
Faculdade
XPe
RELATÓRIO
PROJETO
APLICADO
PÓS-GRADUAÇÃO

XP Educação

Relatório do Projeto Aplicado

Banco de dados para cálculo de inflação de produtos

Victor Augusto Pereira Burgardt

Orientador(a): Ítalo Lucena

[Data]


XPe

Victor Augusto Pereira Burgardt

XP EDUCAÇÃO

RELATÓRIO DO PROJETO APLICADO

TÍTULO DO PROJETO

Relatório de Projeto Aplicado desenvolvido para fins de conclusão do curso [...].

Orientador (a): Ítalo Lucena

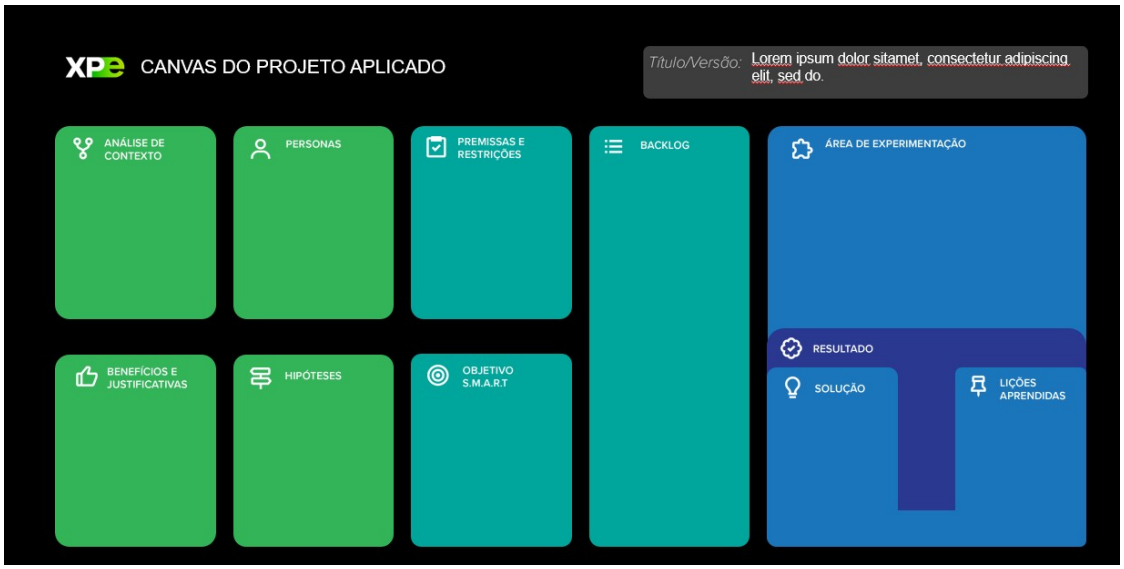
Rio de Janeiro

[Data]

Sumário

1. CANVAS do Projeto Aplicado

Figura conceitual, que representa todas as etapas do Projeto Aplicado.



1.1 Desafio

1.1.1 Análise de Contexto

Nos últimos anos, temos vivenciado um período considerável de instabilidade causada por diferentes crises que vão de origens climáticas, humanitárias, tensões políticas, sanitárias, econômicas até pandemias e guerras. Atualmente, pessoas tem até utilizado o termo *"poly crisis"* para descrever o período atual, que é denotado como um período em que ocorrem múltiplas eventos catastróficos de diversas naturezas em um certo período. Entres as diversas crises mundiais atuais as que mais estão marcando esse período e estão desestabilizando muito o planeta são:

- Pandemia do Corona Vírus
- Crise das redes sociais e Fake News
- Guerra da Rússia e Ucrânia
- Guerra de Israel e Palestina
- Crises climáticas
- Crise de imigração

Dessa forma, não é difícil hoje ler uma manchete de crise nos jornais independe de onde você esteja no mundo. Além disso, com o processo globalização, muito auxiliado pela tecnologia, hoje pequenos conflitos, tensões e crises localizadas, mesmo que distantes e pequenas, afetam diretamente a estabilidade de diversos países fazendo um grande efeito dominó no mundo. Além disso, o Brasil não é imune à crises em seu território, entres elas e talvez a maior, é a tensão política atual que influência o país em diversos aspectos.

