<table>
<tr><td>Título do plano</td><td colspan="3">Energia, eficiência e escolhas leves: o trade-off entre custo computacional e qualidade do modelo</td><td>Sequência<br>19</td></tr>
<tr><td>Etapa/Ano</td><td colspan="4">Ensino Médio: ☐ 1º Ano | ☐ 2º Ano | ☑ 3º Ano</td></tr>
<tr><td>Nível de Maturidade<br>[link]</td><td>Escola</td><td>☐ Emergente<br>☐ Básico<br>☑ Intermediário<br>☐ Avançado</td><td>Docente</td><td>☐ Básico<br>☑ Intermediário<br>☐ Avançado</td></tr>
<tr><td>Competências<br>[link]</td><td colspan="4">☑ C1 ☐ C2 ☐ C3 ☑ C4 ☑ C5</td></tr>
<tr><td>Objetos de Conhecimento</td><td colspan="4">• Custo computacional de modelos de IA: tempo de treinamento, tempo de inferência, consumo de memória e consumo de energia elétrica.<br>• Pegada de carbono associada ao ciclo de vida de um modelo: matriz energética brasileira, fator de emissão ($gCO_2$/kWh), conversão entre Wh, kWh e $gCO_2$.<br>• Diferenças de custo computacional entre famílias de modelos: árvore de decisão (leve), redes neurais MLP (médias), modelos de linguagem de grande porte / LLMs (pesados).<br>• Trade-off entre qualidade do modelo e eficiência: noção de fronteira de Pareto e princípio de suficiência computacional.<br>• Conceitos de Green AI e Red AI: o modelo mais sofisticado nem sempre é o mais adequado para o problema.<br>• Decisão técnica fundamentada em evidência empírica: como argumentar a escolha de um modelo a partir de medições e estimativas.</td></tr>
<tr><td>Recursos Educacionais</td><td colspan="4">• Pacote do projeto integrador entregue no Plano 15: dataset em CSV, datasheet (Plano 8), arquivo .ows do Orange e relatório do MVP.<br>• Modelos comparados no Plano 14: árvore de decisão (Plano 13) e MLP/rede neural — ambos no mesmo arquivo .ows.<br>• Model Card produzido pelos grupos no Plano 18.<br>• Computador ou tablet com Orange instalado, um por grupo.<br>• Calculadora simples ou planilha eletrônica (Google Sheets, Excel) para conversão entre Wh, kWh e $gCO_2$.<br>• Folha de referência impressa<br>• Folha de trabalho "Tabela Comparativa de Custos"<br>• Template do Adendo de Eficiência (uma cópia por grupo).<br>• Cronômetro.</td></tr>
<tr><td>Componentes Curriculares Relacionados</td><td>☐ Arte<br>☑ Língua Port.<br>☐ Ed. Física</td><td>☐ Geografia<br>☐ História<br>☑ Matemática</td><td colspan="2">☐ Língua Inglesa<br>☐ Ciências<br>☑ Computação</td></tr>
<tr><td>Palavras-chave</td><td colspan="4">consumo energético; pegada de carbono; eficiência computacional; trade-off; Green AI; suficiência; árvore de decisão; MLP; LLM; matriz energética brasileira; aprendizagem baseada em projetos</td></tr>
<tr><td>Perguntas Importantes</td><td colspan="4">• Quanta energia foi necessária para treinar o nosso modelo? E quanto custaria, em energia e em emissões de carbono, treinar uma versão muito maior do mesmo sistema?<br>• O modelo mais sofisticado é sempre o melhor? Quando faz sentido escolher um modelo mais simples mesmo que ele seja levemente menos preciso?<br>• Onde estão as escolhas "leves" do nosso projeto — as decisões que economizaram recursos sem comprometer a qualidade — e onde estão as escolhas "pesadas" que poderíamos ter evitado?<br>• Faz sentido usar um LLM para resolver um problema que uma árvore de decisão resolveria? Em que condições sim, em que condições não?<br>• Como o fato de a matriz elétrica brasileira ser predominantemente hidrelétrica muda a conta de pegada de carbono comparada com países como os Estados Unidos?</td></tr>
</table>

