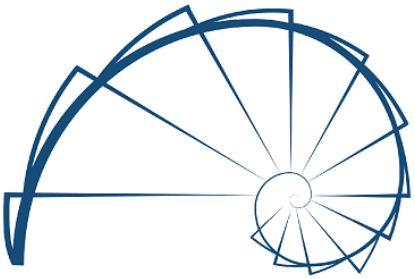

PROJETO 2022.02 - UNDERLYING
Mineração de Dados em Investimentos – Produtos Similares

Versão 0.1:
Tratamento e Limpeza da Base de Dados

Equipe de Projeto Underlying:

Adriel Douglas Nogueira Carlos - 2018012346

Ivan Leoni Vilas Boas - 2018009073

Lucas Tiense Blazzi - 2018003310

COM923 - Tópicos Especiais em Inteligência Artificial

Vanessa Cristina Oliveira Souza

https://github.com/lucasblazzi/market-datalake

IMC - Instituto de Matemática e Computação

Av. BPS, 1303 - Caixa postal 50 - 37500-903

Itajubá - MG - Brasil Telefone: 35-3629-1135

E-mail:imc@unifei.edu.br

1. Tratamento e limpeza da base dados

O código de coleta dos dados, assim como os notebooks para tratamento e limpeza dos datasets estão disponíveis no GitHub, sendo os notebooks no path /etl/notebooks e a coleta em /sources. Como a captura das bases gerou um grande volume de dados (14GB) eles foram carregados no Drive

O tratamento e a limpeza da base de dados foram realizados utilizando a ferramenta jupyter. Cada tratativa o que tange ao tratamento e a limpeza sobre as bases estão passo a passo explicadas nos próprio arquivos do notebook ipynb criados para o desenvolvimento deste projeto. Os notebooks comentados e explicados serão encaminhados em anexo e cada um deste que foram criados serão listados a seguir com os seus objetivos:

- **fixed_income**: Tratará o dataset de registro de renda fixa;
- **fixed_income_prices**: Tratará o dataset de preços de renda fixa;
- **funds**: Tratará o dataset de registro de fundos de investimento;
- **funds_prices**: Tratará o dataset de preços de fundos de investimento;
- **index_prices**: Tratará o dataset de preços conforme os indicadores (CDI, IPCA, SELIC, IGPDI, IGPM);
- **stocks**: Tratará o dataset de registro de produtos de renda variável;
- **stocks_prices**: Tratará o dataset de preços de produtos de renda variável;
- **metrics**: Tratará o dataset das ações, dos fundos e de renda fixa a partir da junção dos datasets de preços no que se refere as métricas dos investimentos (risco, volatilidade, retorno, sharpe, etc)
- **dataset**: Realiza a concatenação e o tratamento final dos dados dos investimentos de todos os datasets (métrica, ações, fundos e renda fixa).

2. Perguntas a serem respondidas pelo modelo

- É possível agrupar os produtos de forma que eles representem o perfil de um investidor?

- Qual produto mais recomendado para um cliente que possui interesses de investimento conhecidos quanto ao objetivo, exposição a risco e renda?

- É possível indicar um produto para um cliente que atenda seus desejos e ao mesmo tempo diversifique sua carteira ou potencialize seus retornos?

- Quais características os produtos dos diferentes mercados possuem em comum?

- Quais são os melhores atributos de investimentos que ajudem a prever os possíveis investimentos desejados pelos clientes?

- Quais os atributos de investimentos são fundamentais para ser considerados numa possível aquisição do investimento?

- Dado um produto de investimento qual é o mais similar a ele com base em suas características?

- Qual produto mais recomendado para um cliente que possui uma carteira de investimentos com n produtos conhecidos?

- É possível indicar um produto para um cliente que atenda seus desejos e ao mesmo tempo diversifique sua carteira?

3. Tarefa de DM

Será utilizado o aprendizado de máquina **não supervisionado**, pois não haverá uma referência (ou critério específico) para o modelo seguir. O modelo também conhecido como modelo exploratório ou de interdependência irá sozinho tentar encontrar semelhanças e diferenças entre os dados de maneira que separe da melhor maneira possível os investimentos em categorias (clusters). Como principais algoritmos levantados para análise inicial estão o **KMeans e a Clusterização Hierárquica (Agglomerative Clustering)**, além disso, foi abordada a hipótese da viabilidade da utilização do **KNN não supervisionado**.

3.1 Validação dos grupos gerados

O primeiro passo para validação dos grupos gerados será a utilização de métricas para determinação inicial do desempenho do modelo, como o Silhouette Score, e análise visual para verificar a dispersão dos clusters.

Passando da etapa inicial será realizada a geração da árvore de decisão com base nos clusters gerados, para determinar se o critério utilizado é coerente com o contexto de cada uma das features que foram selecionadas, e analisar quais regras determinam a formação de cada cluster, contribuindo para uma visão inicial do conteúdo dos clusters.

Por fim, serão retiradas medidas estatísticas (como média, mediana, mínimo, máximo) de cada cluster para verificar o comportamento das features. Assim, pretende-se tentar determinar o conteúdo do cluster a partir de perfis de clientes investidores. Alguns cenários hipotéticos seriam a percepção de produtos voltados para clientes conservadores de baixa renda e com foco no longo prazo ou clientes sofisticados de alta renda e com foco no curto prazo por exemplo. Esses critérios serão analisados conforme o comportamento das features a partir do agrupamento do seu contexto, como: risco do produto (volatilidade, drawdown máximo, value at risk e risco – associado ao suitability do investidor), valor monetário (aplicação mínima, isenção de imposto de renda e tipo de investidor – associado a capacidade financeira do investidor), prazo de aplicação (liquidez – associado ao objetivo do investidor), exposição (estratégia, benchmark)