

FACULDADE IMPACTA DE TECNOLOGIA

IGOR SAKAKI DA SILVA

THIAGO LOPES DE MELO

VALDEMIR MARQUES SEVERIANO

WESLEY ANTONIO OLIVEIRA DA SILVA

**APLICAÇÃO DE ALGORITMO NÃO SUPERVISIONADO PARA
IDENTIFICAR LOCAIS DE RISCO E ALVARÁS SUSPEITOS**

SÃO PAULO

2019

2019

**APLICAÇÃO DE ALGORITMO NÃO SUPERVISIONADO PARA IDENTIFICAR
LOCAIS DE RISCO E ALVARÁS SUSPEITOS**

FIT

IGOR SAKAKI DA SILVA

THIAGO LOPES DE MELO

VALDEMIR MARQUES SEVERIANO

WESLEY ANTONIO OLIVEIRA DA SILVA

**APLICAÇÃO DE ALGORITMO NÃO SUPERVISIONADO PARA
IDENTIFICAR LOCAIS DE RISCO E ALVARÁS SUSPEITOS**

Trabalho Projeto de Big Data Analytics
para obtenção de Pós-Graduação Lato
Sensu em Business Intelligence com Big
Data da Faculdade Impacta de
Tecnologia.

Orientador: Prof. Esp. Thiago
Rolemberg.

SÃO PAULO

2019

IGOR SAKAKI DA SILVA
THIAGO LOPES DE MELO
VALDEMIR MARQUES SEVERIANO
WESLEY ANTONIO OLIVEIRA DA SILVA

**APLICAÇÃO DE ALGORITMO NÃO SUPERVISIONADO PARA
IDENTIFICAR LOCAIS DE RISCO E ALVARÁS SUSPEITOS.**

Trabalho Projeto de Big Data Analytics para obtenção de Pós-Graduação Lato Sensu em Business Intelligence com Big Data da Faculdade Impacta de Tecnologia.

COMISSÃO JULGADORA:

Prof.

Faculdade Impacta de Tecnologia

Prof.

Faculdade Impacta de Tecnologia

Prof. Orientador

Faculdade Impacta de Tecnologia

Prof. Orientador. Thiago Rolemberg – Presidente da Banca Examinadora

São Paulo, _____

Dedicamos aos professores que de alguma forma influenciaram na orientação do desenvolvimento deste projeto, em especial ao Professor Rolemberg por todo o tempo e paciência disposto ao nosso auxílio.

Agradecemos aos familiares e amigos pelo continuo apoio durante o tempo de ausência do desenvolvimento deste projeto.

RESUMO

Dado os últimos acontecimentos envolvendo acidentes e tragédias em estabelecimentos públicos e privados, tivemos o museu Nacional do Rio de Janeiro que foi destruído pelo fogo em um grande incêndio onde foram perdidos 20 milhões de itens, cerca 90% de todo seu acervo. Com mais de 200 anos desde sua fundação, essa enorme perda repercutiu em todos os tipos de mídia e logo após o incêndio ser contido, foi constatado que o Museu estava com seu laudo de vistoria dos bombeiros vencido. Percebemos uma particularidade envolvendo esse acidente do museu com outros acidentes, em que todos estavam com o alvará de funcionamento vencidos e em alguns casos, diferente do Museu, tiveram um grande número de vítimas fatais.

O presente trabalho tem como objetivo a criação de um modelo que irá possibilitar a categorização de alvarás de funcionamento do estado de São Paulo, onde civis, instituições privadas e até mesmo órgãos públicos poderão consultar com o intuito de mostrar os locais que podem apresentar riscos de acidentes. Iremos abordar a importância de ter um alvará em dia, assim como os tipos de alvarás e suas validades.

Os casos mais recentes e de maior repercussão, onde perderam-se muitas vidas e um enorme acervo histórico, foram os acidentes envolvendo a Boate Kiss e o Museu Nacional do Brasil, respectivamente. Ambos funcionavam sem o alvará de funcionamento. Casos como o viaduto da marginal do rio pinheiros que após um estudo da prefeitura, estava na lista de pontes sem risco algum, poderiam ser categorizados pelo algoritmo.

Com o simples pensamento de que nós como usuários ou visitantes de um estabelecimento, não iríamos a um lugar cuja as medidas mínimas para a segurança não foram seguidas, premeditando casos de riscos. A princípio trabalharemos apenas com os estabelecimentos do estado de São Paulo.

Palavras-chave: Alvará, Web Mining, ETL.

