Disponibilizando Dados sobre Resultados Financeiros de Cias Abertas Enriquecidos com Proveniência para a OBInvest

Gilberto Gil F. Gomes Passos[1], Saulo A. Almeida[1], Valquire da S. de Jesus[1], Jorge Zavaleta[1], Sérgio Manuel Serra da Cruz[1]

[1]Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.530 – Rio de Janeiro – RJ – Brasil – 21941-590

{gilbertogilfgp, valquirej, sauloandrade}@gmail.com,

{jorge.zavaleta, serra}@ppgi.ufrj.br

Abstract. *This article describes a study developed based on Data Science techniques, as well as Provenance models and FAIR principles, with the objective of producing a dataset to support OBInvest. For this, open data were used, related to the financial results of publicly traded companies, available on the CVM portal, during the period from 2011 to 2021. Obtaining this dataset made it possible to calculate the Sector Margin metric, which has great relevance in studies of economic growth models.*

Resumo. *Este artigo descreve um estudo desenvolvido com base nas técnicas de Ciências de Dados, assim como, nos modelos de Proveniência e nos princícios FAIR, com o objetivo de produzir um dataset para apoiar a OBInvest: Olimpíada Brasileira de Investimentos. Para isso, foram utilizados dados abertos, relacionados aos resultados financeiros de empresas de capital aberto, disponíveis no portal da Comissão de Valores Mobiliários (CVM), durante o período de 2011 a 2021. A obtenção desse dataset possibilitou o cálculo da métrica Margem do Setor, que possui grande relevância nos estudos de modelos de crescimento econômico.*

1. INTRODUÇÃO

As iniciativas de Educação Financeira, incluindo as que se voltam para jovens, tanto em espaços escolares como em ambientes não formais de ensino, têm sido defendidas e implementadas em vários países, conforme [Aprea et al. 2016], a reboque das ações da Organização para Cooperação do Desenvolvimento Econômico (OCDE), desde 2005 como apresentado em [OECD 2005]. Neste cenário, a abordagem de contextos e noções financeiras e econômicas no currículo de Matemática da Educação Básica tem sido preconizada pelos documentos norteadores nacionais, especialmente com a recente inclusão da Educação Financeira como tema transversal e integrador na Base Nacional Comum Curricular [BNCC 2018].

A [OBInvest 2020] (Olimpíada Brasileira de Investimentos) surge no cenário nacional em agosto de 2020, como um projeto de extensão do [CEFET-RJ 1917], com o objetivo de democratizar o acesso e de promover reflexões acerca de temas e situações econômico-financeiras por meio de uma perspectiva de Educação Financeira para alunos do Ensino Médio de todo o Brasil. A partir de uma lente multidisciplinar e levando

em consideração aspectos didáticos e metodológicos, a OBInvest busca convidar os estudantes a pensar situações e a tomar decisões, que contribuam para o desenvolvimento de habilidades e competências necessárias para a formação crítica, emancipatória e inclusiva do indivíduo, para o pleno exercício da cidadania, e também, para a possibilidade de inserção em um novo mercado de trabalho para os jovens.

Um outro objetivo norteador da Olimpíada, é o desenvolvimento de ferramentas com o intuito de dar acesso de modo facilitado a dados importantes e fundamentais para a tomada de decisão no âmbito de finanças. Assim, tomando um dataset curado e anotado com os metadados de proveniência das demonstrações financeiras das empresas brasileiras de capital aberto, é possível promover aos estudantes e demais interessados em Finanças, o estudo dos comportamentos das séries temporais dos resultados de uma empresa e assim, introduzir uma pesquisa sobre a predição dos resultados futuros.

A ferramenta desenvolvida a partir desse dataset poderá servir para o desenvolvimento de habilidades e competências de jovens talentos interessados em Finanças e Investimentos e poderá ser explorada como uma metodologia ativa pela Olimpíada de Investimentos, preparando profissionais para aprimoramentos e certificações financeiras, bem como o enriquecimento de atividades práticas nacionais da OBInvest.

1.1. Relevância

1.1.1. Demonstração dos Resultados do Exercício (DRE)

Segundo a Comissão de Valores Mobiliários (CVM) [CVM 2017], a Demonstração do Resultado do Exercício (DRE) é um documento contábil onde estão dispostas as apurações de todas as receitas e despesas de uma empresa ao longo de um período, em geral no término de um ano ou de trimestre. Em sua estrutura, a DRE elenca a receita bruta da empresa seguida das deduções contábeis incididas sobre esse valor, até informar, após todas as incorrências, qual foi o lucro ou prejuízo da empresa naquele período. Todas as empresas de capital aberto têm a obrigação de divulgar a DRE e os demais demonstrativos contábeis em períodos trimestrais e no período acumulado ao ano. As demonstrações financeiras divulgadas pelas empresas, e enviadas para a CVM, seguem o padrão internacional de contabilidade, conhecido como IFRS (International Financial Reporting Standards), como expressa a Instrução CVM nº485 [CVM485 2010].

> as demonstrações financeiras consolidadas das companhias abertas deverão ser elaboradas com base em pronunciamentos, plenamente convergentes com as normas internacionais, emitidos pelo Comitê de Pronunciamentos Contábeis – CPC e referendados pela CVM. As demonstrações financeiras consolidadas das companhias abertas serão denominadas "Demonstrações Financeiras Consolidadas em IFRS"

Assim, as demonstrações seguem regras, princípios e fundamentos adotados para unificar os padrões contábeis, e hoje, são mais de 120 países que adotam essa forma padronizada de como a contabilidade é feita [CFC 2010]. Ressaltamos que a CVM disponibiliza os resultados individuais e consolidados, que se diferem ao fato que as demonstrações consolidadas consideram as empresas controladoras e suas subsidiárias [CFC 2016]. Neste trabalho, os experimentos gerados são baseados nas DRE's consolidadas.

1.1.2. Resultados Extraídos

Nas contas disponibilizadas na DRE, podemos destacar a Receita Operacional Líquida, o Custo de Bens e Serviços, cuja diferença é o Resultado Bruto (Lucro Bruto). Temos também as despesas operacionais, cuja diferença para o Resultado Bruto é a conta "Resultado Antes do Resultado Financeiro e dos Tributos", que também é conhecida como LAJIR (Lucro antes de juros e imposto de renda) ou EBIT (Earnings Before Interest and Taxes) e geralmente, é chamado de "lucro operacional" [Ross et al. 2015]. Outro destaque é a linha importante chamada de Resultado Líquido ou de Lucro Líquido.

