Mercado Financeiro Brasileiro - Alternativas de investimento em renda variável

Elysiario Santos, *2220153*

***Abstract*—Com o crescimento exponencial dos dados, há cada vez mais necessidade das organizações e empresas, sejam elas grandes, pequenas ou médias, de se adaptarem a esse aumento de fluxo com o objetivo de otimizar processos e extrair as informações mais relevantes para as respetivas áreas estratégicas do negócio. Torna-se por isso fundamental, dentro do contexto empresarial, a mineração dos dados. Conhecendo os inúmeros benefícios deste processo dentro da organização, revela-se importante os insights que podem ser obtidos através do processo de *data mining*, e permitem que as organizações criem novas oportunidades de negócios e serviços inovadores.**

***Index Terms*—Mineração de dados, Metodologia CRISP-DM, Mercado de ações Brasileiro, Investimentos, Mercado Financeiro.**

I. Introdução

O VOLUME de dados produzidos está dobrando a cada dois anos. Somente os dados não-estruturados compõem 90% do universo digital. Porém, mais informação não significa necessariamente mais conhecimento. A mineração de dados permite separar todos os ruídos caóticos e repetitivos em seus dados; entender o que é relevante para, então, fazer um bom uso dessa informação para avaliar os resultados possíveis e acelerar o ritmo de tomadas de decisões bem-informadas. [1]

Surge assim, a importância de criar um mecanismo que permita às empresas minerar esses dados, a mineração de dados ganhou grande visibilidade nos últimos anos devido ao crescimento exponencial de dados disponíveis e à capacidade de gerar informações relevantes para as organizações. O *Data mining* pode ser utilizado em diversas áreas, como marketing, saúde, finanças, segurança e muitas outras. As informações mineradas também podem ajudar as organizações a entenderem melhor seus clientes, detectar fraudes, prever tendências, tomar decisões estratégicas, estimar demandas, pois se trata de uma área de estudo complexa, que envolve a combinação de várias técnicas, como reconhecimento de padrões, aprendizado de máquina, estatística, processamento de linguagem natural, aprendizado profundo, gerenciamento de banco de dados e outras áreas, sendo uma das principais ferramentas de análise de dados que as organizações podem usar para obter insights valiosos, o que aumenta a competitividade e a lucratividade das empresas.

Neste relatório pretende-se assim, por meio da mineração de dados, ajudar os investidores a tomarem decisões mais assertivas e entender como ela pode ser aplicada a diferentes problemas, como, por exemplo, em quais ativos devo investir no mercado brasileiro para ter maior chance de sucesso. Neste projeto também usaremos a principal metodologia de mineração de dados – CRISP-DM (*C*ross Industry Standard Process for Data Mining).

Exploraremos também algumas técnicas de mineração de dados, previsão e classificação, como análise de clusters e associação. Por fim, exploraremos algumas aplicações da mineração de dados, como segmentação de mercado, detecção de ativos promissoras e previsão de rentabilidade.

Elysiario Santos 2220153@my.ipleiria.pt

II. Metodologia CRISP - DM

CRISP-DM (*C*ross-Industry Standard Process for Data Mining) é uma metodologia para projetos de mineração de dados, desde que foi desenvolvida em 1996, tornou-se a metodologia mais utilizada para mineração de dados. [2]

CRISP-DM (*C*ross-Industry Standard Process for Data Mining) , que significa Processo Padrão de Vários Segmentos de Mercados para Mineração de Dados, é uma forma comprovada pelo mercado para orientar seus esforços de mineração de dados. Como uma metodologia, ela inclui descrições das fases típicas de um projeto, as tarefas envolvidas em cada fase e uma explicação dos relacionamentos entre essas tarefas. Como um modelo de processo, o CRISP-DM fornece uma visão geral do ciclo de vida da mineração de dados.[3]

O modelo CRISP-DM é flexível e pode ser facilmente customizado. Por exemplo, se sua organização planejar detectar a lavagem de dinheiro, é provável que você examine detalhadamente grandes quantidades de dados sem uma meta de modelagem específica. Em vez da modelagem, seu trabalho irá se concentrar na exploração e visualização de dados para descobrir os padrões suspeitos em dados financeiros. O CRISP-DM permite que você crie um modelo de mineração de dados que se encaixe em suas necessidades específicas. Em tal situação, as fases de modelagem, avaliação e implementação podem ser menos relevantes do que as fases de entendimento e preparação de dados. Entretanto, ainda é importante considerar algumas das questões levantadas durante essas fases posteriores para o planejamento de longo prazo e para futuras metas de mineração de dados. O modelo de ciclo de vida é composto de seis fases com setas indicando as dependências mais importantes e frequentes entre as fases. A sequência das fases não é rigorosa. De fato, a maioria dos projetos vão e voltam entre as fases, conforme necessário. [3]

Conforme figura1 a metodologia CRISP-DM é composta por seis fases principais que devem ser seguidas, sendo:

- Fase 1: Entendimento do negócio: É necessário entender o objetivo do estudo e o contexto em que os dados foram coletados.

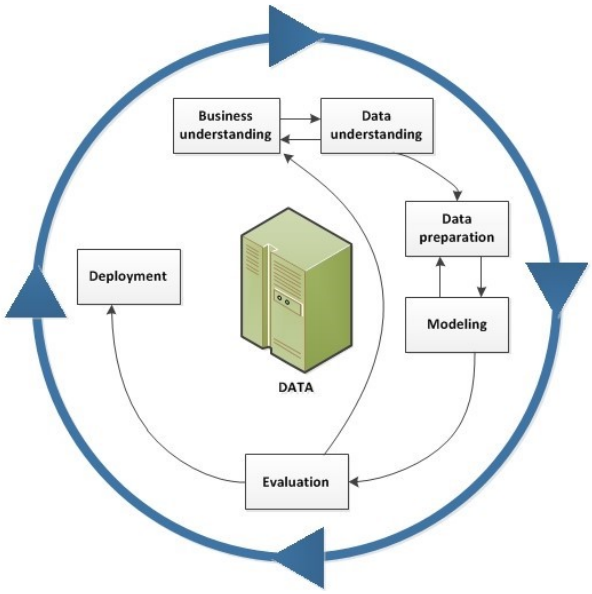


Fig. 1. IBM - CRISP-DM

- Fase 2: Entendimento dos dados: O objetivo é obter um conhecimento mais profundo dos dados coletados, verificando sua qualidade, consistência e completude.
- Fase 3: Preparação dos dados: É necessário preparar os dados para a análise, realizando limpeza, seleção, transformação e integração dos dados.
- Fase 4: Modelagem: Cria-se modelos de análise de dados e explora-se diferentes algoritmos e técnicas para alcançar os resultados que se deseja.
- Fase 5: Avaliação: Os modelos são avaliados em termos de qualidade e relevância para o negócio.
- Fase 6: Implementação: O modelo escolhido é implementado em um ambiente de produção e o processo é monitorado para garantir que os resultados obtidos sejam consistentes.