ABSTRACT

Given the recent events involving accidents and tragedies in public and private establishments, where a fire caused the great majority of them, we have seen an absence of inspection of one of the basic obligations for any public or private organization. The most recent and most significant cases, where many lives were lost and a huge historical collection, the accidents involving the nightclub Kiss and the National Museum of Brazil, respectively. Both worked without the license of operation.

These cases had major repercussions in the media, but hundreds of other accidents that could have their chances of accidents risk decreased considerably, if they had with the operating license on time. With the simple thought that we as a user or visitor of an establishment would not go to a place whose minimum measures for security was not followed, we developed a simple form of disclosure and consultation of any establishment that needs a license, for anyone on the internet. At first we will work with the establishment of the state of São Paulo.

Keywords: License, Web Mining, ETL.

SUMÁRIO

1. Introdução.....	13
1.1 Objetivo.....	14
1.2. Justificativa	14
1.3 Metodologia	17
2. Fundamentação Teórica	17
2.1 Prévia de Funcionamento.....	18
2.2 Alvará de Funcionamento.....	18
2.3 Tipos de Licenças.....	19
2.4 Variações de Alvará de Funcionamento	22
3. Fundamentação Técnica	25
3.1 CRISP-DM	25
3.2 Extensão PDF	28
3.3 Web Scraping e Web Crawling.....	28
3.4 Banco de dados	29
3.5 Algoritmos Supervisionado e Não Supervisionado	31
4. DESENVOLVIMENTO	36
4.1 Download de PDF.....	36
4.2 Extração do texto	37
4.3 Limpeza de dados	38
4.4 <i>Scraping</i> dos dados.....	41
4.5 Armazenamento.....	42
4.6 Arquitetura.....	43
4.7 Iniciando os testes com os algoritmos	44
5. Validação.....	52
5.1 Técnicas de validação para algoritmos não supervisionados	52
5.2 Testes com algoritmo DBSCAN	53
5.3 Testes com algoritmo KMeans	55
5.4 Considerações finais	62
5. Conclusão.....	64
Referências Bibliográficas.....	66

LISTA DE ILUSTRAÇÕES

Figura 1: Museu do rio de janeiro após o incêndio.....	16
Figura 2: Laudo de licença dos bombeiros	20
Figura 3: Alvará para microempreendedor.....	24
Figura 4: Modelo CRISP-DM.....	26
Figura 5: Diferença entre Web Crawler e Web Scraping.....	29
Figura 6: Tipos de algoritmos	31
Figura 7: Gráfico onde o Kmeans foi utilizado.....	33
Figura 8: Ordem lógica extração de dados	36
Figura 9: Arquitetura	43
Figura 10: Exemplo de visualização de dados.....	45
Figura 11: Exemplo de visualização de dados:.....	46
Figura 12: Exemplo de visualização de dados:.....	47
Figura 13: Exemplo de visualização de dados:.....	47
Figura 14: Exemplo do código.....	48
Figura 15: Exemplo de visualização de dados:.....	49
Figura 16: Exemplo de visualização de dados:.....	50
Figura 17: Exemplo de visualização de dados:.....	51
Figura 18: Exemplo do Código.....	53
Figura 19: Exemplo de visualização de dados.....	54
Figura 20: Exemplo de visualização de dados.....	55
Figura 21: Exemplo do Código.....	56
Figura 22: Exemplo de visualização de dados.....	56
Figura 23: Exemplo do código.....	57
Figura 24: Exemplo do código.....	58
Figura 25: Exemplo de visualização de dados	58
Figura 26: Exemplo de visualização de dados	59
Figura 27: Exemplo de visualização de dados.....	59
Figura 28: Exemplo de visualização de dados.....	60

Figura 29: Exemplo de visualização de dados.....	60
Figura 30: Exemplo de visualização de dados.....	61

LISTA DE SIGLAS

SIGLA	DESCRIÇÃO
ETL	Processo de extrair, transformar e carregar de uma fonte de dados para outra.
SCRIPT	Conjunto de comando e tarefas executadas em sequência
SQL	<i>Structured Query Language</i> - Linguagem padrão de um banco de dados.
TOP-DOWN	Cima para baixo - Abordagem de sistema ou assunto iniciando em somente uma visão ou determinada área.
BOTTOM-UP	Baixo para cima - Abordagem de um assunto começando em uma área específica até chegar a visão geral de todas as áreas.