| | | | |
|---|---|---|---|
| | • Que escolhas o nosso grupo recomenda formalmente para a versão final do projeto, e por quê? | | |
| **Encontros** | 1 | **Total de Horas-aula** | 2 |
| **Objetivos** | • Estimar empiricamente, com apoio do Orange, o custo computacional de treinar e usar os modelos do projeto integrador (árvore de decisão e MLP), traduzindo tempo em watts-hora e em gramas de $CO_2$ equivalente.<br>• Comparar o custo medido com valores de referência conhecidos para sistemas baseados em LLMs, evidenciando ordens de grandeza diferentes entre famílias de modelos.<br>• Reconhecer e nomear o trade-off entre qualidade do modelo e eficiência energética como uma decisão de projeto, não como uma propriedade técnica neutra.<br>• Aplicar o princípio de suficiência computacional para fundamentar a escolha do modelo a ser entregue na versão final do projeto (Plano 20).<br>• Produzir o Adendo de Eficiência ao Model Card produzido no Plano 18, integrando medições empíricas, estimativas comparativas e justificativa da decisão final de modelo.<br>• Articular a discussão sobre eficiência energética com a especificidade da matriz energética brasileira, reconhecendo o contexto local como variável relevante na pegada de carbono de sistemas de IA. | | |
| **Habilidades Relacionadas** | **BNCC:**<br>(EM13CNT301) Construir questões, elaborar hipóteses, previsões e estimativas, empregar instrumentos de medição e representar e interpretar modelos explicativos, dados e/ou resultados experimentais para construir, avaliar e justificar conclusões no enfrentamento de situações-problema sob uma perspectiva científica.<br>(EM13CNT206) Discutir a importância da preservação e conservação da biodiversidade, considerando parâmetros qualitativos e quantitativos, e avaliar os efeitos da ação humana e das políticas ambientais para a garantia da sustentabilidade do planeta.<br>(EM13MAT203) Aplicar conceitos matemáticos no planejamento, na execução e na análise de ações envolvendo a utilização de aplicativos e a criação de planilhas para tomar decisões.<br>(EM13MAT302) Construir modelos empregando as funções polinomiais de 1º ou 2º graus, para resolver problemas em contextos diversos, com ou sem apoio de tecnologias digitais.<br>(EM13LP12) Selecionar informações, dados e argumentos em fontes confiáveis, impressas e digitais, e utilizá-los de forma referenciada, para que o texto a ser produzido tenha qualidade técnica e ética.<br><br>**BNCC Computação:**<br>(EM13CO03) Identificar o comportamento dos algoritmos no que diz respeito ao consumo de recursos como tempo de execução, espaço de memória e energia, entre outros.<br>(EM13CO05) Identificar os limites da Computação para diferenciar o que pode ou não ser automatizado, buscando uma compreensão mais ampla dos limites dos processos mentais envolvidos na resolução de problemas.<br>(EM13CO10) Conhecer os fundamentos da Inteligência Artificial, comparando-a com a inteligência humana, analisando suas potencialidades, riscos e limites. | | |
| **Práticas Pedagógicas Inovadoras** | ☑ Aula Enriquecida com Tecnologia<br>☐ Ensino Híbrido: Sala de aula invertida<br>☐ Ensino Híbrido: rotação por estação<br>☐ Ensino Híbrido: rotação individual<br>☐ Aulas Mão na massa<br>☑ Aprendizagem baseada em projetos<br>☐ IA Desconectada | | |

## Sequência Didática

Este encontro é o quarto do Módulo 4 (Reflexão) e opera novamente em formato de Aprendizagem Baseada em Projetos, agora combinada com Aula Enriquecida com Tecnologia, porque a etapa central do plano exige medição empírica direta no Orange. O entregável é o Adendo de Eficiência ao Model Card — não um documento novo, mas uma seção complementar que se anexa ao Model Card produzido no Plano 18. Essa decisão de desenho reforça que o Módulo 4 é um único projeto de documentação responsável que se sofistica a cada encontro: auditar (16), analisar vieses (17), documentar (18), avaliar custos (19), versionar (20). A turma se mantém nos mesmos grupos dos encontros anteriores, em parte porque o trabalho deste plano depende do .ows do Plano 14 (modelos comparados) e do Model Card do Plano 18, e em parte porque a revisão entre pares praticada no Plano 18 ganha aqui uma forma mais leve mas presente: as decisões de modelo são compartilhadas em microapresentação no fechamento.