Exibiremos alguns cálculos importantes neste artigo, um deles é o conceito de Margem. Margem, "em diferentes níveis do balanço, indica o que representam o resultado bruto, operacional e líquido da empresa relativamente à sua receita líquida" [Povoa 2012]. Outros cálculos relevantes são a Análise Vertical, que decompõe o percentual de cada item da DRE em relação à receita naquele período, e a Análise Horizontal que mostra a evolução percentual anual dos números do Demonstrativo de Resultados [Povoa 2012]. Adicionaremos ainda o cálculo das margens de cada setor, que definiremos aqui neste artigo como a porcentagem que representa o somatório dos resultados brutos, operacionais e líquidos das empresas do respectivo setor, relativamente ao somatório das receitas líquidas das empresas do setor.

1.1.3. *Stakeholders*

Muitos são os interessados nas demonstrações contábeis de uma empresa. Chamamos de Stakeholders "qualquer pessoa, entidade ou sistema que afeta ou é afetado pelas atividades de uma organização" [IBGC 2015]. Logo, podemos destacar como stakeholders: os acionistas da empresa, funcionários, clientes, fornecedores, Governo, concorrentes, mídia, Sindicatos, dentre outros.

A busca das informações contábeis é necessária para a análise de empresas feitas por profissionais de Investimentos, e costuma anteceder a tomada de decisão daqueles que pretendem investir numa determinada empresa. Embora as empresas de capital aberto disponibilizem esses resultados nos sites de relação com investidores, muitas vezes, os dados são disponibilizados em formatos distintos, como csv, xls ou pdf e em arquivos separados por ano. Para o stakeholder que pretende acompanhar os resultados trimestrais de mais de uma empresa, este artigo fornecerá um experimento de captação dos dados da DRE enviados à CVM pelas empresas, disponibilizando a série histórica das contas, linha a linha, a partir de 2011. Assim, a partir do dataset gerado será possível gerar a visualização dos seguintes comportamentos ao longo do tempo:

- Desempenho das margens brutas, operacionais e líquidas das empresas;
- Desempenho das margens dos setores ao longo dos trimestres;
- A evolução do lucro líquido e do resultado operacional da empresa ao longo dos trimestres;

Assim, o acesso aos dados para posterior análises dos Stakeholders poderão se tornar mais facilitados de uma maneira geral.

1.2. Estrutura do trabalho

O restante do trabalho é organizado da seguinte forma: Seção 2, apresenta uma lista dos principais trabalhos encontrados que utilizam dados da CVM, bem como alguns repositórios no GitHub. Na seção 3 é apresentado a metodologia utilizada. Na seção 4 é apresentado o dataset OBInvest e informações sobre FAIR, Proveniência e reprodutibilidade. Uma breve discussão é apresentada seção 5. A seção 6 sugere algumas melhorias para trabalhos futuros e na seção 7 é feita a conclusão do trabalho.

2. TRABALHOS RELACIONADOS

Ao levantar trabalhos acadêmicos que se apoiaram em dados disponbilizados pela CVM, os resultados foram pouco expressivos. Em [CHIELLA and RICHARTZ 2019], foi proposta uma análise sobre a geração e a distribuição da riqueza criada pelas empresas listadas na Comissão de Valores Mobiliários, utilizando dados de Demonstrações do Valor Adicionado (DVAs), no período de 2009 à 2017. Em [Catapan and Colauto 2020] foi proposta uma análise comparativa da qualidade dos lucros dos dois padrões contábeis (COSIF e CPC) através de informações disponibilizadas pela CVM e pelo BACEN, no período de 2010 e 2018. [Marques et al. 2020] propôe alguns modelos de *valuation* utilizados por Fundos Mútuos de Investimento de Empresas Emergentes (FMIEE), se apoiando em dados da CVM e em um survey baseado em questionários enviados para os gestores dos fundos da FMIEE. Em [ROVER and BORBA 2006] tentou-se identificar diferenças e semelhanças nas práticas de evidenciação concernentes aos passivos ambientais entre Brasil e Estados Unidos, utilizando Demonstrações contábeis no período de 2002 a 2004.

Na tentativa de identificar uma quantidade maior de trabalhos que utilizassem os dados disponibilizados pela CVM, além da pesquisa por trabalhos acadêmicos, foi realizada uma pesquisa nos repositórios do GitHub, em busca de projetos que utilizassem esse tipo de informação. Os principais projetos encontrados foram o [Quant 2020] que contem uma série de notebooks python para obter diversos tipos de informações financeiras. [dude333 2021] é um projeto feito na linguagem GO Lang da Google, onde é possível obter diretamente da CVM diversas informações através de linhas de comando. O projeto brFinance [Rodrigo 2022] que é uma bibioteca python que pode ser importadas para projetos python, e que busca informações de dados financeiros também diretamente da CVM. O repositório [Vido 2020] contem um conjunto de notebooks python e um deles se apoia nos dados de contas de Demonstrativos de Resultados do Exercício (DREs), o mesmo tipo de dado analisado nesse trabalho.

3. MATERIAIS E METODOLOGIAS

O desenvolvimento deste trabalho se pautou no OSEMN *framework* [Mason and Wiggins 2010], que tem sido muito utilizado em Processos de Ciência de Dados quando se objetiva fazer questionamentos sobre os dados após o fim do processo, o contrário do que é feito no CRISP-DM [Wirth 2000], onde os questionamentos são feitos de forma precoce. Nesse estudo foram desenvolvidas as etapas de Obtenção, Exploração, Modelo (*Workflow*) e Interpretação/Vizualização dos Resultados.

As execuções das etapas foram apoiadas por ferramentas computacionais empregadas para essa finalidade, com destaque para a liguangem de programação Python v.3.9.12, da biblioteca Pandas v1.4.2, da plataforma de *workflow* KNIME v.4.6.1, e dos

ambientes de desenvolvimento integrado Jupyter (executando sobre a plataforma Anaconda3 na versão v2022.5 e Docker v20.10.14). O experimento também foi executado no ambiente de nuvem Google Colabotory.

3.1. O dado bruto

Durante a etapa de Obtenção foram utilizados dados disponíveis no portal de dados abertos da CVM, na área de conjunto de dados, disponível no endereço `https://dados.cvm.gov.br/dataset/`. Esses dados estão marcados com licença aberta (*Open Data Commons Open Database License (ODbL)*) e são gerenciados por meio do *Comprehensive Knowledge Archive Network (CKAN)* [Molloy 2012]. As informações que foram colhidas nesta pesquisa estão listadas no agrupamento de dados do tipo "Companhia", sendo obtidos três conjuntos diferentes, todos em formato *Comma Separated Values(CSV)*:

- *Dataset* Dados Cadastrais de Companhias Abertas (CAD-CIA);
- *Dataset* Formulário de Informações Trimestrais (ITR);
- *Dataset* Formulário de Demonstrações Financeiras Padronizadas (DFP).