III. Compreensão do Negócio

O aluno, Elysiario Santos, faz parte de uma comunidade Católica com sede em Campinas, Brasil, (”https://www.pantokrator.org.br/”) e foi provocado a realizar um trabalho de classificação e previsão para tomada de decisões de investimento. Dentro da Comunidade Pantokrator, que tem mais de 300 membros, foi formado um grupo de estudos do mercado financeiro cuja missão é orientar os demais membros em relação a finanças pessoais. Para isso, foram gravadas algumas aulas que orientam os membros da comunidade a como organizar suas finanças e começar a investir em renda fixa e variável, que é o objeto de estudos desse projeto. O projeto será realizado para esse grupo, interessado em investir em ações e fundos imobiliários brasileiros, podendo estender-se aos familiares e pessoas próximas a eles. O grupo de investidores tem uma estratégia de investimento clara e consistente; querem investir em ações ou fundos imobiliários mais rentáveis do mercado brasileiro. Nosso principal objetivo é proporcionar a esse grupo de investidores informações reais baseada em dados do gênero que, eles poderão tomar decisões mais conscientes e bem fundamentadas, aumentando suas chances de sucesso.

A. Descrição do negócio

O grupo, formado por três membros da comunidade Pantokrator, reuniu-se estrategicamente, analisou vários indicadores financeiros do mercado brasileiro, como IPCA (índice que mede a inflação no Brasil), taxa básica de juros (Selic), entre outros aspectos relevantes e conforme o grupo optou-se por investirem em títulos do tesouro direto, (renda fixa) e ações e fundos imobiliários (renda variável), o grupo também decidiu alocar seus recursos em diferentes ativos de diferentes setores, buscando equilibrar o risco e o retorno esperado, visando sempre os que oferecem uma rentabilidade mais previsível, com menor risco. Para melhor análise do cenário proposto pelos investidores é necessário entender os tipos de investimentos propostos para análise; Fundo imobiliário é uma maneira de investir em imóveis, ou papéis com lastro imobiliário, de forma descomplicada. Uma vez que, indiretamente, quem realmente aloca o dinheiro no mercado é o gestor do fundo, que por sua vez administra o capital de diversos investidores. Trata-se, portanto, de um jeito inteligente e prático de investir no mercado imobiliário sem ter de comprar um imóvel de fato. Os Fundos Imobiliários têm um gestor especializado que, diariamente, faz o acompanhamento do patrimônio e do mercado. Conforme os resultados obtidos, ele faz as alocações necessárias. O objetivo é ter a máxima rentabilidade. Outra forma de obter lucros é com a valorização do bem em si. Então, os lucros vêm dos rendimentos destes ativos e também da valorização das cotas. Basicamente, do ponto de vista do investidor, o processo é muito parecido com o da compra de ações, até porque os fundos são listados na Bolsa de Valores como um ”papel”. [4] Existem diversos tipos de fundos imobiliários no mercado — cada um com estruturas, estratégias e carteiras de ativos diferentes. Contudo, é possível classificar os Fundos imobiliários (FIIS) em três grandes modalidades:

- 1. Fundos de tijolo: São Fundos focados majoritariamente em empreendimentos físicos. A política dos FIIs de tijolo é investir na aquisição, construção ou aluguéis de imóveis comerciais, como: Shopping Centers, Faculdades, Prédios comerciais, agências bancárias, centros de distribuição, galpões e armazéns e Hospitais.[5] O objetivo de um Fundo de tijolo é encontrar pessoas ou empresas interessadas em utilizar os imóveis adquiridos. Em troca, o FII recebe uma renda mensal de aluguel para ser distribuída a seus cotistas.
- 2. Fundos de papel[5] Um FII de papel tem como estratégia investir em títulos financeiros vinculados ao mercado imobiliário, como LCI, CRI, títulos de recebíveis imobiliários, cotas de outros Fundos Imobiliários, entre outros. O lucro do Fundo vem dos juros e dividendos pagos por esses títulos, ou da venda deles.[5]
- 3. Fundos de Fundos (FOFs) Os Fundos de Fundos, também chamados de FOFs, são aqueles que investem em cotas de outros FIIs, sejam eles de papel, tijolo ou em outros FOFs. Então, constituem uma boa forma de diversificação nesse mercado, uma vez que você tem acesso a vários FIIs com um único investimento.[5]

Ações são títulos que representam uma fração do capital de uma empresa. Ao comprar uma ação, você se torna sócio da companhia e pode receber dividendos e lucros. Para entender o que é uma ação, primeiro proponho pensarmos no processo para que ela seja lançada no mercado. Você já deve imaginar que uma empresa, quando decide expandir sua atuação no mercado, precisa procurar recursos para possibilitar este crescimento. Assim, um dos principais caminhos para essa expansão é a abertura de capital e a oferta de suas ações no mercado financeiro. Dessa forma, ações representam uma fração do valor das empresas. Podemos dizer, então, que ação é um pequeno pedaço de uma empresa. E isso significa que, quando você compra uma ação, você se torna sócio dessa organização. As ações são negociadas na Bolsa de Valores, que serve como um ponto de encontro online entre investidores que querem comprar uma ação e investidores que querem vendê-la. Além disso, a Bolsa também possibilita que as empresas disponibilizem suas ações para esses investidores, iniciando essas negociações. [6]

B. Objetivos

Sendo o grupo de investidores pertencente a Comunidade Pantokrator , é importante destacar que o estudo será realizado de forma voluntária, ou seja, não haverá nenhum tipo de contraprestação financeira. Os investidores disponibilizaram um dataset real com dados da BOVESPA (Bolsa de Valores de São Paulo), cujo objetivo do estudo é evidenciar, através técnicas de mineração de dados, se o investimento em ações e fundos brasileiros analisados são atrativos do gênero que, esses investidores tenham análises reais e possam decidir se investir ou não em uma determinada ação/fundo do mercado brasileiro a fim de minimizar o risco de insucesso do investimento. Os principais objetivos de negócio são:

- Identificar ativos com o potencial de oferecer bons retornos financeiros, seja no curto, médio ou longo prazo.
- Avaliar a probabilidade de perda de investimento.
- Avaliar qual(is) ativo(s) poderá ser o mais atrativo a fim de diversificar sua carteira de investimentos e buscar maximizar retornos financeiros.
- Avaliar o desempenho passado de um ativo, o que pode ser usado como um indicador da probabilidade de seu desempenho futuro.

Em resumo, os objetivos de negócio estão alinhados para os investidores obterem melhores retornos e minimizarem os riscos.