### Fase 1 — Enquadramento: por que medir o custo do nosso próprio modelo (≈15 min)

Inicie o encontro retomando, em três a quatro minutos, o que a turma já sabe sobre energia e IA. No Tema 11 do Ano 2, conduzido pela mesma equipe Claudia/Aline/Ricardo, foi trabalhada a pegada de carbono do uso pessoal de IA: cada inferência em um chatbot consome aproximadamente 0,34 Wh, cada imagem gerada por IA consome cerca de 11 Wh, e a matriz elétrica brasileira em 2023 emitiu, segundo o MCTI, 38,5 gramas de $CO_2$ por kWh — número baixo na comparação internacional, graças à predominância hidrelétrica. Esse vocabulário e esses números voltam hoje, mas com uma diferença importante: no Ano 2, a pergunta era genérica ("quanto pesa o meu uso diário de IA?"). No Ano 3, a pergunta é específica ("quanto pesou construir e usar o sistema que NÓS criamos?"). É a passagem do consumo de IA para a produção de IA — e, pedagogicamente, é a passagem do estudante como usuário para o estudante como projetista responsável.

Em seguida, introduza o conceito central da aula: o trade-off entre qualidade do modelo e eficiência computacional. Escreva no quadro três palavras-chave alinhadas: árvore de decisão, MLP (rede neural), LLM (modelo de linguagem de grande porte) e o que cada um significa em termos de "peso":

- **Árvore de decisão (leve):** treina em segundos em um computador comum, ocupa pouca memória, é interpretável. É o modelo do Plano 13, que vocês construíram.
- **MLP / rede neural (média):** treina em segundos a minutos para datasets pequenos, ocupa mais memória, perde explicabilidade. É um dos modelos comparados no Plano 14.
- **LLM / modelo de linguagem de grande porte (pesado):** treinar uma versão de referência como o BERT-base custou aproximadamente 1.438 kWh em estudos

publicados, e modelos como o GPT-3 consumiram cerca de 1.287 megawatts-hora durante o treinamento. Não treinamos um, mas usamos como contraponto numérico.

Apresente o quadro conceitual Green AI versus Red AI, formulado por Schwartz e colegas em 2020. A lógica de "Red AI" é a busca por desempenho a qualquer custo computacional — modelos cada vez maiores, com cada vez mais dados e mais energia, com ganhos marginais de qualidade. A lógica de "Green AI" é o oposto: tratar a eficiência computacional como uma métrica de qualidade, não como uma restrição secundária. Em outras palavras, escolher o modelo mais leve que resolve o problema, e não o mais sofisticado disponível. Apresente, em uma frase, o princípio de suficiência computacional: “o melhor modelo é o modelo suficiente, não o modelo máximo”. Esse princípio orienta toda a aula.

Encerre a abertura com a missão do encontro: cada grupo vai medir empiricamente o custo de treinar a árvore de decisão e a MLP do seu projeto, estimar comparativamente o custo de uma alternativa baseada em LLM, montar uma tabela de trade-offs entre qualidade e eficiência, e — com base nesses dados — recomendar formalmente, no Adendo de Eficiência, qual modelo a turma deve entregar como versão final no Plano 20. A decisão precisa ser tecnicamente fundamentada e ambientalmente consciente.

### Fase 2 — Medição empírica em Orange e estimativa de pegada (≈40 min)

Esta fase é o coração da Aula Enriquecida com Tecnologia. Cada grupo abre, em seu computador ou tablet, o arquivo .ows do projeto, contendo a árvore de decisão (Plano 13) e a MLP (Plano 14). A tarefa é medir empiricamente o custo computacional dos dois modelos e converter a medição em estimativa de pegada de carbono.

**Template Procedimento sugerido (impresso para cada grupo) (≈25 min de execução, ≈15 min de cálculo e registro):**

Durante esta fase, o(a) docente circula entre os grupos com atenção especial a duas questões frequentes. A primeira é a tendência de minimizar a diferença observada ("é só meio segundo"). Reforce que o que importa pedagogicamente é a ordem de grandeza, não o valor absoluto: a MLP custou dezenas de vezes mais energia que a árvore para um ganho de acurácia que pode ser pequeno. A segunda é a tendência inversa, de exagerar a diferença sem contexto. Lembre que valores absolutos baixos (frações de grama de $CO_2$) só ficam preocupantes quando multiplicados por milhões de execuções diárias — que é exatamente o caso de sistemas em produção.