3.2. Aquisição dos dados

Os dados utilizados no trabalho são relativos às Compahias de Capital Aberto que são fiscalizadas pela CVM. Em tempo de planejamento do trabalho, a equipe resolveu analizar informação referentes ao período de 2011 a 2021. Por se tratar de um dado histórico, que passou inclusive por processos de auditorias, a equipe entendeu que não haveria uma ganho em baixar essas informações em tempo de execução do experimento, optando por armazenar os datasets em um diretório de dados, no repositório do projeto. Os dados foram coletados e armazenados no dia 16 de setembro de 2022. Foram coletados três tipos de informações:

1. **Dados Cadastrais de Companhias Abertas (CAD-CIA)** - O *dataset* contém as informações cadastrais das empresas de capital aberto e esse tipo de informação, possibilitaria que fossem realizadas análises agrupadas por áreas de atuação de cada empresa. O Conjunto é composto de apenas um dataset em um arquivo de texto simples e aberto no formato *Comma Separeted Values* (CSV). O Arquivo possui 2550 registros;
2. **Formulário de Informações Trimestrais (ITR)** - A CVM disponibiliza um conjunto de Informações referentes às companhias abertas em uma periódicidade trimestral chamada de Formulário de Informações Trimestrais ou simplesmente (ITR). Os dados disponibilizados nesse conjunto de dados são: Balanço Patrimonial Ativo (BPA); Balanço Patrimonial Passivo (BPP); Demonstração de Fluxo de Caixa - Método Direto (DFC-MD); Demonstração de Fluxo de Caixa - Método Indireto (DFC-MI); Demonstração das Mutações do Patrimônio Líquido (DMPL); Demonstração de Resultado Abrangente (DRA); Demonstração de Resultado do Exercício (DRE) e a Demonstração de Valor Adicionado (DVA). Conforme explicado na seção 1.1.1, os dados de interesse do trabalho são relativos aos DREs das empresas. Para esses dados a CVM disponibiliza dois conjuntos de informações, o DRE Individual, que trata os dados das empresas e suas subsidiarias e coligadas de forma indivudal; e o DRE Consolidado, que trata essas informações entre holding e empresas associados de forma consolidada. Para análise de investimento,

tipicamente são usados os dados de resultados de empresas no formato consolidado. Para o intervalo de 11 anos analizados, os consolidados dos arquivos do tipo CSV, de DREs consolidados totalizou 1.654.787 registros.

3. **Formulário de Demonstrações Financeiras Padronizadas (DFP)** - Informações sobre as contas e o resultado financeiro das empresas a cada ano. Também para o intervalo de 11 anos analizados, o consolidado dos arquivos do tipo CSV de DREs do tipo anual (DFP) totalizou 665.666 registros.

3.3. Redundância de informações

Foi possível observar nos dados brutos de DRE uma considerável redundância de informação. Para todos os registros de contas de DRE, tanto nos demonstrativos de períodicidade trimestral quanto no anual, são apresentados dois registros, onde um contém o dado do período vigente, e um segundo registro com a informação do mesmo período, só que do ano anterior.

Na DRE trimestral (ITR), também foi possível perceber durante a primeira análise que existem registros para os trimestres de forma isolada, mas também existem registros para os acumularos do segundo e do terceiro trimestre, o que de certa forma também cria algum tipo de redundância. Outro ponto observado ainda sobre o ITR, é que não era apresentado o registro do último trimestre (do quarto trimestre), e para obter essa informação, seria necessário obter essa dado do dataset do DRE anual (DFP) e subtrair do acumulado do terceiro trimestes existente no dataset da DRE ITR.

3.4. Pipeline dos dados

A primeira parte da etapa de Exploração foi realizada com o apoio da ferramenta KNIME [Knime 2004]. Foi criado um workflow na ferramenta para iniciar o processo exploratório do conjunto de dados do trabalho. Esse fluxo inicial pode ser visto na Figura 1 .

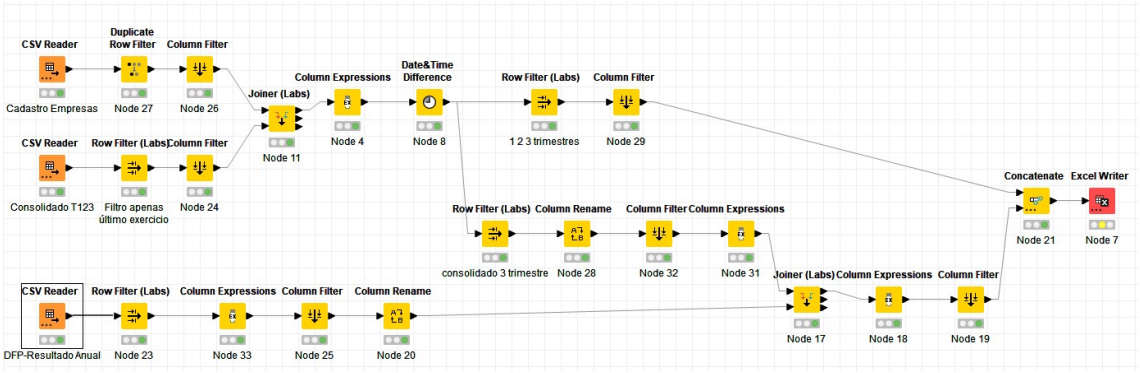

Figura 1. Workflow exploratório montado no KNIME para entender melhor como tratar o conjunto de dados do trabalho.

Na sequência, já utilizando a biblioteca Pandas, foi possível realizar a escolha da estratégia que seria adotada nesse estudo, para montagem de um *dataset* que atendesse ao escopo do trabalho, assim como, agregasse informações que pudessem ser úteis à OBInvest. Dessa forma, foi observado que haveria a necessidade de junção entre o *dataset* CAD-CIA e os *datasets* ITR e DFP, realizando as devidas transformações, composições e agregações julgadas úteis e indispensáveis sobre esses *datasets*. As principais atividades de preparação dos dados desenvolvidas são elecandas a seguir:

- Transformação entre *dataset* CAD-CIA e os *datasets* ITR e DFP com o objetivo de filtrar as empresas com cadastro ativo e agregar a coluna **Setor de Atividade**, ao ITR e ao DFP, oriunda de CAD-CIA;
- Fatiamento do *dataset* ITR, visando à obtenção de dois *datframes*. O primeiro *datframe* contendo informes do primeiro, do segundo e do terceiro trimestre, chamado de TRIM123. O segundo contendo o valor acumulado entre o primeiro e o terceiro trimestre referentes ao ano de vigência do exercício financeiro, denominado ACM3;
- Fatiamento do *dataset* DFP, com o propósito de filtrar dados duplicados e adquirir um *datframe* composto pelo valor acumulado entre o primeiro e o quarto trimestre, chamando de ACM4;
- Subtração entre os valores da coluna **Valor da Conta** dos *datframes* ACM4 e TRIM123, resultando no *datframe* que contém o quarto trimestre, denomindado TRIM4;
- Concatenação entre os *datframes* TRIM123 e TRIM4, obtendo assim o *dataset* OBInvest.

