**Metodologia:**

A pesquisa será realizada em uma das maiores varejistas de moda do Brasil a qual apresenta uma base de dados vasta e consolidada, devido à proximidade entre o pesquisador, empresa e a relevância da empresa perante o mercado brasileiro.

*A priori*, os dados a serem coletados serão extraídos diretamente do banco de dados no formato *.csv* etabulados e operacionalizados no R-Studio. A tabela a ser analisada estará no formato *wide*, onde as observações estarão em linhas e as variáveis em colunas. A quantidade de observações ainda será definida, contudo as observações serão a nível de clientes em todo o território nacional, caso a base fique muito pesada será processada diretamente no *Apache Spark* em linguagem R.

Pensando em uma abordagem metodológica completa, pesquisa estará estruturada de acordo com o modelo de referência *Cross Industry Standard Process for Data Mining* (CRISP-DM) que promove uma revisão de todo o ciclo de vida de um projeto de data mining, onde contempla suas fases, tarefas e relacionamentos entre elas. (Chapman *et. al*, 1999).

Mesmo sendo a metodologia mais utilizada em projetos de mineração de dados, nem sempre seus resultados são positivos, pois não inclui algumas atividades de gerenciamento de projeto (Mariscal *et. al*, 2010).

Contudo, ainda hoje é o modelo padrão *de facto* para mineração de dados e projetos exploratórios em ciência de dados e mesmo com suas limitações não deve ser descartado para projetos que vão dos dados ao conhecimento, quando se tem um claro objetivo de negócios que se traduz dentro de um objetivo de mineração de dados. (Plumed *et. al,* 2021).

O modelo de referência CRISP-DM consiste em 5 fases de acordo com a Figura 1. São Elas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e produção.

Diagrama

Descrição gerada automaticamente

Fig1 – Modelo de referência CRISP-DM (Plumed et. al, 2021)

Primeiramente para que sejam extraídos dados relevantes para a criação do modelo, torna-se necessário o **entendimento do negócio**. Uma entrevista prévia com a gestão e a equipe técnica para nortear quais características das observações podem influenciar ou não na rotatividade dos clientes.

Posteriormente um segundo contato para esclarecer o devido **entendimento dos dados**, verificar os tipos de variáveis, granularidade, estrutura hierárquica, como são extraídas, a relação ao longo do tempo e a relação entre elas.

O próximo passo é todo o processo de limpeza e **preparação dos dados**, a verificação de *missing values* e a retirada ou não caso haja necessidade. Também faz parte do *setup* identificar a necessidade de alteração nos tipos de dados de *object* para *float* e *vise e versa*. Posteriormente as colunas do tipo *object* passarão por um processo de dummização.

Objetificando uma maior confiabilidade nos dados coletados, uma análise de *outliers* deverá ser feita repartindo as observações das colunas numéricas em quartis. E as que apresentarem distância maior que 1,6 dos quartis superior e inferior, serão excluídas.

Em busca de um melhor aproveitamento do número variáveis a serem utilizadas no modelo, será feito um procedimento *step-wise* retirando as que não demonstrarem ser estatisticamente significante em um modelo de regressão linear simples.

Por fim os dados serão padronizados pela fórmula *z-score,* para que fiquem na mesma unidade de medida.

Depois de definidas as características relevantes, vem o processo de **modelagem,** nele o procedimento *step-up* definirá a significância estatística do modelo multinível. Primeiramente, deve-se verificar os *p-values* do modelo nulo, desconsiderando a aleatoriedade nos interceptos e nas inclinações. Passando a 95% de confiança o próximo passo é considerar a aleatoriedade nos interceptos e verificar a significância do modelo. E por fim avaliar estatisticamente uma equação multinível com interceptos e inclinações aleatórias.

No processo de **avaliação,** o modelo de regressão simples refinado pelo processo de *step-wise* será confrontado com o hierárquico e na busca de uma validação mais relevante academicamente, em vista não só a avaliação de modelos estocásticos, bem como determinísticos, também será proposto um confronto com modelos em estado da arte na literatura como o *XGboost* e o *Decision Tree algorithm.*

Por fim, espera-se avaliar a curva ROC de ambos os modelos e os desempenhos de sensibilidade, especificidade e acurácia para que o melhor seja escolhido.