Essas crises trouxeram consigo uma questão que tem ganhado bastante destaque mundialmente, que é a inflação crescente de produtos e alimentos, a qual representa um desafio significativo para a sociedade como um todo, principalmente quando se trata da inflação em alimentos, a qual transcende as fronteiras, afetando tanto as comunidades locais quanto a economia global. Essa questão é mais relevante em um país como o Brasil que possui grande parcela de seu povo em uma situação vulnerável e tem sua economia focada em agronegócio. À medida que a economia oscila, a preocupação com a inflação de alimentos tem se tornado cada vez mais importante, passando de apenas uma preocupação doméstica para a uma reflexão sobre nossa economia como um todo demandando uma análise aprofundada.

Para melhor entender esse cenário é necessário uma base de dados que centralize informações atualizadas e históricas relevantes sobre a flutuação de preços de itens dentro do mercado interno do país, para assim gerar indicadores econômicos relevantes para futuras análises da economia do país. Entretanto, hoje no Brasil

cientistas de dados e jornalistas têm muitas dificuldades de gerar boas análises pela falta desses dados históricos e atualizados, tornando todo esse processo muito devagar e difícil.

Dado esse cenário, a empresa **X**, em conjunto com o governo brasileiro, tem apresentado diversas propostas para combater essa situação por meio de sistemas de informação e processos de coleta de dados de informações de produtos de mercado. Entretanto, até o momento, nenhuma delas retornou informações com um alto nível de granularidade e atualização em tempo real, em vez disso, a coleta de dados ainda é realizada de forma periódica em grandes intervalos e em alguns casos até manualmente.

O objetivo central deste artigo é desenvolver um sistema que automatize a extração de dados de maneira programada, coletando informações de uma ampla variedade de produtos e alimentos anunciados em sites de mercados em todo o país. Essa abordagem visa centralizar as informações em uma única base de dados focada em análises históricas. Dessa forma, estaríamos garantindo o acesso dessas pessoas a uma fonte de dados rica, nos posicionando melhor para construção de indicadores de inflação mais precisos e gerando assim relatórios mais significativos sobre a nossa atual situação econômica.

Para melhor contextualizar o problema foi utilizado a ferramenta matriz CSD (Certezas, Suposições e Dúvidas) que é uma ferramenta que auxilia contextualização do problema através de uma tabela com diferentes percepções a respeito da situação abordada.

Figura 1: Matriz CSD

Outra ferramenta utilizada para auxiliar no entendimento do contexto do problema foi a matriz POEMS (*People*, *Objects*, *Environments*, *Messages* e *Services*) que ajuda a identificar oportunidades de melhoria através de tabela com 5 elementos a serem analisados.

Figura 2: Matriz POEMS

1.1.2 Personas

Para melhor exploração do contexto apresentado será feita a descrição de duas Personas baseadas em pessoas reais envolvidas diretamente com o problema exposto na seção de análise de contexto. Nesta análise, detalharemos uma série de aspectos extraídos durante entrevistas feitas com pessoas que trabalham junto a empresa **X** para assim detalhar o máximo possível suas características, comportamento, sentimentos, dores entre outros. Assim aumentando nosso conhecimento sobre a situação atual e seus agentes.