Ainda durante o processo de preparação foram agregadas mais duas colunas, **Ano** e **Trimestre**. Houve também a necessidade de padronização dos registros contidos na coluna **Valor da Conta**, permancendo todos com valores múltiplos de mil. As etapas do fluxo de trabalho desenvolvido podem ser vistas na Figura 2.

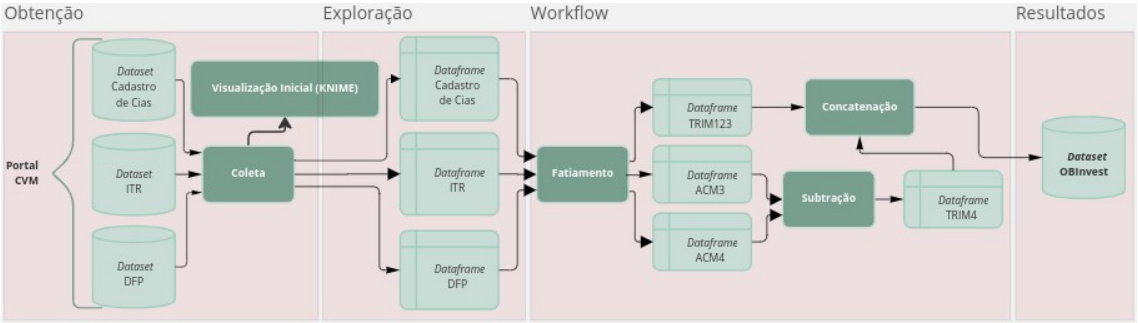

Figura 2. Modelo do fluxo de trabalho desenvolvido.

3.5. Problemas encontrados nos *datasets* brutos

Após o processo de Obtenção, Exploração e *Workflow* foram observadas algumas anomalias nos registros de todos os *datasets* ITR e DFP, que impossibilitaram o uso dos dados em sua totalidade, haja vista que o portal da CVM informa que há uma rotina de atualização semanal dos *datasets*, o que não foi sufiente para evitar as incosistências elencadas a seguir:

- Empresas que possuem exercícios financeiros irregulares, cujo primeiro trimestre do exercício não coincide com o primeiro trimestre do ano;
- Empresas não enviaram os informes trimestrais por serem isentas, impossibilitando o cálculo do desempenho financeiro;
- Empresas somente enviaram os informes financeiros no *dataset* DFP, não sendo possível a definição do trimestre financeiro.

3.6. Discussão sobre os dados, limpeza e vinculação de registros

A estratégia adotada neste estudo foi desconsiderar os dados enviados pelas empresas que se enquadraram nos problemas listados na subseção 3.5. o dataset após todos os tratamentos tem uma quantidade de registros bem inferior aos númneros iniciais. O principal fator responsável por essa diminuiçao dos registros é a redundância de informações existentes nos dados brutos, como explicado na seção 3.3. Além da redundância , após o tratamento de todos os problemas encontrados nos *datasets* houve um redução tanto quantidade de empresas quanto na quantidade de registros totais, mas não impossibilitando a continuidade do estudo. As seguintes reduções foram constatadas no *dataset* OBInvest:

- Os *datasets* ITR e DFP registraram, respectivamente, 621 e 623 empresas distintas. Esse quantitativo recuou para 425 empresas;
- O somatório de registros únicos encontrados nos *datasets* ITR e DFP foi de 1.637.176, sendo esse valor reduzido para 413.641;

Uma possível solução para os problemas, encontrados nos registros anômalos das empresas, seria fazer um tratamento de forma individual desses registros, de forma que posssibilitasse a inclusão desses dados no *dataset* OBInvest, o que não faz parte do escopo deste trabalho. Mas que poderiam ser implementada em futuros trabalhos.

3.7. Análise e Vizualização dos Resultados

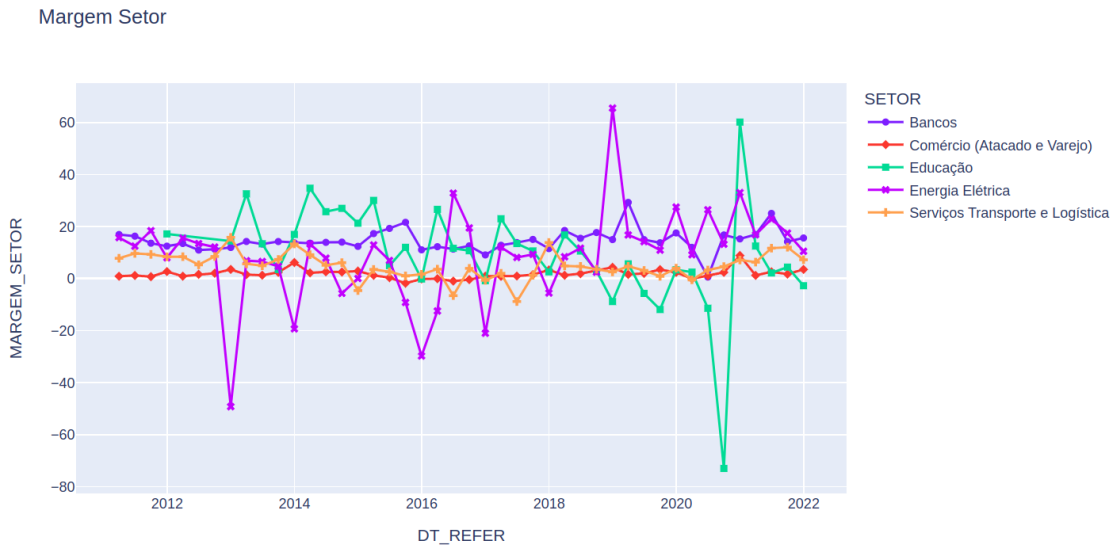

Figura 3. Margem Líquida Setorial

Como exemplo de uma visualização de dados, trazemos uma visualização da margem líquida dos setores. Com um procedimento análogo podemos determinar a margem operacional e a margem bruta tomando respectivamente as contas referentes ao lucro bruto e operacional ao invés do lucro líquido. A Figura 3 refere-se à série histórica da margem líquida trimestral dos setores aos quais as empresas estão classificadas de acordo com a CVM. Essa margem líquida dos setores foi obtida como a divisão do somatório dos Lucros Líquidos pelo somatório das Receitas Líquidas, por trimestre, das empresas que compõem determinado setor. Esse é um resultado que poderá rastrear oportunidades de investimentos em setores que indicam lucratividade crescente, e isso pode ser de grande valia para investidores de um modo geral.