C. Critérios de Sucesso

Com base em análises estatísticas e financeiras, o estudo será capaz de:

- Identificar quais ativos apresentam um potencial de bons retornos financeiros.
- Apresentar uma avaliação precisa da probabilidade de perda de investimento em cada ativo. Consideramos a taxa de acerto de no mínimo 80%.
- Identificar o ativo com maior potencial de valorização. O retorno financeiro esperado do ativo identificado como o provavelmente mais rentável deve ser superior ao retorno financeiro esperado do conjunto de ativos analisados.
- Deve fornecer sugestões de diversificação da carteira de investimentos, as sugestões serão coerentes com os objetivos de maximização de retornos financeiros e minimização de perdas. Para isso, será possível ver os ativos mais interessantes por setor ou tipo.
- Apresentar uma análise do desempenho passado do ativo identificado como mais rentável, cuja análise deve indicar que o desempenho passado do ativo é um bom indicador de seu desempenho futuro.
- Deve fornecer um relatório completo, com as análises estatísticas e financeiras realizadas, incluindo as sugestões de diversificação da carteira de investimentos e outras informações relevantes. O relatório deve ser claro e de fácil entendimento, permitindo que os investidores possam tomar decisões de forma clara e assertiva baseada em dados fidedignos.

Como última meta e critério de sucesso do nosso projeto está a satisfação do cliente, o projeto deve atender às expectativas dos investidores, demonstrando que as análises e sugestões fornecidas são úteis e confiáveis. Os resultados do projeto devem encorajá-los em suas decisões de investimento.

D. Cenário

Tendo em vista que, o projeto baseia-se em prover análise de qualidade para auxiliar os investidores a tomar decisões assertivas sobre onde investir seu dinheiro, os investidores providenciaram o dataset com informações da BOVESPA (Bolsa de Valores de São Paulo). Os especialistas da área de negócio serão o grupo de investidores que solicitaram a análise proposta neste projeto. No âmbito de softwares/ferramentas a serem utilizadas neste projeto de mineração de dados, serão Python, com diversas bibliotecas para análise estatística e criação de modelos e Excel para a manipulação e geração do dataset que será utilizado na aplicação. As informações provenientes da BOVESPA (Bolsa de Valores de São Paulo) e disponibilizados para análise pelo grupo de investidor aos autores, contém dados financeiros, históricos de preços de ações e fundos imobiliários do mercado brasileiro. Tendo como requisito deste trabalho, os desenvolvedores denominados autores do projeto deverão ter acesso aos dados financeiros, bem como históricos para a realização de análises, que deverá ser entregue pelos investidores. Como restrição teremos os resultados obtidos através deste estudo que serão válidos apenas para o período de análise realizado e não podem ser extrapolados para períodos futuros, bem como ser utilizado para seu devido fim. Os Riscos e contingências encontradas poderão ser a falta de acesso a dados financeiros relevantes ou estes dados serem de baixa qualidade, neste caso será preciso buscar alternativas para obter os dados necessários ou ajustar os dados disponíveis, para não comprometer o projeto. No contexto de terminologia, definiu-se dos termos para garantir que todos os membros da equipa tenham uma compreensão comum dos objetivos e das técnicas aplicadas neste projeto. No que se refere a termos do negócio:

- Investimentos rentáveis: investimentos que geram retornos financeiros positivos em relação ao risco assumido;
- Fundos imobiliários: investimentos em imóveis, recebíveis ou outros títulos negociados na bolsa de valores;
- Ações: investimentos negociados na bolsa de valores, compra de uma pequena parte de uma empresa;
- Grupo de investidores: Grupo composto por 3 pessoas que auxiliarão a Comunidade Pantokrator na tomada de decisões de investimentos.
- Dataset – conjunto de dados que será utilizado na análise.
- Ativos - São títulos e contratos negociados no mercado de capitais e no mercado financeiro e são adquiridos por meio de negociações de compra e venda.

Termos do projeto:

- Código - Ticker que permite identificar qual a empresa o fundo estamos a consultar ou negociar na bolsa de valores de São Paulo(Bovespa).
- Tipo - Determina se estamos a consultar uma ação (ligada a empresas) ou fundo.
- Setor - Informa qual a área de atuação da empresa ou o que é característica do fundo. Ex: Petr4 (Empresa de Petróleo e gás), VGIA11(Fundo de investimentos no Agronegócio)
- Cotação Atual - Preço atual da ação ou fundo. Quando o mercado está aberto esse valor oscila a cada segundo.
- Preço Mínimo - Valor mínimo dos últimos 12 meses
- Preço Máximo - Valor máximo dos últimos 12 meses
- Dividend Yield - Percentual de Distribuição dos lucros em forma de dividendos nos últimos 12 meses.
- Liquidez Diária - Volume financeiro médio diário de negociação nos últimos 12 meses
- P/VP - Preço sobre valor patrimonial da empresa ou fundo. Consiste na divisão do patrimônio líquido pelo total de ações disponível no mercado.
- CAGR Lucro - Evolução do lucro nos últimos 5 anos (ações) e 3 anos (fundos)
- Patrimônio Líquido - Valor disponível após calcular as receitas menos as despesas. Neste caso o ano de referência é 2022.
- Caixa - Disponibilidade financeira imediata. Nesse caso o ano de referência é 2022.
- Lucro - Resultado do exercício financeiro, nesse caso o ano de referência é 2022.
- Investir - Define se é pertinente ou não investir na empresa, ou fundo mediante critérios estabelecidos pelos clientes.

Termos de mineração de dados:

- Mineração de dados: processo de extração de informações úteis e conhecimento a partir de grandes conjuntos de dados.
- Técnicas de mineração de dados: métodos matemáticos e estatísticos utilizados para descobrir padrões em dados.

Em relação ao custo-benefício do projeto, levou-se em consideração os custos:

- Mão de obra: contratação de dois cientistas de dados na modalidade freelance para realizar a análise de investimento. Custo médio de 50 euros por hora de trabalho.
- Software - sendo o Python uma linguagem de programação de código aberto, o programa e as bibliotecas desenvolvidas em Python são gratuitos para utilização, da mesma forma o Microsoft Excel não terá custo por já fazer parte do pacote do Microsoft Office.

Em contrapartida, os benefícios são significativos para o grupo de investidores, como uma maior precisão na escolha das ações e fundos para investir, resultando em possíveis ganhos financeiros, em termos de benefícios, a utilização da mineração de dados pode trazer eficiência da tomada de decisão. É importante saber que, a análise dos dados financeiros da BOVESPA (Bolsa de Valores de São Paulo) irá identificar tendências e padrões que não seriam visíveis por meio de análises convencionais, permitindo que os investidores tenham informações precisas com grande potencial de sucesso em suas decisões.