Persona 1 - Andrei Bandeira, Cientista de dados

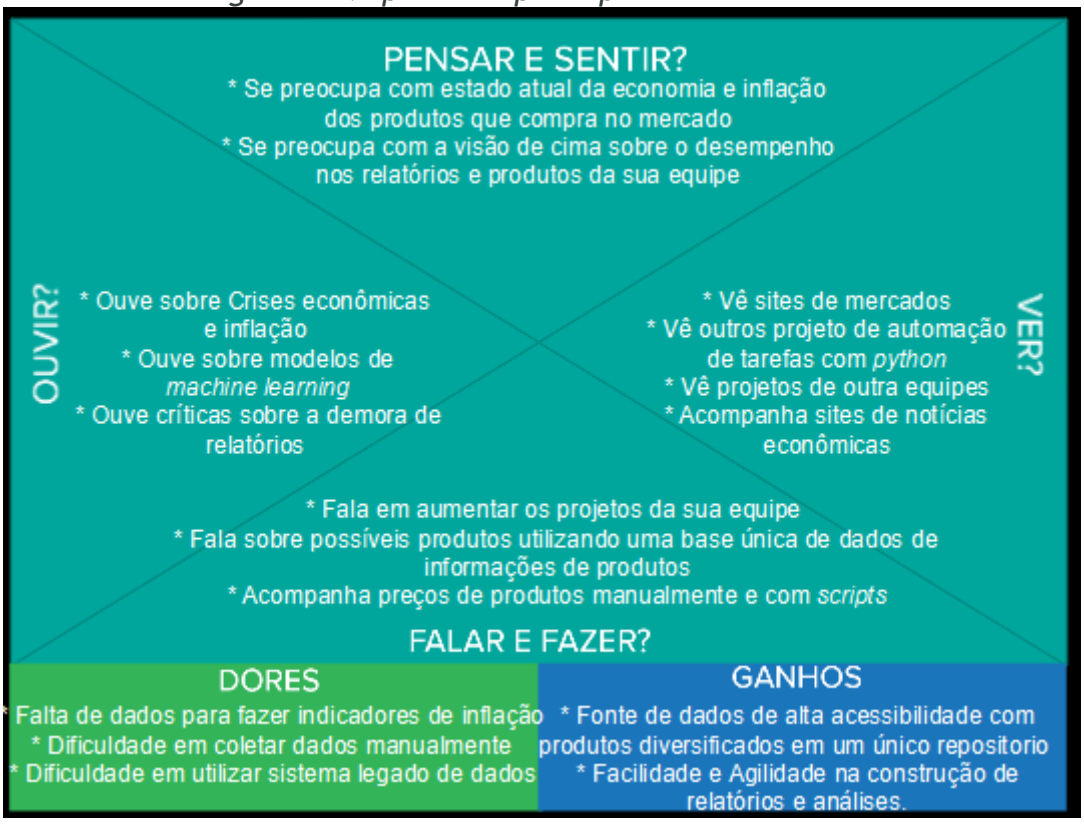
Andrei Bandeira é um profissional experiente no campo da ciência de dados e Inteligência Artificial. Ele tem 40 anos, é casado e também é pai de dois filhos pequenos, ele dedica bastante tempo à sua carreira, mas também ao bem-estar de sua família. Ele trabalha na empresa X há 8 anos e atualmente está como cientista de dados sênior do time de Inteligência artificial, possui ensino superior completo e é mestre em ciência da computação. Além disso, detém um sólido conhecimento em ferramentas de *Business Intelligence* e auxilia diversos times a utilizá-las. Andrei atualmente tem o objetivo de expandir a equipe de inteligência artificial e seus projetos, ele reconhece a importância da construção de novas bases de dados e extração de dados com alto potencial exploratório, particularmente na coleta de dados econômicos do país e tem grande paixão por utilizar modelos de aprendizagem de máquina em seus projetos, mas percebe que atualmente algumas bases de dados não estão tão ricas para esses tipos de projeto. Andrei acredita que a atual base de dados de preços de mercado é insuficiente para seus projetos e lhe dá muita dor de cabeça pela falta de informações e pela desestruturação dessas informações dentro da empresa. Ele tem uma meta clara que é reunir informações econômicas robustas e utilizá-las para desenvolver produtos e relatórios de excelente qualidade, impulsionando ainda mais a inovação e a eficácia de sua equipe.

Persona 2 – Daniel Ribeiro, Professor de Estatística

Daniel Ribeiro é um professor de estatística, com 57 anos de idade e com uma carreira acadêmica rica em experiência. Daniel faz pesquisas acadêmicas há mais de 30 anos e possui participação em diversos projetos governamentais, atualmente ele faz diversas pesquisas com foco na situação econômica do país, sempre gerando relatórios para aumentar a transparência da situação econômica no país como para o bem-estar público. Atualmente ocupa o cargo de professor-assistente na universidade federal de pernambuco no centro de informação. Além disso, exerce o papel de professor pesquisador na empresa X, a qual ele possui acesso a diversas bases de dados da empresa para fazer seus relatórios e também realiza reuniões periódicas com diferentes time de *analytics* para dar conselhos e

direcionamentos em projetos. Ele possui um bom *background* em computação, entretanto possui maior familiaridade com ferramentas voltadas para projetos estáticos com a linguagem R. Como cidadão e professor de estatística ele tem grandes preocupações com a economia do país e acompanha atentamente as notícias e tendências econômicas. O desejo de Daniel é que a empresa amplie seus horizontes e colete mais dados em diferentes fontes de dados para aumentar o potencial da empresa, gerando um ambiente rico e diversificado de dados. Dessa forma, os pesquisadores conseguiam monitorar aspectos financeiros do país, gerando assim melhores documentos e *insights* mais precisos, o que no momento atual não está conseguindo ser feito por conta da falta de qualidade de dados coletados. Um dos principais projetos que o mesmo tem acompanhado é o coletor de preços de mercado, o qual tem para si um grande peso, pois além de professor pesquisador, Daniel é um pai e um cliente habitual de diversos mercados e deseja como todo cidadão brasileiro saber se os produtos que gosta estão ficando mais caros ou mais baratos.

