# Aplicação da Metodologia CRISP-DM para a Modelagem da Arquitetura do Sistema de Análise de Dados no Ambiente Acadêmico

## Entendimento do Negócio

Os objetivos técnicos e necessidades do projeto foram compostos na Sprint 1, Imersão Inicial  e Método de Estruturação de Problema.

**Problemas:**

- Dificuldade em monitorar o desempenho acadêmico e outros indicadores da educação superior;
- Dados acadêmicos incompletos e/ou dispersos em várias fontes de dados;
- Falta de acesso rápido e fácil às informações cruciais para tomada de decisão;
- Dificuldade para planejar e alocar recursos educacionais;
- Falta de integração e centralização dos dados acadêmicos;
- Dificuldade para identificar tendências e padrões dos dados no ambiente acadêmico;
- Dificuldade para avaliar o impacto de mudanças e identificar melhorias no ambiente acadêmico;

**Objetivos:**

- Permitir acesso rápido e fácil às informações tempestivas;
- Prover uma arquitetura de sistema robusta, escalável e eficiente para análise de desempenho acadêmico;
- Permitir identificar tendências e padrões ao longo do tempo e validar quais estratégias de ensino são mais eficazes;
- Disponibilizar dados e relatórios dinâmicos úteis para tomadas de decisão;
- Prover capacidade para monitorar do desempenho acadêmico em diversas escalas;
- Facilitar a comunicação entre as partes, mantendo todos informados sobre o progresso acadêmico;
- Garantir a integridade dos dados, além da segurança e conformidade do sistema;
- Auxiliar no planejamento e alocação de recursos educacionais;
- Garantir a integridade dos dados, além da segurança e conformidade do sistema;
- Garantir o acesso a dados completos de maneira unificada, em um único sistema;

# Entendimento dos Dados

Levando em consideração as limitações técnicas impostas para acesso aos dados e o cunho do projeto ser a modelagem de arquitetura, utilizaremos a visão do autor [10] em relação ao entendimento dos dados. Haja vista que, o acesso foi concedido para utilização dos dados acadêmicos na Predição de Evasão Acadêmica e que hoje fazem parte do projeto Sissa.

Assim, para que seja possível o entendimento dos dados foi preciso coletar, analisar e descrever os dados (i.e. quanto aquisição, inicialmente as informações eram obtidas por meio de arquivos no formato de valores separados por vírgulas CSV) oriundos da plataforma Analisa UFG. Porém, dada a participação do autor e os seus pares em um projeto de pesquisa, em apoio com o Ministério da Educação do Brasil, que visa buscar soluções para evasão acadêmica na educação superior, as informações passaram a ser consumidas de um datalake.

Não obstante, em relação à organização dos dados, estes foram coletados em dois grandes conjuntos tabulares: informações referentes ao ingresso (i.e. aspectos gerais, demográficos e trajetória acadêmica do aluno) e desempenho (i.e. ocorrência de matrículas em disciplinas). Considerando essa arquitetura, verificações de qualidade foram realizadas, os problemas foram encontrados e resolvidos na fase de preparação.

Alguns dos problemas encontrados pelo autor dizem respeito à: alta quantidade de atributos com informações faltantes, duplicatas de atributos e registros, inconsistências na formatação de valores decimais, registros em branco, e campos multivalorados (i.e. detalhado na Tabela 3.1: Problemas de qualidade nos dados coletados, apresentado em [10]).

# Preparação dos Dados

Nesta etapa, compreendeu-se ao autor o ajuste dos registros para submissão à modelagem. Como o processo de modelagem utilizou diferentes abordagens, com variadas configurações de atributos, foi desenvolvida uma estrutura de dados intermediária, genérica o suficiente para:

1. ser adequada a diferentes espécies de análises, sem a necessidade de consultas contínuas as tabelas de ingresso e desempenho;
2. armazena os atributos essenciais para criação dos modelos preditivos;

Dessa maneira, com base no modelo Student Sequence (i.e. é a preservação do progresso de um indivíduo ao longo do tempo, de modo que ao predizer informações, os

padrões de sua história sejam captados e contribuam para a obtenção de melhores resultados em relação a abordagens unicamente vetoriais) de Mahzoon.