4. DATASETS CURADOS, REPRODUTÍVEIS E ENRIQUECIDOS COM PROVENIÊNCIA DE METADADOS

O dataset OBInvest contendo os DREs trimestrais das empresas de capital aberto listados na CVM é obtido como resultado do experimento. Procurou-se seguir as melhores práticas relacionadas aos princípios F.A.I.R. Além disso, foi feita uma descrição mais detalhada do experimento, baseada em proveniência, na tentativa de garantir tanto a interoperabilidade quanto a reprodutibilidade.

Essa seção do artigo apresenta os detalhes sobre a condução da construção da proveniência e da reprodutibilidade. A versão executável do experimento pode ser acessada através do *doi*: `https://doi.org/10.5281/zenodo.7114963` e reproduzida através da utilização de container Docker.

4.1. Dicionário de dados

O Dicionário de dados do dataset OBInvest gerado pela ao final do pipilene dos dados difere muito pouco do dicionário fornecido nas informações de metadados que constam no site de CVM. Alguns poucos campos do dicionário da DRE da CVM foram suprimidos, e os campos Ano e Trimestre foram adicionados. O Dicionário pode ser observado em maiores detalhes na tabela 1.

4.2. Aplicando FAIR

Em [Wilkinson et al. 2016], uma série de questionamentos relevantes, mas ao mesmo tempo óbvios, sobre a urgência de mudança na forma como as pesquisas deveriam ser conduzidas, para garantir com clareza e facilidade no reuso dos trabalhos acadêmicos. A preocupação esta relacionada tanto com a execução quanto com os dados utilizados nas pesquisas. O trabalho deixa em voga uma série de reflexões na comunidade científica.

Os princípios, em inglês, Findable, Accessible, Interoperable e Reusable, criando o acrônimo F.A.I.R.(Recuperável, Acessível, Interoperável e Reutilizável), a partir de 2017, passaram a ser organizados e promovidos pelo movimento GO FAIR, em diversos países do mundo, incluindo o Brasil.

O experimeto realizado nesse artigo, tentou seguir os princípios definidos em [GO-FAIR 2017] durante a sua criação. A tabela 2 apresenta o entendimento sobre quais os princípios que foram atingidos durante o processo. Na coluna de Situação, foi identificado como ”Implementado”, apenas os princípios que a equipe julgou que houve uma implementação completa. Para aqueles em que permaneceram dúvidas sobre a completude da implementação, optou-se por marcá-los como ”Não implementado”.

4.3. Proveniência no Dataset Curado

A proveniência gerada, tentou representar, de forma mais detalhada possível, todo o ciclo de vida do experimento: desde a origem e a localização dos dados utilizados; todos os agentes envolvidos; e um passo a passo da execução do experimento, onde os metadados que permitem a interoperabilidade e imagem da proveniência são criados em tempo de execução do experimento.

A proveniência foi criada utilizado a biblioteca python PROV v2.0.0, que respeita o modelo estabelecido pela padrão [PROV-Overview 2013] mantido pela W3C. Houve

Dicionário de dados do Dataset OBInvest			
Nome campos	**Descrição**	**Tipo**	**Categoria**
CNPJ_CIA	CNPJ da companhia	varchar(20)	
DT_REFER	Data de referência do documento	date(10)	AAAA-MM-DD
DENOM_CIA	Nome empresarial da companhia	varchar(100)	
CD_CVM	Código CVM	char(6)	
GRUPO_DFP	Nome e nível de agregação da demonstração	varchar(206)	
MOEDA	Moeda	varchar(100)	
ESCALA_MOEDA	Escala Monetária	varchar(100)	
ORDEM_EXEC	Ordem do exercício social	varchar(9)	
DT_INI_EXERC	Data início do exercício social	date(10)	AAAA-MM-DD
DT_FIM_EXERC	Data fim do exercício social	date(10)	AAAA-MM-DD
CD_CONTA	Código da conta	varchar(18)	
DS_CONTA	Descrição da conta	varchar(100)	
VL_CONTA	Valor da conta	decimal(29,10)	
ST_CONTA_FIXA	Indica se é conta fixa ou não	varchar(1)	'S':Sim; 'N':Não
ANO	Ano da data de referência do documento	decimal(4,0)	
SETOR_ATIV	Setor de atividade	varchar(100)	
TRIMESTRE	Trimestre data de referência do documento	decimal(1,0)	

Tabela 1. Representação do dicionário de dados de DRE trimestral OBInvest

Princípios da iniciativa GO F.A.I.R.	
Princípio	**Situação**
(F) – Findable – recuperável	
(F1) para metadados devem ser atribuídos identificadores globais, persistentes e identificáveis	Implementado
(F2) os dados são descritos fazendo uso de metadados enriquecidos.	Implementado
(F3) os metadados incluem clara e explicitamente os identificadores dos dados que são descritos.	Implementado
(F4) os metadados são registrados ou indexados por intermédio de um recurso pesquisável.	Não implementado
(A) – Accessible – acessíveis	
(A1) Os metadados são recuperáveis pelos seus identificadores usando se um protocolo de comunicação padronizado.	Implementado
(A1.1) O protocolo é livre, aberto e universalmente implementável.	Implementado
(A1.2) O protocolo permite procedimentos de autenticação e autorização, quando necessário.	Implementado
(A2) Os metadados são acessíveis mesmo quando os dados não estão mais disponíveis.	Implementado
(I) – Interoperable – interoperáveis	
(I1) os metadados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento.	Implementado
(I2) Os metadados usam vocabulários que seguem os princípios FAIR.	Não implementado
(I3) Os metadados incluem referências qualificadas para outros metadados;	Implementado
(R) – Reusable –reutilizáveis	
(R1) os metadados são descritos com uma pluralidade de atributos precisos e relevantes.	Não implementado
(R1.1) os metadados são liberados com licenças de uso de dados claras e acessíveis.	Implementado
(R1.2) os metadados estão associados à proveniência detalhada.	Implementado
(R1.3) os metadados fazem parte de domínios com uso de padrões compartilhados em comunidades	Implementado

Tabela 2. Princípios FAIR implementados

um grande cuidado em utilizar namespaces que pudessem atribuir semântica em cada etapa do processo de criação da proveniência. A proveniência pode ser visualizada e analisada em três partes principais.