Figura 3: Mapa de empatia para Personas 1 e 2



1.1.3 Benefícios e Justificativas

Nesta seção será abordado benefícios e justificativas que fazem do projeto de coleta de dados de preços de produtos de mercado uma iniciativa fundamental para a empresa, expandindo as bases de dados da economia brasileira. Será exposto os impactos positivos da adoção desse projeto e as motivações dos agentes por trás da construção desse sistema. Para realizar esta análise foi utilizado duas técnicas para facilitar a representação dos dados obtidos. Sendo a primeira a *Blue print*, que é uma ferramenta que permite analisar a rotina de agentes com o objetivo de encontrar possíveis *insights* para construir uma solução mais assertiva. A segunda técnica utilizada foi o Canva de proposta de valor, que faz a representação visual para facilitar o entendimento da interação das dores e ações do público envolvido com as ações e vantagens da solução proposta. As duas técnicas aplicadas no projeto podem ser observadas abaixo respectivamente nas figuras X.X e X.X

Tabela 1: Blueprint de ações do cliente e solução proposta

Ações do Cliente	• Realizar coleta de dados disponíveis em bases incompletas. • Faz coleta manual de dados em sites de mercado quando se deseja saber sobre dados faltantes.
Objetivos do Cliente	• Fazer indicadores de inflação em diversos produtos. • Desenvolver produtos em cima de uma base de dados completa de informações de produtos/alimentos.
Atividades do Cliente	• Procurar dentro de diversos bancos de dados as informações necessárias para análise. • Construção de scripts para extrair dados de forma pontual. • Consultas nos bancos para extração da informação. • Pré Processamento de dados extraídos
Questões do Cliente	• Os dados disponíveis na base de dados da empresa X estão atualizados? • Quando foi a última vez que verificaram os dados da base da empresa X. • Existe uma forma de unificar esses dados em algum lugar?
Barreiras do Cliente	• Falta de informação nas bases da empresa X • Lentidão na hora de fazer coleta manual de dados

	• Dificuldade na construção de *scripts* de extração de dados e scripts de pré-processamento.
Saída desejável da solução	• Construção de uma base de dados que centraliza as informações sobre produtos e alimentos para construção de indicadores de inflação.
Funcionalidades da solução	• Coleta periodicamente informações sobre produtos dentro de diversos sites de mercado • Pré-processamento dos dados extraídos • Centraliza as informações obtidas em tabelas
Interação com a solução	• Os usuários apenas fazem consultas nas tabelas do banco de dados
Mensagem solução	• Informações coletadas de produtos dentro dos sites de mercados
Onde ocorre a solução	• Solução ocorre dentro do ambiente nuvem (AWS)
Tarefas aparentes da solução	• Usuários podem acessar um banco de dados RDS
Tarefas escondidas da solução	• Coletores *web-scraping* rodando periodicamente extraindo informações dos sites

Figura 4: Canva de proposta de valor

Em suma pode-se notar que o projeto traz diversos impactos para a empresa, sendo o principal a criação de uma base de dados com grande potencial de exploração e criação de novos projetos para as partes interessadas no projeto, além de substituir sistemas velhos e bases de dados legado.

1.1.4 Hipóteses

Para melhor compreendimento da solução proposta, nesta seção introduziremos uma série de hipóteses, analisadas durante o processo de entendimento do contexto do problema e que foram fundamentais para o direcionamento da formação do sistema proposto. Para auxiliar a documentação de análise de hipóteses foram utilizadas duas técnicas: Matriz de observações e hipóteses e tabela de priorização de ideias. Podemos observar as duas aplicadas ao nosso contexto nas figuras abaixo respectivamente.