1. Introdução

Alvará, um documento que parece não ser tão importante, muitas vezes deixado de lado pelas organizações e que acaba ficando irregular, até mesmo o governo deixa de pedir a regularização do mesmo. Entretanto, esse documento é de extrema importância, pois ele é o que irá garantir que o local onde a empresa está instalada é apropriado para o desenvolvimento de suas atividades, e se o mesmo atende as medidas de segurança que garantem a integridade do estabelecimento e das pessoas que ali trabalham ou frequentam. Também fica responsável por verificar se o estabelecimento não irá causar qualquer impacto a aqueles que já estão ali instalados, por exemplo, se uma empresa que trabalhe com transportes ou manuseio de produtos químicos seja instalada em uma área residencial, o risco que ela pode trazer aos que estão ao seu redor é enorme.

Existem diferentes tipos de alvarás que regularizam diferentes tipos de atividades. Esses alvarás devem ser regularizados depois de um certo período de tempo, o documento deve ser emitido pela prefeitura municipal e ser regulamentado e emitido para todos e qualquer estabelecimento. Todos os tipos de alvarás serão descritos mais a frente nesse trabalho. Deixar de cumprir com obrigação de ter o alvará e mantê-lo em dia, pode acarretar em multa e consequentemente o fechamento do estabelecimento.

Frequentemente vemos notícias de estabelecimentos irregulares que na maioria das vezes sofrem algum incêndio, e que logo após foi constatado que o lugar em questão estava com o alvará de funcionamento expirado. Vimos a importância desse documento e utilizando os dados públicos, iremos identificar os estabelecimentos no estado de São Paulo que estão com o alvará expirado e através disso identificar possíveis locais de risco.

O trabalho tem como foco três tipos de público alvo: Empresas Privadas, Órgãos Públicos e o público final onde qualquer cidadão pode verificar se o local é de risco ou não. Todos serão melhor explicados na justificativa do projeto. Espera-se ao término do projeto a criação de uma ferramenta onde será possível ver as categorias de risco de determinado local.

1.1 Objetivo

O principal objetivo deste projeto é criar um modelo que categorize empresas com seus respectivos riscos. Temos como hipótese a possibilidade de classificar esses locais por meio de seus alvarás.

Por meio de um processo de extração de dados, pretendemos conseguir capturar informações pertinentes para a solução e utilizar um algoritmo de *machine learning* para conseguir classificar cada um desses locais.

Utilizando o mesmo em uma aplicação, pretendemos divulgar para que qualquer indivíduo, empresa ou instituição possa utilizar esses dados para melhorar a segurança de pessoas, reduzir impactos ambientais, econômicos e auxiliar a sociedade a obter mais qualidade de vida.

1.2. Justificativa

Observando as notícias (Celular, internet, TV e Rádio) nos deparamos com notícias de acidentes ou tragédias (provocadas ou não), sejam elas em locais públicos ou privados.

Recentemente, um incêndio destruiu o Museu Nacional do Rio de Janeiro, a mais antiga instituição científica do país. Estima-se que 90% dos mais de 20 milhões de itens do acervo foram perdidos. Nas investigações, foi constatado que o alvará de funcionamento estava vencido a anos.

Após pesquisarmos sobre tragédias ou acidentes ocorridos, identificamos uma potencial falha ou até mesmo negligência no seguimento de normas e leis vigentes para o funcionamento de estabelecimentos privados ou públicos em seu local de instalação. Tivemos mais casos além do Museu Nacional, como por exemplo, o incêndio do Museu da Língua Portuguesa na Estação da Luz e um caso mais antigo e com bastante repercussão que foi o incêndio na Boate Kiss, localizado na cidade de Santa Maria no Rio Grande do Sul, onde 242 vidas foram perdidas e 680 pessoas saíram feridas. Ambas com o alvará vencido.

Percebemos que muitas dessas tragédias poderiam ter sido evitadas se tivessem tido a devida atenção, seja ela por parte da fiscalização que não ocorreu ou foi negligente, do proprietário que não se atentou para o prazo das

vistorias e renovação do alvará ou da população que não se preocupou em obter maiores informações sobre o local.

Diante do cenário descrito, percebemos que seria de urgente importância, uma análise preditiva onde locais com alto risco devem ser vistoriados com mais detalhes.

Problemas identificados:

- Iniciativa privadas

Evitar que a empresa tenha suas operações interrompidas devido à falta de alvará. Evitando gastos com multas e acidentes de trabalho.

- Público Final

Recentemente o ministério público promoveu mudanças para evitar incêndios e acidentes com vítimas além das medidas já utilizadas, os responsáveis pela gestão de casas de shows e boates deverão divulgar para os clientes o número e a data de validade do alvará de funcionamento e da proteção contra incêndios. Tudo deverá ser impresso nos ingressos e demais materiais de divulgação das festas.

A ferramenta pretendida será mais um auxílio para a população.