A primeira parte, apresentada nas figuras 4 e 5, tenta demonstrar a origem das informações, disponibilizadas pelo agente da CVM, como ela é disponibilizada e como foi armazenada em nosso repositório.

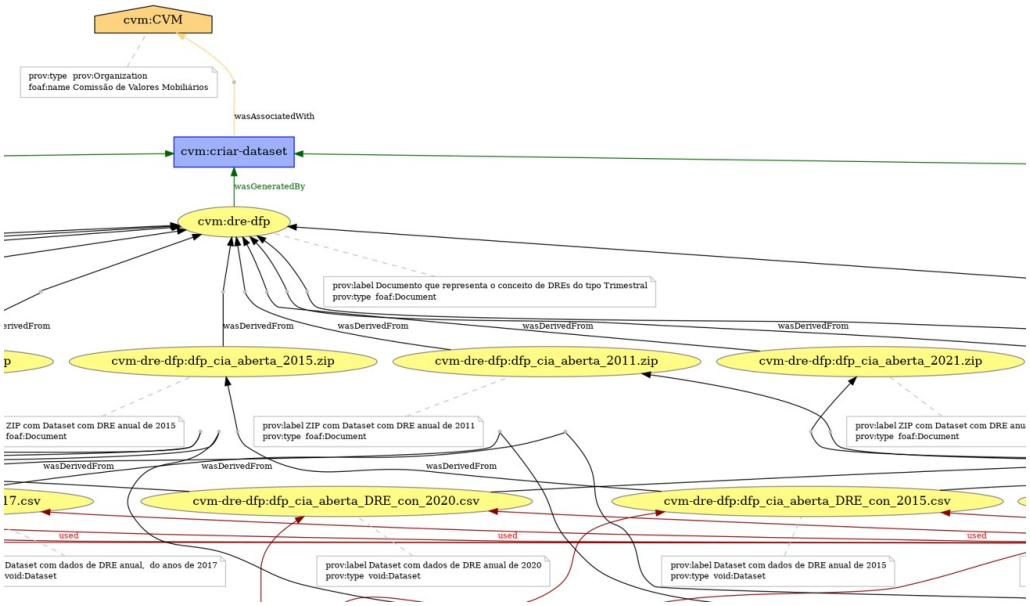

Figura 4. Agente CVM disponibilizando datasets utilizados no experimento.

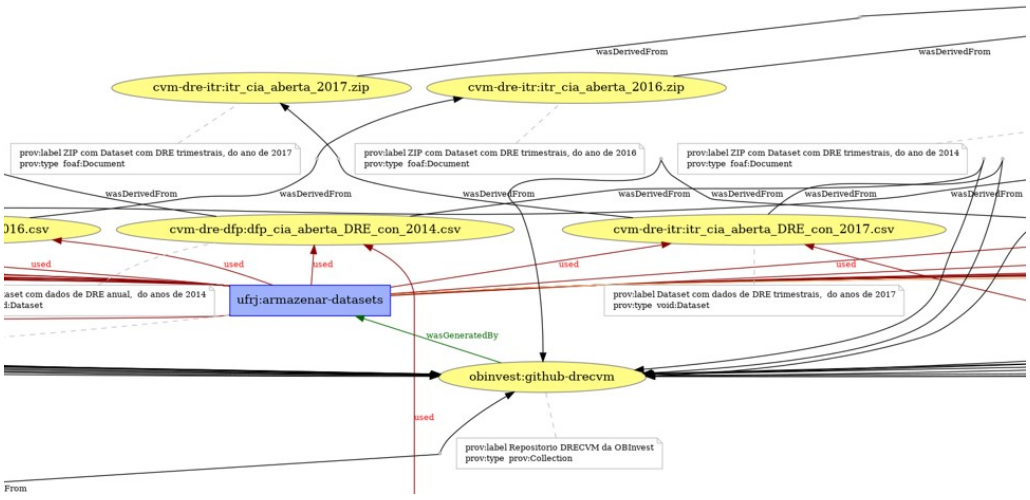

Figura 5. Datasets sendo armazenados no repositório do Github da OBInvest.

Na Segunda parte apresentada na figura 6, pode-se visualizar a hierarquia de agentes envolvidas na definição e execução do experimento. Finalmente, na terceira e última parte, apresentada na figura 7, foi criado um agente referente ao notebook python do experimento, do tipo ”Agente de Software”que detalha toda a execução do experimento, incluindo os timestamps de cada etapa durante a sua execução.

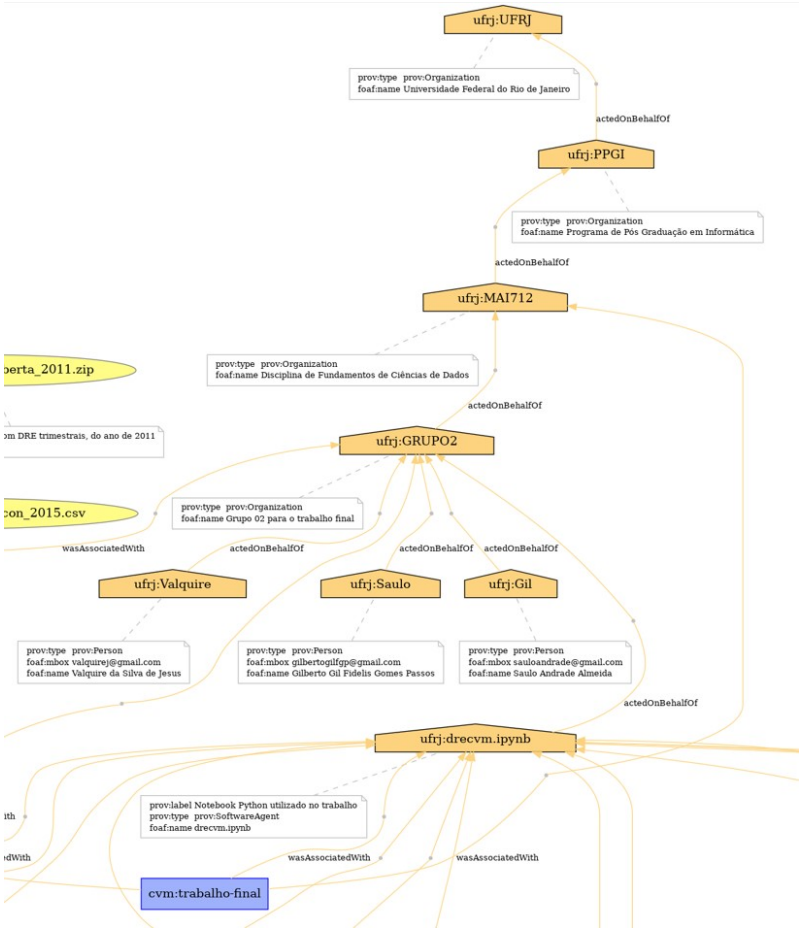

Figura 6. Hierarquia dos agentes envolvidos no experimento.