Matriz de observações para hipóteses

Observações	Hipóteses
Dados sobre produtos e alimentos no Brasil podem ser obtidos nos sites de mercado	É possível extrair essas informações através de web-scrapers
Sites de mercado estão em constante atualização	É possível mapear a estrutura de um site para descobrir a melhor forma de consumir seus dados
Alimentos e produtos em mercado estão em constante mudança	É possível extrair dados diariamente para monitorar a flutuação nos valores
É necessário uma forma de consultar dados sem muito pré-processamento	Podemos criar um banco de dados estruturado e criar *procedures* e funções para facilitar o consumo das informações
Alguns sites podem ter tecnologias anti *web scraping*	É possível contornar esse problema através de algumas ferramentas (+complexidade)
Cada site de mercado possui sua variedade de produtos e preços	É possível centralizar essas informações em um banco de dados

Ideias:

- Banco de dados estruturado com preços de produtos de mercado.
- Data Lake com dados de mercado.
- Parceria com Supermercados para Coleta de Dados em Tempo Real.
- Continuar sistema atual da empresa **X.**

Ideia	B	A	S	I	C	O	Somatório
Banco de dados estruturado com preços de produtos de mercado	5	5	5	4	4	3	26
Data Lake com dados de mercado	5	5	4	4	3	1	23
Parceria com Supermercados para Coleta de Dados em Tempo Real	5	4	5	4	3	2	23
Continuar sistema atual da empresa X	4	3	3	5	2	1	18

Tabela 2: Priorização de ideias graduação BASICO.

1.2 Solução

1.2.1 Objetivo SMART

Após exposto o detalhamento completo do contexto do projeto, nesta etapa iremos apresentar os objetivos do projeto e para isso foi utilizado a metodologia SMART em que são definidos objetivos respeitando 5 critérios: Específico "*Specific*", Mensurável "*Measurable*", Atingível "*Attainable*", Relevante "*Relevant*" e Temporal "*Time-based*". Dessa forma, os objetivos do projeto são:

Construir um banco de dados estruturado com informações sobre a variedade de produtos vendidos e seus preços em múltiplos supermercados no país. Tudo deve ser feito, através de um pipeline de dados formado de coletores de dados "*scrappers*" que serão acionados diariamente para percorrer os sites fazendo a extração da informação do site completo, logo após os dados serão tratados e repassados para dentro do banco de dados para outros usuários consumirem. Dessa forma, será possível criar diversos produtos a partir desse banco de dados e dará aos pesquisadores uma forma de monitorar certos produtos dentro do país.

Lista de Objetivos:

- O banco de dados deve estar em um ambiente com alta disponibilidade (nuvem).
- O banco deve ser projetado para facilitar as pesquisas sobre preços de itens.
- Os dados devem sempre ter os preços atuais dos produtos.
- Os coletores de dados devem percorrer os sites diariamente.
- O projeto deve ser concluído até o fim de novembro de 2023.
- Os dados extraídos devem ser processados pelos coletores para depois repassar para o banco.

1.2.2 Premissas e Restrições

Para melhor curso no desenvolvimento de projetos é recomendo fazer uma análise de premissas e descobertas e restrições associadas à construção de um sistema, à medida que as premissas são fatos que orientam nossas ações e as restrições são delimitadores do escopo da solução. Nesta seção detalharemos essas premissas e restrições através de uma matriz de riscos que descreve os riscos identificados associados os a impactos possíveis e ações para remediá-los e evitá-los. Nossa matriz de risco pode ser observada na figura X.X abaixo.

Risco identificado	Impacto potencial	Ações preventivas	Ações corretivas
Sites com bloqueio de web-scrapers	Alto	Fazer mapeamento de estrutura de sites extenso.	Trocar de sites ou utilizar outras técnicas de *scrapping* (selenium)
Estrutura de site mudou durante o projeto	Alto	Monitorar constantemente os coletores	Atualizar coletor de dados
Sites apresentando instabilidades	Alto	Não depender de um único site / Adicionar fluxo quando o site está fora do ar	Re executar o coletor quando o site voltar ao ar
Agendador de atividades não funcionando	Médio	Fazer validação do agendador de tarefas	Executar o coletor manualmente
Coletores extraindo dados inconsistentes para o RDS	Médio	Validar coleta de dados antes de enviar para RDS	Remover dados inconsistentes do banco e atualizar coletor

Tabela 3: Matriz de riscos

1.2.3 Backlog de Produto

Após feita a delimitação da solução, definindo seu escopo, objetivo e documentado suas características, detalharemos nesta seção os passos para a desenvolvimento da solução proposta através da construção de um backlog de atividades detalhando cada atividade e qual sprint de desenvolvimento ela pertence. Para facilitar o acompanhamento do backlog de atividade será utilizado o trello que permitirá o rastreio e o gerenciamento das atividades nos garantindo que todas sejam executadas no momento certo do desenvolvimento. Abaixo podemos observar um snapshot do painel trello utilizado durante o desenvolvimento do projeto com as atividades listadas divididas em sprints.