### Fase 3 — Estimativa comparativa com LLM e tabela de trade-offs (≈30 min)

Com a primeira coluna da tabela preenchida com dados empíricos (árvore e MLP), o grupo agora estima por analogia o que custaria resolver o mesmo problema com um modelo de linguagem de grande porte. Esta fase não é técnica de medição: é exercício de leitura de literatura aplicada e de raciocínio por ordens de grandeza.

**Template da Tabela Comparativa de Custos (folha de referência fornecida pelo professor para cada grupo):**

Com a tabela preenchida, o grupo discute a leitura comparativa em três a quatro minutos e responde por escrito, no rodapé do template, as três perguntas-chave.

É no terceiro item que o conceito de fronteira de Pareto se torna visível para a turma: em escala pequena, a diferença absoluta entre uma árvore e uma MLP é desprezível; em escala grande, a mesma diferença pode equivaler a toneladas de $CO_2$ por ano. A decisão de modelo, portanto, depende do contexto de uso, não apenas do desempenho em laboratório.

### Fase 4 — Decisão de modelo e Adendo de Eficiência ao Model Card (≈25 min)

A última fase consolida o trabalho do encontro em um documento curto e formal: o **Adendo de Eficiência ao Model Card**. O Adendo não é um documento independente — é uma seção que se anexa ao Model Card produzido no Plano 18, ampliando-o. A turma deve ter os dois documentos juntos sobre a mesa: Model Card de ontem, Adendo de hoje.

**Estrutura do Adendo de Eficiência (template fornecido)**

O grupo dispõe de aproximadamente 15 minutos para preencher o Adendo, com o Model Card do Plano 18 ao lado. Reforce que o Adendo deve usar a mesma linguagem cuidadosa do Model Card: honesta com os limites, transparente com a fonte dos números, acessível para um leitor não-técnico.

**Microcompartilhamento (≈8 min):**

Cada grupo apresenta, em até 75 segundos, três pontos de seu Adendo: (a) qual modelo escolheu para a versão final, (b) qual foi o argumento decisivo da escolha em termos de trade-off, e (c) uma recomendação de uso eficiente que considera a mais original do grupo. O(a) docente anota termos-chave no quadro, formando um pequeno mosaico do que a turma como um todo está recomendando para o projeto.

**Conexão para o Plano 20 (≈2 min):**

Encerre indicando o que muda no horizonte. No próximo e último encontro do módulo (Plano 20), a turma vai consolidar a versão final do projeto integrador, incorporando todas as decisões dos Planos 16 a 19: a auditoria do fluxo, a Carta de Mitigação, o Model Card e o Adendo de Eficiência. Será o momento de decidir, coletivamente, o versionamento do sistema (qual versão entregar, qual data de revisão, qual ciclo de atualização) e de fazer uma retrospectiva crítica do projeto inteiro. Recolha os Adendos finalizados e fixe-os, junto com os Model Cards, no mural ou pasta digital da turma.

### Resumo temporal do encontro (≈120 min)

| Tempo | Fase do PBL | O que acontece e produto da fase |
|---|---|---|

| | | |
|---|---|---|
| ≈15 min | **Enquadramento** | Recuperação dos números do Tema 11 (Ano 2). Apresentação do trade-off entre qualidade e eficiência. Quadro Green AI vs Red AI. Princípio de suficiência computacional. Definição da missão. |
| ≈40 min | **Medição empírica em Orange** | Aula com Tecnologia: medição de tempo de treinamento e inferência da árvore de decisão e da MLP no .ows do Plano 14. Conversão para Wh e $gCO_2$ usando o factor MCTI. Preenchimento da primeira parte da Tabela Comparativa. |
| ≈30 min | **Estimativa comparativa** | Estimativa por ordens de grandeza para um LLM hipotético, com base em Strubell, Patterson e referências do Tema 11. Tabela completa com cinco linhas (tempo, energia, pegada, acurácia, explicabilidade). Discussão do trade-off em pequena e grande escala. |
| ≈25 min | **Adendo de Eficiência + Síntese** | Preenchimento do Adendo (medições, decisão de modelo justificada, recomendações de uso eficiente). Anexação ao Model Card do Plano 18. Microcompartilhamento de 75 s por grupo. Gancho para o Plano 20. |
| ≈10 min | ***Folga de manejo*** | *Margem para transições, gerenciamento de turma e ajustes técnicos no Orange. Recomenda-se preservá-la: a Fase 2 frequentemente exige mais tempo do que o estimado, especialmente em turmas com pouca familiaridade com a ferramenta.* |