- Órgãos Públicos

Ao invés de vistoriar todos de uma vez, seria priorizado aqueles que estão com maior risco, ou seja, fazer uma triagem na hora de vistoriar de acordo com a classificação.

Dependendo do local em que o acidente vier a ocorrer pode atrapalhar toda uma região, pois um incêndio em um local movimentado que dá acesso a outras regiões faria com que a região fosse interditada e com que os motoristas procurassem outras rotas para chegar ao destino, sobrecarregando outra região, gerando congestionamento, multas e uma população insatisfeita e com raiva. Como foi o caso da ponte que caiu recentemente e não foi vistoriada, causando todo esse transtorno para a cidade, mesmo não sendo um estabelecimento; foi um caso de negligência.

Na figura 1 temos a imagem do museu Nacional do Rio de Janeiro interditado após o incêndio:

Figura 1: Museu do rio de janeiro após o incêndio



Fonte: <https://odia.ig.com.br/rio-de-janeiro/2018/09/5572740-museu-nacional-divulga-lista-de-1-450-milhao-de-itens-nao-atingidos-por-incendio.html>

- **A importância deste projeto.**

Melhorar a visibilidade de riscos. Nossa solução poderá confirmar se o local tem a documentação necessária para garantir a medida de segurança mínima para cada ramo de atividade de uma empresa.

Após o término do trabalho, teremos uma análise que classifica os locais de risco que estão com pelos dados obtidos nos alvarás, como foi visto no capítulo 1.2 Justificativa.

- **Manter o local regular perante as leis**

Como vemos na introdução, aquele que deixar de cumprir com a norma de ter o alvará em dia está sujeito a multa e até mesmo o fechamento do

estabelecimento, esse é um excelente motivo para deixar o mesmo em dia com as normas vigentes do estado.

- **Prevenção de gastos financeiros**

Se o local é seguro, logo estará menos sujeito a incêndios e outros tipos de eventos que possam danificar a empresa e seus bens, não havendo assim, gastos na recuperação. Além de evitar que a empresa seja multada por não ter tomado as medidas necessárias para a segurança e integridade do local.

- **Auxílio ao governo**

Ajudará o governo na tomada de decisão em poder ver os locais que estão com a maior probabilidade de risco e assim solicitar a regularização do mesmo.

1.3 Metodologia

Para o desenvolvimento do trabalho, teremos como base a metodologia de pesquisa descritiva e para a coleta dos dados utilizaremos sites do governo que possuem dados abertos. Nesta fase do projeto, utilizaremos os dados disponibilizados no portal imprensaoficial.com.br, nele encontraremos as informações sobre as expedições de alvarás em São Paulo que será parte essencial do trabalho.

Para a coleta dos dados nos sites mencionados, utilizaremos de técnicas de *web mining*: uma metodologia de recuperação da informação que usa ferramentas de mineração de dados para extrair informações tanto do conteúdo das páginas e de sua estrutura de relacionamentos (links). Após a coleta das informações, usaremos técnicas de *machine learning* para agrupar perfis de alvarás e com isso identificar possíveis riscos.

2. Fundamentação Teórica

Neste capítulo será apresentado todo o conteúdo encontrado relacionado ao funcionamento ditados por normas, decretos e leis vigentes na cidade de São Paulo, porém, em uma linguagem mais comum e simplória para obtenção de um alvará de funcionamento.

2.1 Prévia de Funcionamento

Um dos primeiros passos para todo empreendedor, seja ele de pequeno ou grande porte, é fazer uma consulta prévia do local onde deseja abrir sua empresa. Para verificar e confirmar se um local ou endereço é passível para instalação da atividade comercial de sua empresa ou não, é possível realizar uma consulta junto a Prefeitura ou ao Órgão de Administração Regional, utilizando o portal de transparência da região a ser consultada (em casos de grandes municípios e em outros casos será necessário o comparecimento presencial para realizar a consulta).

Todo município brasileiro possui diretrizes para a construção de determinados tipos de imóveis e abertura de determinadas empresas em cada região da cidade. Estas diretrizes são chamadas de plano diretor ou plano de zoneamento urbano e consiste em realizar uma consulta prévia na região para que os novos estabelecimentos não prejudiquem os que ali já estão instalados. Sem a consulta prévia é impossível conseguir seu alvará de funcionamento no futuro. A Consulta Prévia de Funcionamento para o imóvel onde se pretende instalar a atividade, poderá ser efetuada através do Endereço do logradouro e respectivo número, ou número do Código de Endereçamento Postal - CEP do logradouro, em seguida o número do imóvel, ou ainda, número do Setor, Quadra e Lote - SQL constante do IPTU do imóvel.

O que é um Plano Diretor Estratégico?