Dessa forma, propôs-se no trabalho um mecanismo de divisão temporal dos dados destinados a treino e teste, que avalie o desenvolvimento de um aluno até um ponto específico da sua história e como detalha o autor, a informação predita será a evasão ou graduação de um indivíduo, independente do tempo de permanência na instituição.

# Modelagem

Nesta etapa, levando em conta o cunho de modelagem de arquitetura do projeto, é realizada a análise dos dados preparados anteriormente e pode-se utilizar técnicas de modelagem estatística, mineração de dados e machine learning para buscar padrões, identificar insights e gerar informações relevantes para a gestão acadêmica.

Após reunião com uma parte interessada que representa a figura de um coordenador de curso, parte que pode ser influenciada diretamente pela solução proposta, apresentou que uma de suas dores poderia ser tanto os problemas que foram apresentados anteriormente como também a falta de uma ferramenta que possibilite a utilização de ferramentas de suporte a solução de problemas no ambiente acadêmico. Dentre eles podemos citar a dificuldade para auxiliar um aluno em situação que esteja fora do fluxo acadêmico e precise de suporte do coordenador para construir uma matriz curricular que faça com que ele possa retornar ao fluxo convencional proposto pela unidade acadêmica.

Assim, essas análises podem ser direcionadas a diferentes temas, como a análise de desempenho dos alunos, padrões de evasão, efetividade de programas acadêmicos, entre outros. Portanto, é importante utilizar técnicas adequadas ao contexto acadêmico e validar os resultados obtidos.

# Avaliação

Após a modelagem, os resultados obtidos são avaliados em relação aos objetivos e requisitos definidos no início do projeto. Assim, é necessário verificar se as análises feitas são consistentes e se os insights gerados podem realmente contribuir para a gestão acadêmica.

Assim como descrito no Roadmap, as etapas de avaliação permeiam o seu direcionador em relação a validação de conformidade com o que é esperado para que se tenha um Produto Mínimo Viável (MVP) logo após a estruturação da modelagem. Entre elas, as funcionalidades mínimas necessárias com funcionamento completo são: do cadastro, login de usuário, alteração de dados pessoais, telas de listagem de Dashboards e Relatórios, tela para listagem de perguntas, tratamento de erros e mensagens de aviso. Já em com o funcionamento parcial estão as funcionalidades de chat orientado por contexto e a ferramenta para gerar Dashboards e relatório com base no prompt fornecido pelo usuário.

Caso seja necessário, é possível retornar às etapas anteriores para refinar os processos de extração e preparação de dados, ou até mesmo redefinir os requisitos e objetivos, se houver necessidade.

## Implantação e Manutenção

Nessa fase, os resultados obtidos são implantados na instituição acadêmica, seja por meio da criação de relatórios e dashboards interativos, da disponibilização de APIs para integração com outros sistemas ou da geração de alertas automáticos (i.e. é importante considerar as necessidades e limitações da instituição, bem como a capacitação dos usuários para que saibam como utilizar e interpretar as informações geradas e também levar em conta o escopo do projeto que visa a modelagem de arquitetura).

Após a implantação, é necessário garantir a manutenção contínua da arquitetura de extração e visualização de dados. Isso inclui atividades de monitoramento, atualização dos modelos, correção de possíveis problemas, evolução com base no feedback dos usuários e na mudança de necessidades e objetivos da instituição acadêmica.

Portanto, a aplicação da metodologia CRISP-DM para a extração e visualização de dados em um sistema de gestão acadêmico permite uma abordagem estruturada, metódica e iterativa, garantindo que os resultados obtidos sejam relevantes e utilizáveis para a tomada de decisões e análises na instituição acadêmica.