Sprint 1:

- Fazer modelagem de banco de dados
- Construir RDS na Nuvem
- Fazer Análise de estrutura de sites de mercado
- Fazer script de extração de dados em jupyter notebook

Sprint 2:

- Fazer script web scraper na nuvem
- Estudar melhor forma para rodar web scraper na nuvem EC2 / Lambda
- Fazer funções de processamento de dados
- Fazer teste de ingestão de dados no RDS através de código python
- Ajustar permissões IAM entre scrapper e RDS

Sprint 3:

- Fazer fluxo completo na nuvem
- Estudar melhor forma de fazer o scheduling dos processos
- Fazer Scheduling de web scraper
- Validar fluxo e dados
- Fazer consultas SQL (simulação de usuários)

TODO: Adicionar foto do trello

2. Área de Experimentação

O que significa esta seção?

Esta seção tem o objetivo de apresentar as evidências do planejamento dos requisitos selecionados do Backlog de Produto, além de mostrar a maneira como eles foram desenvolvidos e registrar os resultados alcançados.

É necessário expor a execução e a validação dos experimentos relacionados ao desenvolvimento da solução, ou seja, testar se você está no caminho certo ou se algo precisa ser modificado (pivotar).

Quais etapas já devem estar finalizadas no momento do preenchimento desta seção? (Pré-requisitos)

No momento do preenchimento, é esperado que você já tenha cursado a disciplina de Inovação e Design Thinking, em especial as etapas do processo de Design Thinking, além de estar se preparando para desenvolver a solução idealizada no seu Projeto Aplicado.

Você também já deve ter preenchido o primeiro capítulo deste relatório (CANVAS do Projeto Aplicado).

Como esta seção deve ser preenchida?

Esta seção é a área mais dinâmica do CANVAS do Projeto Aplicado. Nela você deverá inserir os experimentos necessários para desenvolver e validar cada Sprint. Ao final do experimento, você deverá preencher o item **"Solução"** da seguinte maneira:

- **Evidência do Planejamento**: comprove que os requisitos referentes à Sprint foram efetivamente planejados. Para isso, utilize o Trello e adicione, neste campo, uma cópia da tela da ferramenta com a Sprint planejada.
- **Evidência da Execução de cada Requisito**: para cada requisito planejado, adicione um artefato que comprove o cumprimento da etapa. Podem ser anexados, por exemplo, códigos, documentos, modelos, scripts, capturas de tela, entre outros. *Importante: o número de artefatos adicionados deve ser o mesmo que o número de requisitos planejados.*
- **Evidência da Solução**: os requisitos implementados contribuem para o alcance de um resultado geral, que deverá ser comprovado neste campo. Isso

será feito por meio de capturas de tela, gráficos, modelos, textos, figuras, tabelas, testes, entre outros.

Para cada Sprint, cite no item **"Lições Aprendidas"** o que não foi validado, mas forneceu insights para ajuste da rota.

Quais ferramentas devem ser utilizadas?

Obs.: Para realização desta seção você deverá utilizar o Trello.

2.1 Sprint 1

2.1.1 Solução

Evidência do planejamento:

Evidência da execução de cada requisito:

Evidência dos resultados:

2.1.2 Lições Aprendidas

2.2 Sprint 2

2.2.1 Solução

Evidência do planejamento:

Evidência da execução de cada requisito:

Evidência dos resultados:

2.2.2 Lições Aprendidas

2.3 Sprint 3

2.3.1 Solução

Evidência do planejamento:

Evidência da execução de cada requisito:

Evidência dos resultados:

2.3.2 Lições Aprendidas

3. Considerações Finais

3.1 Resultados

Por meio de um texto detalhado, apresente os principais resultados alcançados pelo seu Projeto Aplicado.

Cite os pontos positivos e negativos, as dificuldades enfrentadas e as experiências vivenciadas durante todo o processo.

3.2 Contribuições

Apresente quais foram as contribuições que o seu Projeto Aplicado trouxe para que o Desafio proposto fosse solucionado.

Cite, por exemplo, as inovações, as vantagens sobre os similares, as melhorias alcançadas, entre outros.

3.3 Próximos passos

Descreva quais são os próximos passos que poderão contribuir com o aprimoramento da solução apresentada pelo seu Projeto Aplicado.