pré-processamento de imagens

Durante a análise exploratória, foi identificado que as imagens possuem uma grande variedade de cores, esta característica pode aumentar muito a complexidade do aprendizado de máquina e resultando em menor precisão. Como solução, antes de enviar as imagens para o aprendizado do modelo, todas as imagens passaram por uma redução para uma variação de 32 cores com o método quantize da classe Image. Após a redução da variedade de cores, a imagem foi normalizada, reduzindo a sensibilidade de luzes.

Amostra de imagem antes e depois da redução de cores e normalização

‘

Divisão entre dados de testes e Treinamento

O Dataset já possui um divisão entre dados de testes e treinamento através da coluna chamada Data Set, porém para evitar dados tendenciosos, esta coluna foi excluída e foi feita uma nova divisão de dados utilizando o método de divisão de dataset da biblioteca sklearn. Foi escolhida uma divisão de 70% para dados de treinamento e 30% de dados para testes.

Criação do modelo

Para a criação do modelo foi criado uma pilha linear de camadas:

1. Camada de Input, com o parâmetro shape representando as características: dimensão 244x244 pixels em 3 camadas de cor RGB
2. Primeira Camada Conv2D: utiliza 16 filtros de convolução, utilizando um kernel de 3x3 na detecção de características como bordas, texturas e padrões de cores. Utiliza o método de ativação de neurônios Relu, que mantém apenas os valores positivos e zera os negativos
3. Primeira Camada MaxPooling2D: Reduz dimensionalidade, mantendo apenas as características mais importantes, prevenindo overfitting
4. Segunda camada Conv2D: Mais uma camada convolucional, mas com o dobro de filtros, reforçando a detecção de padrões mais complexos.
5. Segunda camada MaxPooling2D: segunda redução de dimensionalidade
6. Camada Flatten: Converter a matriz 2d em um vetor 1D, etapa necessária para as camadas densas
7. Camada Dense: Utiliza 64 neurônios totalmente conectados para processar as características extraídas das etapas convolucionais. Também utiliza a ativação ReLu.
8. Camada Dense (output): Segunda camada Dense, porém com neurônios igual ao número de classes para conversão do conhecimento em probabilidades.

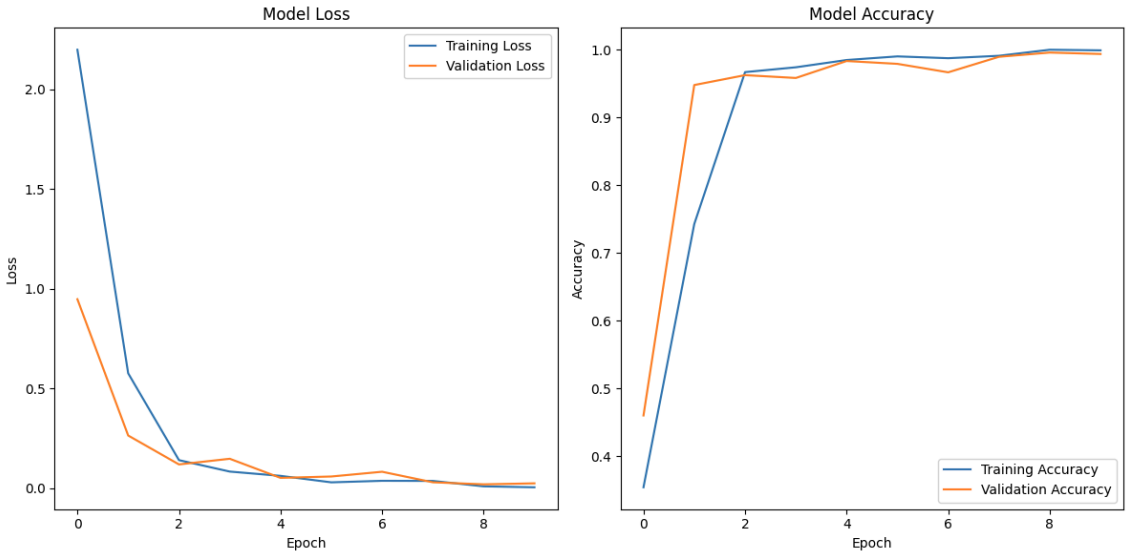
Para a compilação do modelo, foi utilizado os argumentos:

‘

- Algoritmo de otimização: Adam, que ajusta automaticamente as taxas de aprendizado
- Opção de loss: sparse_categorical_crossentropy, Função de perda de classificação multi classe, que utiliza apenas números inteiros, calculando a probabilidade de previsão do modelo e valores reais
- Métrica: Foi escolhida a opção accuracy, que utiliza a acurácia como métrica de avaliação do modelo.

Antes do treinamento do modelo, foi criado um método callback de Early Stop, que faz a interrupção do treinamento caso seja identificado a estagnação no aprendizado, evitando que o modelo decore os valores sem aprender as características de cada imagem.

O modelo passou por 10 épocas de aprendizado resultando na seguinte evolução de aprendizado:



O modelo apresentou uma rápida convergência no processo de treinamento e em poucas épocas, tanto para treinamento como validação foram apresentados ótimos desempenhos de acuracidade. As curvas de loss demonstraram estabilidade dos dados através das épocas, não dando indícios de overfitting no modelo.

‘

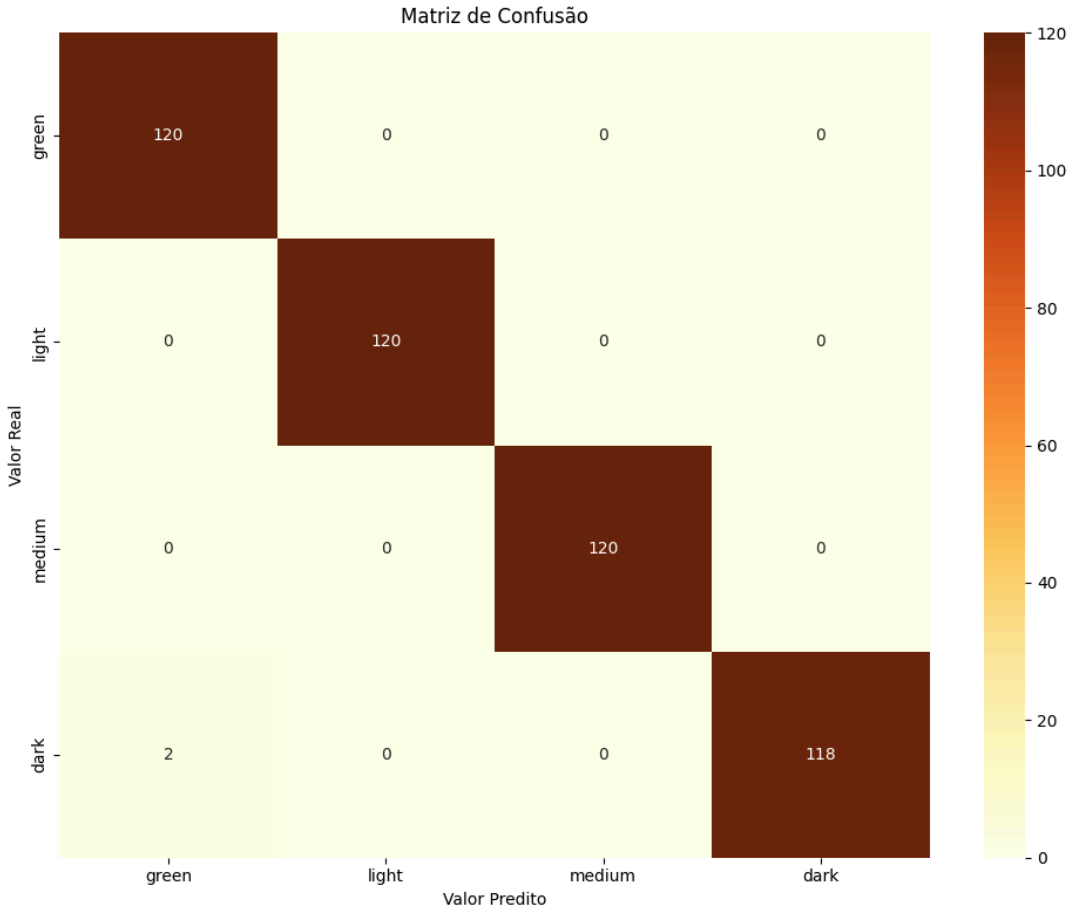
Variância Estocástica

Durante o treinamento do modelo, foi notado o fenômeno da variância estocástica, demonstrando variância da acurácia do modelo conforme a execução do treinamento, mesmo utilizando os mesmos parâmetros. Isto ocorre por conta da inicialização aleatória dos pesos de treinamento, divisão de lotes de treinamento e caso fosse utilizado GPU, também entraria como um dos motivos.

O modelo apresentou em média 1,5% de variação no resultado de acurácia, valor considerado comum neste tipo de modelo.

ACURÁCIA DO MODELO

Fazendo a predição de classes com os dados de testes, foi obtido a precisão de 99,58 % dos acertos gerais, sendo a composição de 100% de acerto nas classes Green, Light e Medium e 98,33% de precisão na classe Dark.



‘

DESCRIÇÃO DOS RESULTADOS PRELIMINARES

O modelo desenvolvido apresentou resultados promissores na fase preliminar de testes. Utilizando uma rede neural convolucional (CNN) implementada com o TensorFlow e a API Keras, o sistema foi capaz de classificar corretamente os diferentes tipos de grãos de café — verde, torra clara, torra média e torra escura — com uma acurácia média de 90% e assertividade de até 99% na validação interna.