4.4. Reprodutibilidade do experimento

Um outro ponto importante que foi tratado durante o desenvolvimento do trabalho, estava ligado a ideia de repetibilidade e reprodutibilidade.

Quando pensamos em repetibilidade e reprodutibilidade, a ideia de executar o experimento em ambientes de nuvem, como o Google Colaboratory ou o MyBinder, sempre desperta muita atenção. Entretanto, alguns questionamentos surgiram para os integrantes da equipe. Os principais estavam relacionados à falta de garantia que esse tipo de ambiente oferece, sobre a evolução das versões da linguagem e das bibliotecas, e como isso poderia impactar na repetibilidade futuramente.

Dessa forma, optou-se por investir em ações de repetibilidade e reprodutibilidade utilizando containeres Docker e ambientes onde a versão da linguagem e das principais bibliotecas seriam gerenciados pelo Conda.

A ideia consistiu em criar uma imagem Docker, de tal maneira que pudesse ser contralada qual a versão exata do python que seria utilizada, bem como as versões das principais bibliotecas utilizadas no experimento. Todas definidas de forma explícita, garantindo dessa forma um ”congelamento”tanto do experimento quanto de todo o ambiente utilizado.

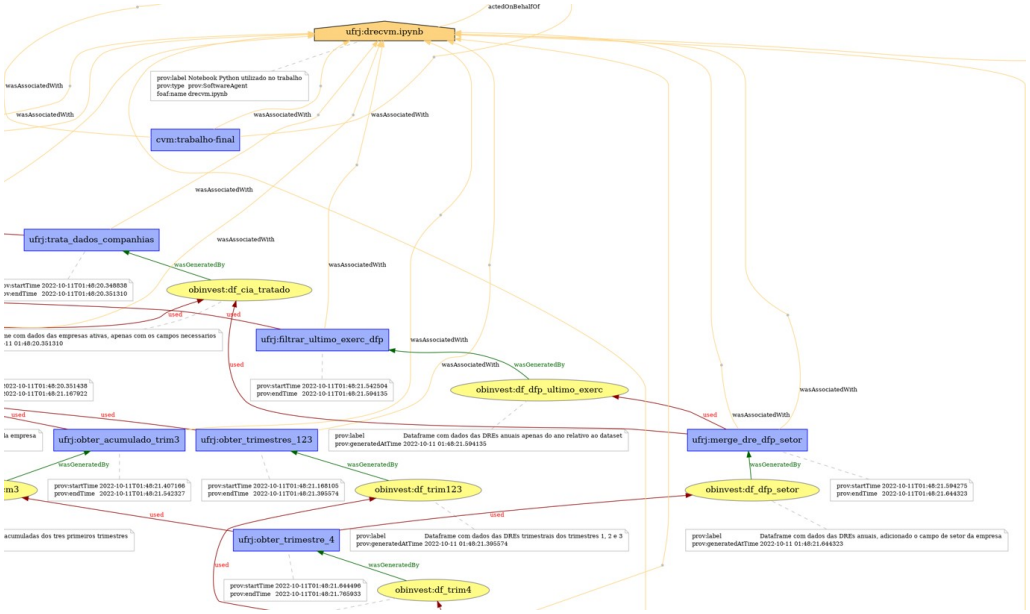

Figura 7. Detalhamento das operações do experimento, criada em tempo de execução.

Para isso, durante a criação da imagem docker, alguns dos passos foram focados na definição de um environment Conda, amarrando as versões do python e das principais bibliotecas utilizadas. Outra verificação criada pela equipe, foi a criação de um método que é executado antes do código principal do experimento, onde são verificados o ambiente Conda e se as versões esperadas das bibliotecas são as concontradas no ambiente Jupyter.

Um outra etapa durante a criação da imagem docker, é o clone do repositório do github da OBInvest na versão específica que foi definida (utilizando uma tag do controle de versão).

O diretório onde o repositório é clonado, serve como caminho para a instância do Jupyter que será executado durante a execução dessa imagem e será usado durante a criação do container docker, de onde é possível reproduzir todo o experimento (incluindo a criação do dataset OBInvest e a criação dos metadados de proveniência em tempo de execução).

A imagem Docker criada foi enviada para o Docker Hub, um repositório remoto de imagens Docker, que permite que um container de uma imagem seja baixado e executado de forma simplificada. Bastando para isso executar o comando: *docker run -p 8888:8888 obinvest/drecvm*.

5. DISCUSSÃO

As informações de Demonstração de Resultados trimestrais oferecidos como resultado na criação do Dataset curado da OBInvest não tinha sido encontrado em nenhum dos trabalhos relacionados. Mesmo com a exclusão das situações de exceção apresentadas na seção 3.5, entendemos que o conjunto oferecido vai auxiliar uma série de atividades da OBInvest.

Apesar da falta de maturidade com os princípios GO FAIR e a falta de experiência com questões relacionadas a reprodutibilidade utilizando containeres, entendemos que o trabalho foi razoavelmente bem sucedido nesses aspectos. É possível de forma simples e clara, executar o procedimento a partir de qualquer computador com conexão para a Internet. Uma vez que esse ambiente tenha sido executado uma primeira vez, o trabalho pode inclusive ser utilizado de maneira *off-line* nas próximas execuções.

Os metadados gerados durante o processo de desenvolvimento do trabalho, conseguiram representar de forma detalhada, qual a origem dos dados, quais os principais agentes envolvidos, bem como o passo a passo detalhado do experimento, onde toda essa informação que possibilita a interoperabilidade é criada em tempo de execução nos formatos RDF Turtle, XML além de apresentada em uma imagem em formato png.

A análise de margem de setor, apesar de simples, ilustra de maneira clara, o potencial de informação que pode ser obtida do dataset criado no experimento.

