Projeto Potencial de Mercado

Objetivo: Prever quais empresas poderão se tornar clientes nos próximos anos.

Objetivos específicos: Determinar segmentos de mercado, microregiões e mesoregiões com potencial para crescimento.

Descrição do Problema

O problema do potencial de mercado pode ser visto como um problema de classificação que é um problema muito comum na área de Mineração de Dados. Dado uma base de dados com registros não classificados, um algoritmo de classificação atribui a cada registro uma classe ou categoria dentre as classes/categorias existentes.

No problema do potencial de mercado, podemos definir, por exemplo, duas classes: Uma classe representa um cliente em potencial e a outra classe representa um cliente não considerado potencial. Nesse caso, todas as empresas não clientes da Marluvas seriam classificadas como um cliente em potencial ou não. É importante destacar que poderiam ser utilizadas mais classes para representar potenciais clientes em diferentes níveis como, por exemplo, indo desde um não potencial cliente, passando por um com potencial baixo até chegar no nível com grande potencial.

A classificação dos dados irá se basear em diversos atributos das empresas como segmento de mercado, tamanho da empresa, exigência de EPI, crescimento da Marluvas na microregião da empresa, entre outros, assim como pode se basear também em atributos mais subjetivos como conhecimento da marca Marluvas, confiança na marca, entre outros. Um importante atributo é o fato da lei exigir que uma empresa forneça EPIs para seus funcionários.