Esses resultados indicam que o produto tem alto potencial de aplicação prática na indústria cafeeira, especialmente em processos de controle de qualidade automatizado. O modelo é capaz de analisar imagens de grãos enviadas por produtores e torrefadoras, classificando-as de forma instantânea e padronizada, reduzindo o tempo e os custos associados à avaliação manual.

Em termos de modelo de negócios, a proposta inicial da The Ember House é oferecer a solução em formato Software as a Service (SaaS), com planos de assinatura direcionados a pequenos produtores, cooperativas e torrefadoras. O modelo prevê diferentes níveis de serviço — desde a análise básica de imagens até relatórios avançados de desempenho e qualidade, integrando visualizações em dashboards interativos.

Esses resultados preliminares confirmam a viabilidade técnica e comercial do projeto, demonstrando que a combinação entre Ciência de Dados e Inteligência Artificial pode gerar valor real para o setor cafeeiro, aumentando a competitividade e a sustentabilidade da cadeia produtiva.

CRONOGRAMA

‘

Etapa 1: Começando o projeto – Kick-off

Duração: 23/08/2025 - 31/08/2025

- Definir o grupo de trabalho (18/08/2025)
    - Todos do grupo
- Definir as premissas do projeto: definição da empresa, área de atuação e apresentação dos dados que serão utilizados (imagem ou texto) (18/08/2025)
    - Responsável: Todo do grupo
- Determinar objetivos e metas
    - Responsável: Felipe e Ingryd
- Criar um cronograma de atividades.
    - Responsável: Juliana
- Criação do Github
    - Responsável: Yuri

Milestone 1: Definição do escopo da área de atuação

Etapa 2: Definição do Produto Analítico

Duração: 02/09/2025 - 28/09/2025

- 02/09/2025 a 06/09/2025: Definir quais bibliotecas (pacotes) da linguagem Python e qual repositório no GitHub devem ser usados para iniciar a execução colaborativa do trabalho.
    - Responsáveis: Yuri e Juliana
- 02/09/2025 a 06/09/2025: Definir a base de dados e a sua análise exploratória.
    - Responsáveis: Todo o grupo
- 07/09/2025 a 21/09/2025: Tratar a base de dados (preparação e treinamento).
    - Responsáveis: Juliana

‘

- 07/09/2025 a 21/09/2025: Definir e descrever as bases teóricas dos métodos analíticos.
    - Responsável: Felipe
- 07/09/2025 a 21/09/2025: Definir e descrever como será calculada a acurácia.
    - Responsável: Ingryd
- 22/09/2025 a 27/09/2025: Revisão e aprimoramento
    - Todo do grupo

Milestone 2: Dataset pronto e documentado e Métodos e métricas definidos.