E. Objetivos da Mineração de Dados

Os objetivos de mineração de dados deste projeto estão conforme as expectativas do grupo de investidores. No entanto, de maneira geral, temos como objetivo principal extrair informações relevantes e gerar insights a partir da base de dados investimentos, bem como:

- Identificar oportunidades de investimento com potencial de rendimentos atrativos: através do processo de *Data Mining* iremos analisar os dados financeiros, e identificar ativos com o potencial de oferecer bons retornos financeiros, seja no curto, médio ou longo prazo.
- Prever comportamentos futuros para gerenciar riscos: através do processo de *Data Mining* também permite nos identificar os riscos associados a diferentes ativos e avaliar a probabilidade de perda nos investimentos realizados. Isso será útil para os investidores que buscam gerenciar riscos e minimizar suas perdas.
- Comparar diferentes ativos: Através do processo de *Data Mining* iremos comparar o desempenho de diferentes ativos e avaliar qual deles é mais rentável. Sendo este um dos objetivos dos investidores que buscam diversificar sua carteira de investimentos e maximizar seus retornos financeiros.
- Identificar padrões e tendências, através do processo de Data Mining iremos avaliar o desempenho passado (último ano 2022) de um ativo, o que pode ser usado como um indicador da probabilidade de seu desempenho futuro. Isso pode auxiliar os investidores a identificar padrões e tendências de mercado e a tomar decisões assertivas sobre onde investir seus recursos.

Em resumo, o objetivo da mineração de dados é possibilitar os investidores a tomada de decisões sobre onde investir seu dinheiro, avaliar o desempenho passado e buscar prever com a maior assertividade possível, o futuro dos ativos, comparar diferentes alternativas de investimentos e gerenciar riscos, sendo a mineração de dados muito valiosa para os objetivos almejados por meio deste projeto. Os critérios de sucesso deste projeto de mineração de dados estão atrelados ao negócio, bem como os objetivos da mineração de dados. Definiram-se alguns critérios de sucesso:

- Qualidade dos dados: A qualidade dos dados irá garantir que a nossa análise seja precisa e confiável e para isso é necessário que a base de dados tenham dados limpos, livres de erros e organizados.
- Objetivos mensuráveis: Definiu-se a taxa de acerto de no mínimo 80%.

Seleção adequada de técnicas de mineração de dados: As técnicas de mineração de dados escolhidas para este projeto são:

- Estatística: para analisar dados financeiros ao longo do ano, como preços de ações/fundos e identificar padrões e tendências.
- Classificação: essa técnica pode ser usada para classificar as ações em categorias, como ações de alto ou baixo risco, com base em diferentes critérios, como indicadores financeiros, histórico de preços, volume de negociações, bem como se é recomendado o investimento ou não.
- Regressão: essa técnica pode ser usada para prever o desempenho futuro de uma ação/fundo com base em dados históricos.
- Clusterização: agrupamento e segmentação das informações de setores de investimento.

O sucesso do nosso projeto de mineração de dados depende da aplicação dos insights gerados, o que é importante para garantir que as recomendações geradas através deste estudo sejam implementadas e que as decisões sejam tomadas com base nele.

F. Plano de projeto

Definimos o planeamento do projeto em quatro principais ações:

1. Definição dos objetivos do projeto: Os objetivos listados foram definidos com base na necessidade de negócio dos investidores. Identificar oportunidades de investimento promissoras. Prever comportamentos futuros para gerenciar riscos. Comparar diferentes ativos. Identificar padrões e tendências dos ativos.

2. Definição do escopo do projeto:

a. Os dados a serem analisados incluem registros de ações e fundos imobiliários negociados na bolsa de valores de São Paulo (BOVESPA).

b. As técnicas de mineração de dados aplicadas neste projeto são: análise estatística, classificação, regressão e clusterização.

Atividades e Sub-atividades a serem realizadas para garantir o objetivo do projeto:

- Definição do objetivo;
- Coleta dos dados;
- Limpeza da base de dados;
- Preparação dos dados;
- Análise exploratória dos dados;
- Seleção da técnica de modelagem;
- Avaliação do modelo;
- Implantação do modelo;
- Monitoramento contínuo;
- Construção da visualização e relatório final;
- Apresentação ao cliente.

c. A equipa do projeto terá dois profissionais de dados, podendo ter apoio de um consultor financeiro e um analista de investimentos, esses últimos são da comunidade que solicitou esse projeto.

Cronograma do projeto:

	Prazos			
	2023			
Atividades	**Abril**	**Maio**	**Junho**	**Julho**
Definição do objetivo				
Coleta dos dados				
Limpeza dos dados				
Preparação dos dados				
Análise exploratória dos dados				
Seleção da técnica de modelagem				
Avaliação do modelo				
Implantação do modelo				
Monitoramento contínuo				
Construção relatório final				
Apresentação ao cliente				

Fig. 2. Cronograma

IV. Compreensão dos Dados

A. Recolha dos Dados

A coleta dos dados, a princípio foi realizada pelo grupo de investidores que informou que todos os dados que constam na base de dados foram coletados da BOVESPA (Bolsa de Valores de São Paulo). Após a primeira avaliação dos dados a equipa notou que algumas informações não constavam no dataset e seriam importantes para a realização do projeto, por isso, foi desenvolvida uma automação usando as bibliotecas pyAutogui e Selenium que estão no início do notebook, comentadas, e podem ser consultadas e utilizadas para a atualização das informações, quando houver necessidade. Nessa etapa, foram obtidas informações relevantes acrescidas aos dados, como cotação máxima, mínima e setor, sendo este último fundamental para a realização do projeto.

B. Descrição dos Dados

A base de dados contém 897 instâncias, 14 atributos sendo um desses a variável alvo (Investir Sim ou Não). Os atributos são tanto nominais como numéricos e o alvo pode assumir duas classes, 1 ou 0, que indica S= Sim para recomendação de investimento e N= Não recomendação de investimento, respectivamente.

Dos atributos, 10 são numéricos e 4 são nominais. Os dados obtidos contêm informações financeiras da bolsa de valores de São Paulo e são dados públicos. Com uma análise prévia, já é possível identificar os dados faltantes e necessidade de tratamento nos dados. A base de dados não possuía um atributo alvo no qual foi definido através do cálculo (se liquidez diária for maior que 1000000; CAGR Lucro maior que 0,05; P/VP maior que 0,3;P/VP menor que 1,2; então classifica-se ”SIM”

ao contrário classifica-se ”NÃO”). A Análise Exploratória de Dados será realizada durante a preparação dos dados, e as observações tratadas inicialmente serão analisadas mais profundamente.

C. Exploração dos Dados

A análise exploratória dos dados foi realizada em python e a partir dos objetivos da mineração de dados, será possível que as questões levantadas sejam respondidas analiticamente ou por meio de gráficos.

- Quais ativos com o potencial de oferecer bons retornos financeiros?
- Quais os riscos associados a diferentes ativos e qual a probabilidade de perda de investimento?
- Qual o desempenho de diferentes ativos e qual deles é mais rentável atualmente?
- Quais setores são mais ou menos promissores?