## Avaliação

A avaliação é processual e baseada em evidências de medição, raciocínio comparativo e justificativa de decisão. Em um plano que articula PBL com investigação empírica, a qualidade da medição e a qualidade do argumento têm peso comparável. Critérios e peso:

- **Medição empírica e cálculo de pegada — Fase 2 (20%):** registros consistentes de tempo de treinamento e inferência; conversão correta de Wh para kWh e para $gCO_2$ com o fator MCTI; documentação clara do método (potência assumida, fonte do factor de emissão).
- **Tabela Comparativa de Custos — Fases 2 e 3 (20%):** tabela completa com as cinco linhas (tempo, energia, pegada, acurácia, explicabilidade) e três colunas (árvore, MLP, LLM hipotético). Estimativas para LLM em ordem de grandeza defensável e com fonte declarada.
- **Adendo de Eficiência — Bloco 2 (decisão de modelo justificada) (25%):** argumento técnico e ambiental claro para a escolha do modelo da versão final; uso explícito do princípio de suficiência computacional; reconhecimento honesto do que se ganha e do que se perde com a escolha.
- **Adendo de Eficiência — Bloco 3 (recomendações de uso eficiente) (15%):** três recomendações específicas ao projeto, não genéricas, focadas em eficiência operacional do sistema.
- **Articulação com o Model Card do Plano 18 (10%):** Adendo coerente em linguagem, formato e nível de detalhe com o Model Card existente; anexação visível dos dois documentos como pacote único.
- **Postura crítica e ambiental durante o encontro (10%):** observação docente sobre engajamento na medição empírica, qualidade da discussão de trade-offs e capacidade de articular custo computacional com responsabilidade ambiental.

Recomenda-se que a rúbrica seja apresentada à turma no início do encontro e impressa em cada mesa.

## Material Complementar

**Sobre custo energético de modelos de IA:**

STRUBELL, Emma; GANESH, Ananya; MCCALLUM, Andrew. Energy and Policy Considerations for Deep Learning in NLP. Proceedings of ACL, 2019. Estudo seminal sobre o custo energético de treinar modelos como BERT. Disponível em: https://aclanthology.org/P19-1355/

SCHWARTZ, Roy; DODGE, Jesse; SMITH, Noah A.; ETZIONI, Oren. Green AI. Communications of the ACM, v. 63, n. 12, 2020. Texto que cunhou o termo "Green AI" e propôs eficiência como métrica de qualidade. Disponível em: https://cacm.acm.org/research/green-ai/

PATTERSON, David et al. Carbon Emissions and Large Neural Network Training. arXiv:2104.10350, 2021. Atualização com dados sobre treinamento de LLMs maiores como GPT-3. Disponível em: https://arxiv.org/abs/2104.10350

HENDERSON, Peter et al. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. Journal of Machine Learning Research, 2020. Disponível em: https://jmlr.org/papers/v21/20-312.html

LACOSTE, Alexandre et al. Quantifying the Carbon Emissions of Machine Learning. Calculadora online ML $CO_2$ Impact, 2019. Útil para estimar pegada de treinamentos hipotéticos. Disponível em: https://mlco2.github.io/impact/

**Matriz energética brasileira e fator de emissão:**

BRASIL. Ministério da Ciência, Tecnologia e Inovação (MCTI). Fator de emissão de $CO_2$ na geração de energia elétrica no Brasil em 2023. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/02/fator-de-emissao-de-co2-na-geracao-de-energia-eletrica-no-brasil-em-2023-e-o-menor-em-12-anos

EPE — Empresa de Pesquisa Energética. Anuário Estatístico de Energia Elétrica. Atualização anual sobre composição da matriz brasileira. Disponível em: https://www.epe.gov.br/

INICIATIVA VERDE. Calculadora de carbono. Disponível em: https://iniciativaverde.org.br/calculadora

SOS MATA ATLÂNTICA. Calculadora de emissão de $CO_2$. Disponível em: https://www.sosma.org.br/calcule-sua-emissao-de-co2