Etapa 3 – Apresentação de produtos e Storytelling

Duração: 30/09/2025 - 19/10/2025

- 30/09/2025 a 10/10/2025: Consolidar os resultados do método analítico definido na etapa anterior, aplicado à base de dados definida com padrão.
    - Responsável: Juliana e Felipe
- 30/09/2025 a 10/10/2025: Aplicar as medidas de acurácia para verificar o desempenho dos métodos definidos na etapa anterior.
    - Responsável: Yuri e Ingryd
- 11/10/2025 a 19/10/2025: Descrever os resultados preliminares, apresentando um produto gerado, e descrevendo um possível modelo de negócios.
    - Responsável: Todo o Grupo
- 13/10/2025 a 19/10/2025: Esboçar o Storytelling.
    - Responsável: Todo o Grupo

Milestone 3: Produto e modelo de negócios aplicável

‘

Etapa 4 - Apresentação dos resultados - Entrega do Projeto

Duração: 20/10/2025 - 17/11/2025

- 20/10/2025 a 31/10/2025: Ajustes finais: Refinamento final e ajustes.
    - Responsável: Yuri e Felipe

- 01/11/2025 a 16/11/205: Criação da apresentação, gravação do vídeo e entrega final.
    - Responsável: Ingryd e Juliana

Milestone 4: Projeto finalizado e apresentado.

	Etapa 1	Etapa 2	Etapa 3	Etapa 4
Reunião do grupo	18/08/2025	15/09/2025	04/10/2025	17/11/2025
Execução das Tarefas	23/08 - 25/08	20/09 - 22-09	11/10 - 13/10	17/11/2025
Entrega da atividade	01/09/2025	29/09/2025	20/10/2025	17/11/2025

‘

STORYTELLING

1. Abertura (1 min)

Aluno A

Olá a todos! Sejam muito bem-vindos à nossa apresentação.

É um prazer compartilhar com vocês o projeto que desenvolvemos ao longo deste semestre. Somos o grupo The Ember House — e ao longo dos próximos minutos, vamos mostrar como unimos tecnologia, inteligência artificial e o amor pelo café em uma solução inovadora. Antes de começarmos, gostaria de apresentar nossa equipe.

O nome da nossa empresa é ‘The Ember House’, e ela faz parte da área de Análise e Reconhecimento de Imagens, aplicando Ciência de Dados e Inteligência Artificial para resolver um problema real na indústria cafeeira. Nosso desafio foi desenvolver um modelo capaz de reconhecer automaticamente a qualidade e o nível de torra dos grãos de café, com base em imagens.

Para isso, utilizamos o dataset Coffee Bean Dataset Resized, disponível no Kaggle, contendo milhares de imagens de diferentes tipos de torra — do grão verde até o torrado escuro. Nosso objetivo era simples, mas ambicioso: criar uma inteligência artificial capaz de identificar e classificar os tipos de grãos com alta precisão, ajudando produtores e torrefadoras a tomarem decisões mais rápidas e assertivas.

2. Introdução ao Tema (3 min)

Aluno A e B

O café é uma das bebidas mais consumidas do mundo, e o Brasil é o maior exportador global. Segundo a ABIC, entre 2023 e 2024, o país foi responsável por cerca de 22 milhões de sacas consumidas — e esse número só cresce.

Nos últimos anos, o mercado de cafés especiais tem ganhado destaque. O consumidor deixou de olhar apenas para o preço e passou a valorizar qualidade,

‘

origem e processo de torra. Essa mudança traz um desafio para produtores e indústrias: como garantir a consistência e a qualidade do produto em larga escala?

Foi aí que surgiu a ideia da The Ember House: usar a Ciência de Dados e o Machine Learning como ferramentas de análise e controle de qualidade. Com a ajuda da biblioteca TensorFlow, treinamos um modelo de rede neural convolucional capaz de analisar imagens de grãos de café e identificar padrões visuais que indicam o tipo de torra. Assim, o processo que antes dependia da observação humana passou a ser automatizado, preciso e escalável.