Inicialmente, as análises feitas foram: Verificou-se a quantidade de papéis (ações ou fundos) com recomendação de investimento classificadas como “SIM”, foram identificados 255 papéis. Por outro lado, também verificou-se a quantidade de papéis (ações ou fundos) recomendação de investimento classificadas como “NÃO”, foram identificados 642 papéis.

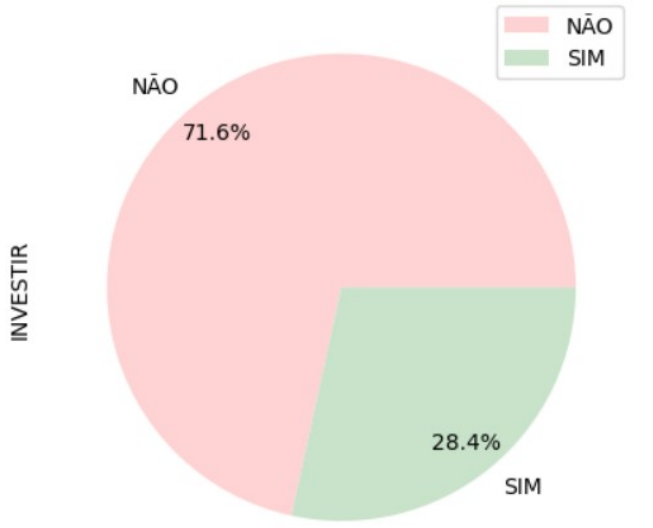


Fig. 3. Investir S N

Obtivemos as primeiras impressões dos dados, através do resumo estatístico. Verificou-se a maior frequência de preços (cotação atual) e a concentração maior está entre 0 e 20 reais (moeda brasileira), pode-se observar que há na base de dados um grande números de ações que podemos classificar como acessíveis devido ao custo baixo para compra. Verificou-se também o dividend yield, no qual observou que o intervalo é de 0 a 20 por cento na maioria dos papéis, o que pode-se concluir que o mercado que paga pouco dividendos (distribuição do lucro). Em relação ao CAGR (evolução do lucro nos últimos 5 anos (ações) e 3 anos (fundos), no qual verificamos a evolução do lucro, a maiores incidências se concentram entre -10 e 10%, o que evidencia poucas empresas e títulos com crescimento considerável no lucro. Aqui já se obtém uma informação relevante, que indica que a maioria das empresas está com prejuízo nos últimos 5 anos, mostrando a necessidade e relevância deste projeto para a mitigação dos riscos.

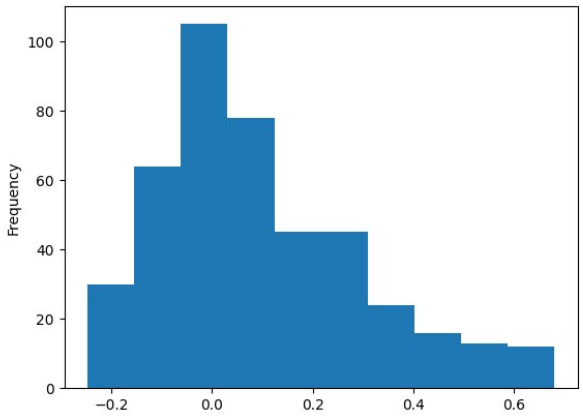


Fig. 4. Evolução do lucro

Após transformar a variável preditora em numérica pode-se buscar a sua correlação com as demais variáveis. Infelizmente, não houve correlação forte para a variável em questão, mas é importante destacar outras correlações que servirão como sugestão ao grupo de investidores para possíveis evoluções desse projeto, a saber: Patrimônio Líquido X Caixa; Luco X Caixa; Liquidez diária X Patrimônio Líquido e Patrimônio Líquido x Lucro. Todas com correlação moderada/alta.

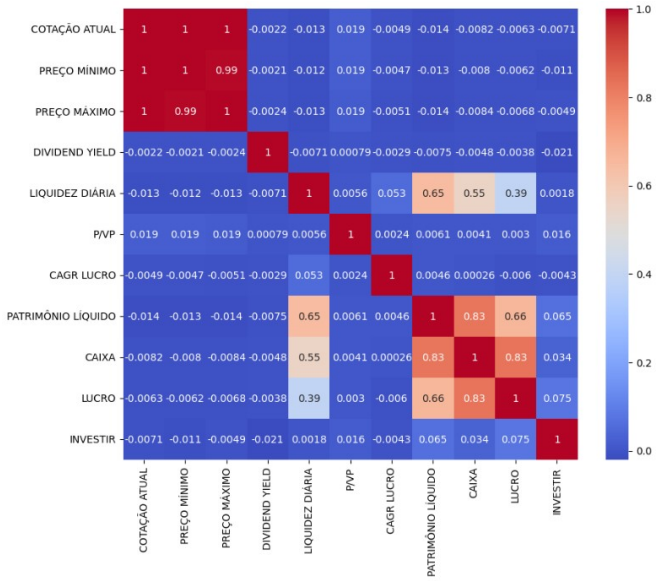


Fig. 5. Correlação das variavéis

Por fim, cabe salientar que foi realizada uma profunda e minuciosa análise por setor tendo uma divisão por ações e fundos, o que revelou alguns pontos críticos, como, por exemplo, a informação que a maioria dos fundos imobiliários tem

lucro próximo a zero ou prejuízo. Isso serve de alerta para que os investidores sejam ainda mais criteriosos nas decisões para essa modalidade de investimentos. Abaixo temos um exemplo crítico para fundos que mostra a evolução negativa do lucro e um exemplo bastante positivo para ações, mostrando força da liquidez do setor de petróleo e afins, onde denota a importância desse trabalho, mesmo num estado ainda incipiente:

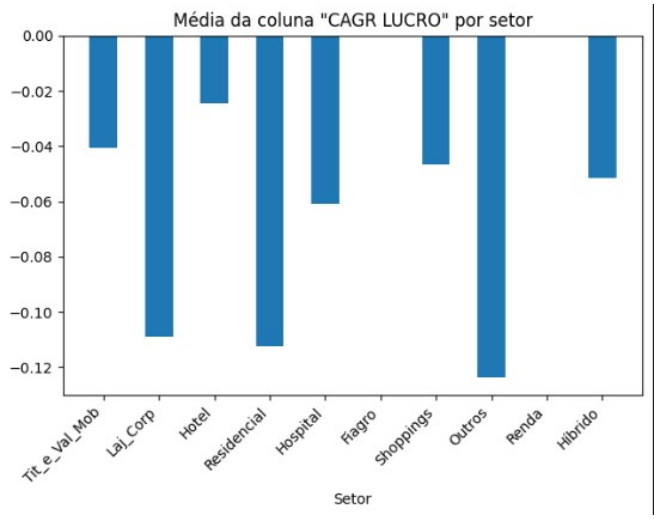


Fig. 6. CAGR Fundos

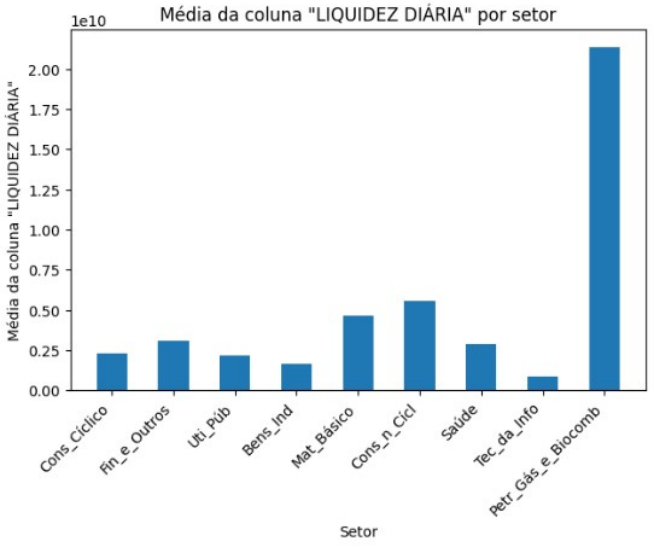


Fig. 7. CAGR Ações

D. Qualidade dos Dados

Inicialmente observamos a qualidade dos dados na base de dados, sendo possível identificar que seria necessário um atributo alvo para de desenvolver as demais análises, bem como a questão final que permeia todo esse projeto que é identificar as ações com mais potencial de rentabilidade e recomendar o investimento ao grupo de investidores. Para solucionar esse problema, criamos uma variável baseada no cálculo proposto pelo grupo de investidores (se liquidez diária for maior que 1000000; CAGR Lucro maior que 0,05; P/VP maior que 0,3;P/VP menor que 1,2; então classifica-se ”SIM” do contrário classifica-se ”NÃO”) assim obtivemos que S= Sim para recomendação de investimento e N= Não recomendação de investimento, respectivamente.

Para realizar a exploração acima destacada, foi preciso realizar algumas análises a fim de verificar a qualidade do dataset, inicialmente apresentou o erro abaixo, devido à base de dados estar no formato csv, para tanto, realizou-se a alteração do formato da base de dados para xlsx.

```
# ParserError: Error tokenizing data. C error: Expected 5 fields in line 3, saw 7
#df=pd.read_csv('projeto_dm.csv')
```

Fig. 8. Erro

Verificou-se também:

- A dimensão da base de dados (897,14)
- Cabeçalho da base de dados
- As últimas instâncias do data set
- Overview das variáveis

Entretanto, nenhum erro de formato foi apresentado. A seguir foram verificados os tipos dos dados para buscar e corrigir inconsistências, sendo necessário transformar atributos de texto em atributos numéricos. Tendo feito essas alterações, foi possível iniciar a análise exploratória dos dados.

V. Preparação dos Dados

Na etapa de preparação dos dados foram realizados os seguintes procedimentos:

- Normalização da variável ”Liquidez diária”: Esta variável tinha um range muito grande de valores, oscilando entre centenas de milhões negativos e dezenas de bilhões positivos. Partindo do princípio que as demais variáveis utilizadas no modelo eram percentuais, estabeleceu-se que a Liquidez receberia valores apenas entre 0 e 1 e isso já melhorou bem a visualização dos dados antes da próxima etapa, mesmo assim ficava clara a presença de outliers que precisariam ser removidos:

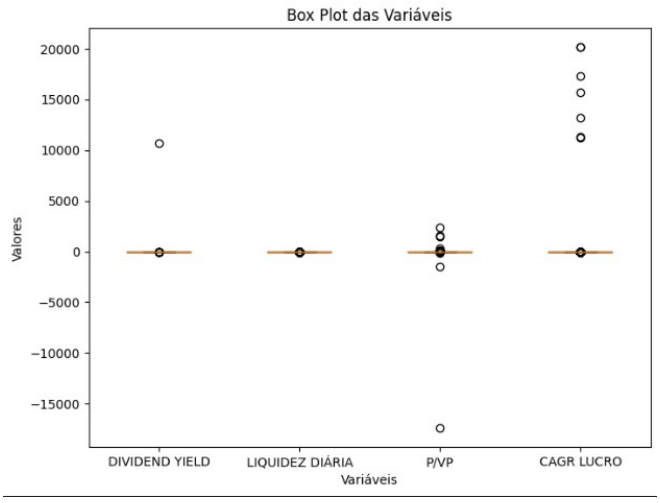


Fig. 9. Outliers antes

- Dando continuidade, foram removidos os outliers, numa das etapas mais trabalhosas do projeto. Inicialmente, foi aplicada a técnica de IQR para remover os outliers, mas as instâncias permaneceram as mesmas, o que entendeu-se não ser o correto, dado o gráfico acima, após isso, foi usada a técnica de Hampel, que deixou apenas 252 instâncias e entender-se que esse número seria insuficiente para as próximas etapas (clusterização e testes de modelos) sendo assim, optou-se pela criação de uma função específica, para a remoção de outliers, def remove$_rows(df1)$: $condition = (df1['DIVIDENDYIELD'] > 1)|(df1['P/VP'] > 2)|(df1['P/VP'] < -0.5)|(df1['CAGRLUCRO'] > 2)|(df1['CAGRLUCRO'] < -1)df_filtered = df1[\ condition]returndf_filtered$
Aplicada a função restaram 736 variáveis, onde esta foi a melhor opção para remoção de outliers. A seguir o gráfico com os outliers removidos:

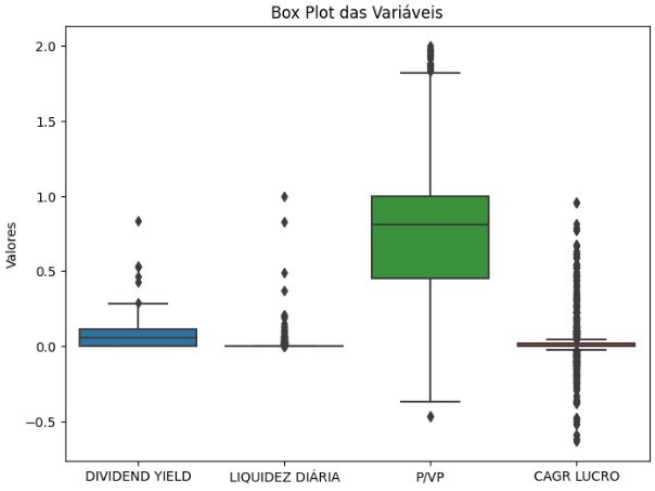


Fig. 10. Outliers depois

- Concluídas as alterações acima e notando-se o claro sucesso desta, o projeto seguiu para a etapa de clusterização.