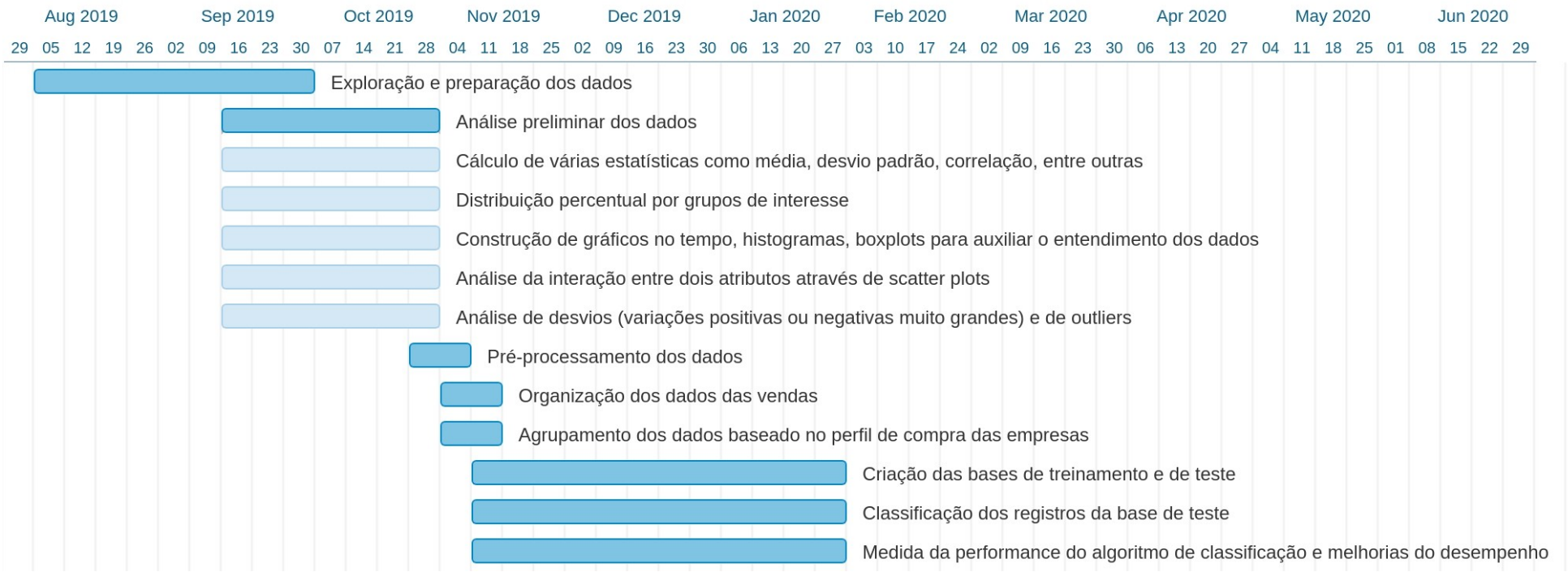
Cronograma

O desenvolvimento do projeto Potencial de Mercado pode ser dividido em duas etapas:

1. Análise.
2. Mineração de dados.

A etapa de análise compreende várias atividades para o entendimento dos dados e preparação desses dados para as etapas posteriores. A etapa de mineração é a responsável, dentre outras coisas, por particionar os dados em subgrupos mais homogêneos, por classificar os dados e procurar padrões nos dados.

O cronograma do projeto com o tempo para realização de cada atividade é dado no gráfico de Gantt na figura abaixo. A descrição de cada atividade é dada a seguir.



Análise

1. Exploração e preparação dos dados: Os dados das bases de dados do CNAE e CAGED serão importados para um banco de dados PostgreSql para que possam ser efetuadas consultas sobre essas bases. Um exemplo de consulta seria o cruzamento de dados dessas duas bases para listar todos os empregados de uma empresa. Será realizado um

estudo sobre os dados para determinar atributos importantes e definir as formas como os dados serão organizados. Uma forma de organizar os dados, por exemplo, é separar os clientes inativos dos ativos. Os clientes inativos são potenciais clientes que podem ser agrupados de várias formas (por exemplo, por segmento de mercado). Serão implementadas diversas consultas para preparação dos dados para as etapas posteriores.

2. Análise preliminar dos dados:

   Inicialmente, serão implementadas as seguintes análises:

   - Distribuição percentual por grupos de interesse (organizar os dados, por exemplo, por segmentos de mercado, por microregião e por mesoregião);
   - Cálculo de várias estatísticas como média, desvio padrão, correlação, entre outras;
   - Construção de gráficos no tempo, histogramas, boxplots para auxiliar o entendimento dos dados;
   - Análise da interação entre dois atributos através de scatter plots;
   - Análise de desvios (variações positivas ou negativas muito grandes) e de outliers.

Mineração de dados

1. Pré-processamento dos dados: Normalização dos dados, tratamento de valores inexistentes, nulos e outras transformações necessárias para preparar os dados para os algoritmos de classificação.
2. Organização dos dados das vendas: Separar os dados de venda de cada linha de calçados por segmento de mercado, por microregião e por mesoregião. Separação desses dados também por trimestre, semestre e ano.
3. Agrupamento dos dados baseado no perfil de compra das empresas:
   - Agrupar as empresas pelas linhas de calçados que são adquiridas;
   - Agrupar as empresas pelos períodos do ano nos quais determinadas linhas de calçado são compradas.
4. Criação das bases de treinamento e de teste: Devem ser criadas as bases de treinamento e teste para classificação das empresas em potenciais clientes. A criação das bases será feita utilizando o algoritmo de fragmentação chamado Cross-Validation.
5. Classificação dos registros da base de teste: Classificação dos dados utilizando uma ou mais técnicas de classificação. Inicialmente, pretende-se utilizar árvores de decisão para essa tarefa.
6. Medida da performance do algoritmo de classificação e melhorias do desempenho: Medida do número de erros, falsos positivos e falsos negativos na classificação. Análise dos resultados para tentar melhorias no modelo. Variação dos parâmetros do modelo avaliando o impacto no desempenho.

Destaca-se que as etapas 4, 5 e 6 serão realizadas várias vezes durante a classificação dos registros com o objetivo de diminuir o número de erros, aumentando dessa forma o desempenho do algoritmo de classificação.