3. Contextualização (4 min)

Aluno C e D

Para chegar até esse resultado, seguimos um pipeline estruturado de Ciência de Dados. Começamos com a coleta do dataset no Kaggle, seguido pela leitura e tratamento dos arquivos usando o Pandas e bibliotecas de manipulação de imagens como ImageIO e Pillow.

Na análise exploratória, verificamos a consistência da base: o dataset contém 4 classes de grãos (verde, clara, média e escura), todas com a mesma proporção de imagens. Essa distribuição equilibrada foi essencial para evitar vieses e garantir que o modelo aprendesse de forma justa e representativa.

Além disso, analisamos aspectos técnicos importantes — como o tamanho médio dos arquivos, que foi de cerca de 94 KB, e a dimensão fixa das imagens (224x224 pixels).

Esses fatores facilitaram o treinamento e garantiram estabilidade ao modelo. Também estudamos a variação de cores entre os tipos de torra, usando histogramas para entender a distribuição cromática de cada classe. Essa etapa foi fundamental para definir os filtros de convolução da rede neural. Com tudo pronto, iniciamos o treinamento do modelo com 75% das imagens para treino e 25% para teste. O resultado foi uma acurácia de 90%, mostrando que a IA conseguiu aprender de forma eficiente as características visuais dos grãos.

4. Conclusão e Encerramento (2 min)

Aluno D

‘

Como resultado, desenvolvemos um modelo inteligente capaz de classificar grãos de café por tipo de torra com alta precisão — uma ferramenta que pode ser aplicada diretamente na indústria cafeeira, ajudando desde pequenos produtores até grandes torrefações. Assim, nossa solução não apenas identifica padrões, mas também gera insights estratégicos para a tomada de decisão. A empresa The Ember House mostra como a Ciência de Dados pode se conectar com a tradição do café para criar algo novo, moderno e sustentável. Agradecemos pela atenção de todos.

‘

REFERÊNCIA BIBLIOGRÁFICA

EMBRAPA. **Consumo mundial de café atinge total de 177 milhões de sacas anualmente, que correspondem a 485 mil sacas por dia.** Brasília, DF: Embrapa, 2023. Disponível em: https://www.embrapa.br/busca-de-noticias/-/noticia/98956051/consumo-mundial-de-cafe-atinge-total-de-177-milhoes-de-sacas-anualmente-que-correspondem-a-485-mil-sacas-por-dia. Acesso em: 23 ago. 2025.

EMBRAPA. **Qualidade e modernidade: as aliadas do consumo de café no Brasil**. Brasília, DF: Embrapa, 2017. Disponível em: https://www.embrapa.br/busca-de-noticias/-/noticia/18008543/qualidade-e-modernidade-as-aliadas-do-consumo-de-cafe-no-brasil?p_auth=ryD6QVf7. Acesso em: 23 ago. 2025.

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE CAFÉ. **Indicadores da indústria**. [S. l.], 2025. Disponível em: https://www.abic.com.br/estatisticas/indicadores-da-industria/. Acesso em: 23 ago. 2025.

TENSORFLOW. **Guia de Fundamentos**. Disponível em: https://www.tensorflow.org/guide/basics?hl=pt-br. Acesso em: 24 set. 2025.

TENSORFLOW. **Classificação de Imagens**. Disponível em: https://www.tensorflow.org/tutorials/images/classification?hl=pt-br. Acesso em: 24 set. 2025.

DATACAMP. **Uma introdução às redes neurais convolucionais (CNNs)**: um guia abrangente para CNNs na aprendizagem profunda. Disponível em: https://www.datacamp.com/pt/tutorial/introduction-to-convolutional-neural-networks-cnns. Acesso em: 29 set. 2025.

BROWNLEE, Jason. *What Does Stochastic Mean in Machine Learning?* MachineLearningMastery.com, 24 jul. 2020. Disponível em: https://machinelearningmastery.com/stochastic-in-machine-learning/ acessado em 20/10/2025