VI. Modelagem

Nesta etapa, realizou-se a clusterização e iniciou-se o processo de teste dos modelos, que foi concluído apenas na etapa de avaliação devido a esse ser o ponto central do projeto:

- No primeiro momento foi utilizada com sucesso a metodologia de Elbow, que previu a presença de 4 clusters. Em seguida, foi realizada sem sucesso a tentativa de visualização dos clusters com a metodologia k-means. Estima-se que devido à normalização dos dados estes tenham ficado muito concentrados, atrapalhando a visualização, por isso, foi utilizada a técnica de dendograma e silhouette para melhor visualização dos clusters.

  O número de 4 clusters foi confirmado pelo comando AgglomerativeClustering que retornou 4 grupos (0,1,2 e 3), confirmando a quantidade de cluster vista no gráfico de Elbow.

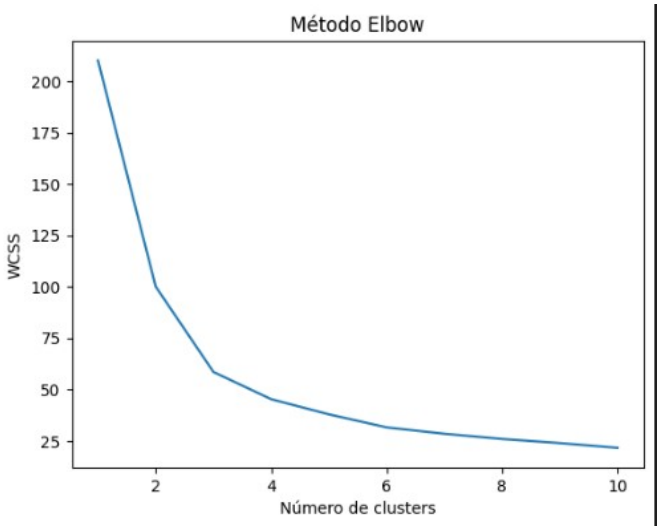


Fig. 11. Gráfico de Elbow

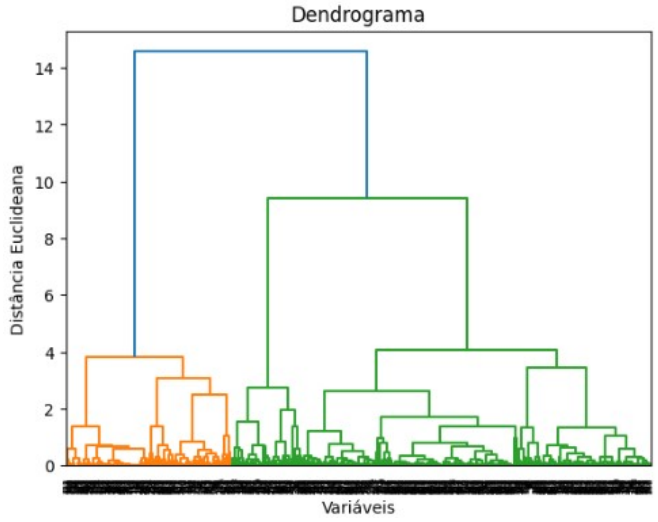


Fig. 12. Gráfico de dendograma

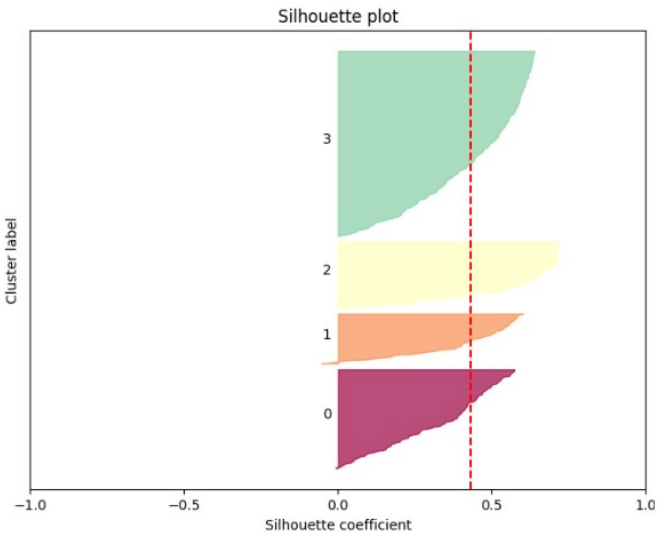


Fig. 13. Gráfico de silhouette

- A próxima etapa talvez seja a mais relevante, já que consiste na criação dos modelos. Foram testados 7 modelos com diferentes resultados. Os resultados serão detalhados na etapa de avaliação, onde agora compete mencionar quais foram os modelos utilizados:
  - Regressão linear
  - Regressão logística
  - Regressão logística com PCA
  - KNN
  - Naive Bayes
  - KPMN
  - Árvore de decisão
- A motivação para escolha desses modelos foi que as variáveis eram todas numéricas, portanto, se poderia testar a regressão linear e as demais técnicas que foram utilizadas em aula ou já testadas pelo aluno Elysiario, em seu trabalho voluntário para a Comunidade Pantokrator. Não foram testados mais modelos, por entender que o resultado era suficientemente bom e nessa altura já havia mais de um modelo com acurácia acima de 80%, que era o objetivo do projeto.

VII. Avaliação

Um aprendizado importante foi desenvolvido nesta etapa. Segundo Vitor Rodrigues a medida mais importante para um caso de investimentos é a precisão, que pode ser calculada pela fórmula (VP/VP+FP), onde VP= Verdadeiro Positivo e FP=Falso Positivo. Entende-se que a presença de falsos negativos não é tão relevante quanto a de falsos positivos, uma vez que, estes podem levar o grupo de investidores a uma decisão de investimentos equivocada e acarretar numa perda financeira. Após essa reflexão pode-se tirar melhor proveito das avaliações dos modelos:

Modelo	Acurácia	Precisão	Revocação
Regressão Linear	nan	nan	nan
Regressão Logística	0.6264	0.4078	0.0667
KNN	0.8221	0.7223	0.8178
Naive Bayes	0.6997	0.6755	0.5595
KMeans	0.1536	nan	nan
Árvore de Decisão	0.9728	0.9616	0.9684

Fig. 14. Comparação dos modelos

Como fica evidente na tabela acima, o modelo com melhor desempenho foi a árvore de decisão, seguida pelo KNN. Importante mencionar que, no critério mais relevante que é a precisão, o KNN não supera os 80% almejados, mas pode-se atualizar facilmente o dataset para poder buscar novos insumos, mas por hora entende-se que deve seguir com a árvore de decisão para a implementação do modelo. A decisão de investir ou não num papel (ação ou fundo) é, sem dúvida, muito relevante e deve ser muito bem refletida, onde, esse trabalho tem também como proposta levar os investidores a perderem o medo do mercado financeiro, obtendo através desses estudos maior segurança em relação aos investimentos em renda variável. Para concluir esta etapa foi realizada a análise da curva roc-auc, que apenas ratifica o que foi mencionado até agora em relação aos modelos utilizados.