6. Trabalhos futuros

Entende-se que diversas melhorias podem ser incorporadas ao trabalho apresentado, algumas sugestões de trabalhos futuros são elencandas a seguir:

- Realizar análises mais substanciais sobre o dataset criado, como por exemplo a predição de resultados baseado no histórico dos demostraivos de resultados;
- Tratar caso a caso, as situações de exceção de empresas que apresentaran resultados de DREs fora do padrão, conforme apresentado na seção 3.5 ,para que as mesmas possam ser adicionadas ao Dataset OBInvest;
- Obter outros tipode se informações disponibilizadas pela CVM, para serem utilizados durantes as atividades da Olimpíada Brasileira de Investimento;
- Adaptar a imagem docker para que seja possível executar a mesma no ambiente de nuvem da My Binder, visando a mlehoria ds questões de reprodutibilidade. Como o ambiente My Binder já trabalha sobre o ambiente de containeres, esse forma de execução iria propor um ambiente completamente reprodutível e executável diretamente de um ambiente de nuvem.

7. Conclusão

O trabalho proposto nesse artigo teve por objetivo criar e disponibilizar um dataset com informações de Demonstração de Resultados trimestrais de Empresas de capital aberto, originalmente disponibilidados pela CVM, para que pudesse ser utilizado nas atividades da OBInvest. Como forma de demostração do potencial do OBInvest, foi feita uma análise de resultados trimestais de margem de setor, agrupadas por setor industrial, para exemplificar de maneira simplificada, como essas informações podem ser exploradas.

Durante o desenvolvimento do trabalho, todas as atividades tentaram seguir aos princípios GO FAIR e houve uma grande preocupação em garantir que o experimento pudesse ser reproduzido de forma simplificada. Para isso, tanto o experimento quanto o ambiente utilizado foram empacotados através da utilização de imagens Docker e com apoio do Gerenciador de ambientes do Conda.

Acredita-se que esses objetivos foram atingidos durante a execução do trabalho.

Referências

Aprea, C., Wuttke, E., Breuer, K., Koh, n., Davies, P., Greimel-Fuhrmann, B., and Lopus, J. (2016). *International Handbook of Financial Literacy*.

BNCC (2018). Base nacional comum curricular. Disponível em: http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf, Acessado em: 20/09/2022.

Catapan, A. and Colauto, R. D. (2020). Governança corporativa: uma análise de sua relação com o desempenho econômico-financeiro de empresas cotadas no brasil nos anos de 2010–2012. *Contaduría y administración*, 59(3):137–164.

CEFET-RJ (1917). Centro federal de educação tecnológica celso suckow da fonseca. Disponível em: http://www.cefet-rj.br, Acessado em: 20/09/2022.

CFC (2010). Ifrs ganha espaço e estará em vigor em 140 países num prazo de dois anos. Disponível em: https://cfc.jusbrasil.com.br/noticias/2448777/ifrs-ganha-espaco-e-estara-em-vigor-em-140-paises-num-prazo-de-dois-a Acessado em: 20/09/2022.

CFC (2016). Nbc tsp 6 - demonstrações consolidadas e separadas. Disponível em: https://cfc.org.br/wp-content/uploads/2016/02/NBC_TSP_6_Demonstracoes_Consolidadas_e_Separadas.pdf, Acessado em: 20/09/2022.

CHIELLA, F. and RICHARTZ, F. (2019). Mfc283-análise da geração e distribuição do valor adicionado das empresas registradas na cvm nos anos de 2009 a 2017.

CVM (2017). *ANÁLISE DE INVESTIMENTOS - Histórico, Principais Ferramentas e Mudanças Conceituais para o Futuro*. Disponível em: https://www.investidor.gov.br/portaldoinvestidor/export/sites/portaldoinvestidor/publicacao/Livro/livro_TOP_analise_investimentos.pdf, Acessado em: 20/09/2022.

CVM485 (2010). InstruÇÃo cvm nº 485. Disponível em: https://conteudo.cvm.gov.br/export/sites/cvm/legislacao/instrucoes/anexos/400/inst485.pdf, Acessado em: 20/09/2022.

dude333 (2021). Rapina. Disponível em: https://github.com/dude333/rapina, Acessado em: 12/10/2022.

GO-FAIR (2017). Go fair. Disponível em: https://www.go-fair.org/fair-principles/, Acessado em: 12/10/2022.

IBGC (2015). Código das melhores práticas de governança corporativa. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4382648/mod_resource/content/1/Livro_Codigo_Melhores_Praticas_GC.pdf, Acessado em: 20/09/2022.

Knime (2004). Knime workflow. Disponível em: https://www.knime.com/, Acessado em: 12/10/2022.

Marques, V. A., da Cunha, J. V. A., and do Carmo Mário, P. (2020). Modelos de valuation utilizados pelos fundos mútuos de investimentos em empresas emergentes (fmiee).

Mason, H. and Wiggins, C. (2010). A taxonomy of data science. Disponível em: https://https://web.archive.org/web/20211219192027/http://www.dataists.com/2010/09/a-taxonomy-of-data-science/, Acessado em: 26/09/2022.

Molloy, J. C. (2012). The Open Knowledge Foundation: Open Data Means Better Science. Working Papers id:4686, eSocialSciences.

OBInvest (2020). Olimpíada brasileira de investimentos. Disponível em: https://www.obinvest.org, Acessado em: 20/09/2022.

OECD (2005). Recommendation on principles and good practices for financial education and awareness. Disponível em: https://www.oecd.org/finance/financial-education/35108560.pdf, Acessado em: 20/09/2022.

Povoa, A. (2012). *Valuation - Como Precificar Acoes*.

PROV-Overview (2013). An overview of the prov family of documents. Disponível em: https://www.w3.org/TR/prov-overview/, Acessado em: 12/10/2022.

Quant, C. (2020). Python para investimentos. Disponível em: https://github.com/codigoquant/python_para_investimentos, Acessado em: 12/10/2022.

Rodrigo, E. (2022). brfinance. Disponível em: https://github.com/eudesrodrigo/brFinance, Acessado em: 12/10/2022.

Ross, S., Westerfield, R., Jaffe, J., and Lamb, R. (2015). *Administração Financeira*.

ROVER, S. and BORBA, J. A. (2006). A evidenciação das informações ambientais nas demonstrações contábeis das empresas que atuam no brasil e que negociam adr's na bolsa de valores dos estados unidos: uma análise das dfp's (cvm) e do relatório 20-f (sec). In *VI CONGRESSO USP DE INICIAÇÃO CIENTÍFICA EM CONTABILIDADE*, volume 6.

Vido, L. (2020). Dre cvm. Disponível em: https://gist.github.com/Vido/cbc33862dd27a22790df633f1d113ae6, Acessado em: 12/10/2022.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Wirth, R. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39.