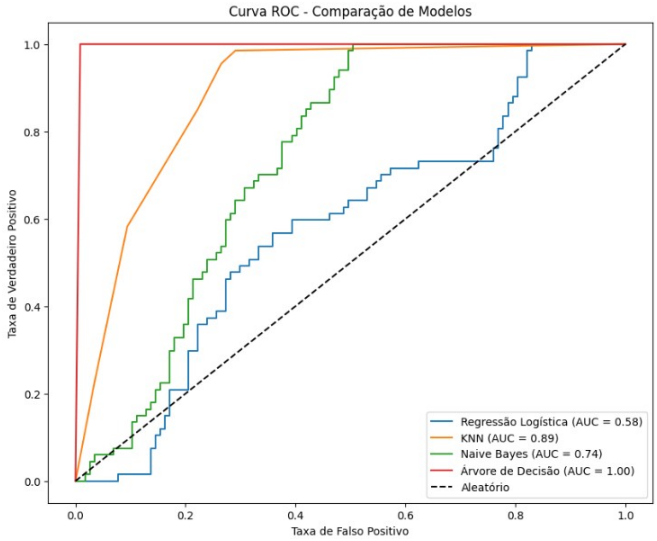


Fig. 15. Curva ROC-AUC

VIII. Implementação

Ao iniciar esta etapa é importante salientar que esta ferramenta será um apoio na tomada de decisão e não deve ser a única a se considerar. Antes de utilizá-la é importante que o cliente tenha conhecimento do mercado e já tenha em mente algumas ações e fundos para colocar seus múltiplos (dados financeiros) na aplicação, de forma a obter mais segurança ao investir ou não num determinado papel.

Na etapa final, foi utilizado o Streamlit para implementar uma aplicação. Foi utilizada uma base de dados reduzida, apenas com as variáveis x e y do modelo, sendo a y a variável preditora. A primeira camada contém o modelo de árvore de decisão, sendo o melhor avaliado, juntamente com os inputs que o cliente poderá realizar. O front-end contém os campos abaixo relacionados:

- Liquidez diária
- P/VP
- Dividend Yield
- Cagr
- Após a inserção desses valores o cliente deverá pressionar o botão de output (enviar) que retornará com a recomendação ou não de investimento.

Essa aplicação, mesmo que simples a princípio pode ajudar muitas pessoas a perderem o medo de investir no mercado de renda variável, que traz consigo um certo grau de risco e também boas oportunidades.

O uso do Streamlit foi um desafio adicional, devido ao fato do desconhecimento desse recurso por parte do desenvolvedor desse projeto, mas com o apoio da docente, Maria Beatriz, e com a consulta aos materiais adicionais disponibilizados por, ela foi possível concluir a aplicação e consequentemente, concluir o projeto com sucesso. A inspiração para o nome da aplicação se dá devido ao aluno Elysiario, ter ao longo do Mestrado, um grande apoio por parte do seu filho que mesmo sendo muito jovem gosta muito de programação e análise de dados, e de um amigo que cursa Estatística e o apoio com dúvidas pontuais nos períodos de provas e trabalhos. Ambos

se chamam Felipe.

Abaixo está a primeira versão da aplicação batizada como Fe²:

Fig. 16. Aplicação Fe²

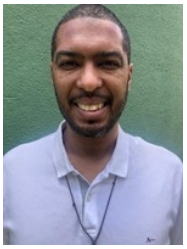
IX. Conclusão

Para a conclusão deste prazeroso projeto cabe destacar alguns pontos de extrema relevância:

- Obtenção de insights importantes: pode-se obter através dos estudos desse projeto uma excelente visão dos riscos do mercado de renda variável. A descoberta do lucro baixo e evolução negativa do lucro se tornam muito relevantes para uma tomada de decisão mais assertiva. Em posse dessas informações, o range de busca por papéis seguros fica consideravelmete menor de mais fácil entendimento, porque os papéis que possuem recomendação de ”não investimento” podem ser descartados com segurança.
- Apoio no âmbito pessoal: O aluno Elysiario Santos tem grande interesse no tema por ser investidor e ensinar educação financeira para seus filhos. Na atualização do projeto que aconteceu com os dados de maio e junho foram situadas alternativas de investimento interessantes que se tivessem sido adquiridas na ocasião, estas já estariam trazendo bons resultados. Na atualização de julho, pretende-se comprar ações e fundos que o estudo sugerir.
- Alcance imensurável do projeto: este conhecimento será compartilhado no github e LinkedIn e não se pode medir o alcance e interesse de pessoas de todo o mundo pelo tema. O tema ”investimentos” está sempre entre os top trends das redes sociais brasileiras e qualquer trabalho que possa ajudar o público a perder o medo de investir certamente poderá ganhar grande relevância.
- Interesse em seguir com uma tese nessa linha: Elysiario tem interesse em realizar sua tese sobre o mercado de renda variável brasileiro, sobretudo, utilizando modelos de predição, análise de grandes volumes de dados e séries temporais. Por isso se espera que esse projeto seja o início bem-sucedido de uma preparação para a conclusão do Mestrado em Ciência de Dados.

References

[1] Sas, “Qual a importância da mineração de dados.” [Online]. Available: //www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html#dmworld
[2] Estatidados, “Crisp-dm (processo padrão inter-indústrias para mineração de dados).” [Online]. Available: http://estatidados.com.br/crisp-dm-processo-padrao-inter-industrias-para-mineracao-de-dados/
[3] IBM, “Visão geral da ajuda do crisp-dm.” [Online]. Available: https://www.ibm.com/docs/pt-br/spss-modeler/18.4.0?topic=dm-crisp-help-overview
[4] X. Investimentos, “Fundos imobiliarios.” [Online]. Available: https://conteudos.xpi.com.br/fundos-imobiliarios/
[5] ToroInvestimentos, “Fundos imobiliarios.” [Online]. Available: https://blog.toroinvestimentos.com.br/bolsa/fundos-imobiliarios-fiis
[6] ——, “Ações.” [Online]. Available: https://blog.toroinvestimentos.com.br/bolsa/o-que-e-acao



Elysiario Virginio dos Santos É licenciado em Administração e Letras, com Pós-Graduação Em Gestão Industrial. Atuou por mais de 10 anos como Professor e Coordenador Pedagógico em concomitância com atuações na área de investimentos e liderança de equipa de customer experience. Há 4 anos começou a estudar e realizar projetos voluntários na área de Business Intelligence e decidiu migrar de carreira para a área de dados, por esse motivo está a cursar Ciência de Dados no Politécnico de Leiria atualmente.