Seria um plano que, a partir de um diagnóstico científico da realidade física, social, econômica, política e administrativa da cidade, do município e de sua região, apresentaria um conjunto de propostas para o futuro desenvolvimento socioeconômico e futura organização espacial dos usos do solo urbano, das redes de infraestrutura e de elementos fundamentais da estrutura urbana, para a cidade e para o município. Propostas estas definidas para curto, médio e longo prazo, e aprovadas por lei municipal. (Villaça, 1999, p.238).

2.2 Alvará de Funcionamento

Existem diferentes tipos de documentação que são necessárias para a atuação e prestação de serviços oferecidos por uma empresa em determinado local, que variam de acordo com o tipo de atividade que será fornecido. Neste projeto será abordado e descrito com maior profundidade, somente o documento

de total obrigatoriedade em qualquer atividade que uma empresa poderá exercer, o Alvará de Funcionamento. Mas a solicitação de um Alvará de Funcionamento, poderá depender de Licenças de Funcionamento previamente solicitadas e aprovadas ao seu respectivo órgão de regulamentação.


2.3 Tipos de Licenças

Atualmente existem 5 tipos de licenças, tendo como indispensável a licença AVCB.

- **Auto de Vistoria do Corpo de Bombeiros (AVCB)**

Realizado pelo Corpo de Bombeiros serve para verificar as normas e condições mínimas de segurança do local. Todo o tipo de empresas deve obter essa licença. De acordo com a legislação de segurança contra incêndio e pânico, a validade do Auto de Vistoria do Corpo de Bombeiros (AVCB) é de 05 (cinco) anos, com exceção das edificações de recepção de público, cujo AVCB possui validade de 03 (três) anos.

Figura 2: Laudo de licença dos bombeiros



AVCB

AUTO DE VISTORIA DO CORPO DE BOMBEIROS

WWW.BOMBEIROS.MG.GOV.BR

Nº: 201700 **VALIDADE: 07/02/2023**

O Corpo de Bombeiros Militar de Minas Gerais certifica que a edificação, ou área de risco, abaixo descrita, possui as medidas de segurança previstas na legislação estadual* de Segurança Contra Incêndio e Pânico vigente, considerando as informações no respectivo Processo de Segurança Contra Incêndio e Pânico (PSCIP).

Endereço: _____

Bairro: NOVO ALVORADA **Município:** Sabará

Ocupação: D-4 **Público:** *****

Proprietário: _____

Responsável pelo Uso: _____

Responsável Técnico: _____

Área Total: 1151.79

Área Liberada: 1151.79

Emitido em: 07/02/2018

Última Atualização: 07/02/2018 15:49:56

*Lei 14.130/2001 - Dispõe sobre prevenção Contra Incêndio e Pânico do Estado de Minas Gerais e dá outras providências.

CORPO DE BOMBEIROS MILITAR DE MINAS GERAIS

Fonte: <https://insertprojetos.com/auto-de-vistoria-do-corpo-de-bombeiros-avcb-belo-horizonte.html>

- **Licença Ambiental**

O órgão expedidor são órgãos municipais e estaduais do meio ambiente e Ibama. Essa licença deve ser obtida por empresas como: indústria metalúrgica, mecânica, indústrias de material de transporte, têxtil, vestuário, calçados, madeira, química, produtos alimentares, obras civis, turismo, atividades agropecuárias, entre outras;

Licença Prévia (LP) – validade, no mínimo, do prazo estabelecido pelo cronograma de elaboração dos planos, programas e projetos relativos ao empreendimento ou atividade, não podendo ser superior a 5(cinco) anos;

Licença de Instalação (LI) – validade, no mínimo, do prazo estabelecido pelo cronograma de instalação do empreendimento ou atividade, não podendo ser superior a 6 (seis) anos;

Licença de Operação (LO) – deverá considerar os planos de controle ambiental e será de, no mínimo, 4 (quatro) anos, e no máximo, 10 (dez) anos.

- **Licença Sanitária**

Envolve órgãos municipais, estaduais e federais de vigilância sanitária (ANVISA). Deve ser obtido por empresas que atuam em setores de alimentos, cosméticos, produtos para higiene e perfumes, medicamentos e/ou insumos farmacêuticos, saneantes e produtos para saúde;

A Renovação da Autorização de Funcionamento será anual.

A inobservância da Resolução configura infração sujeita às penalidades previstas na Lei 6.437/77, que prevê multas de R\$ 1.500,00 a R\$ 2.000.0000,00 sem prejuízo de outras sanções de natureza civil ou penal cabíveis.

- **Produtos de origem animal**

A licença deve ser obtida pelo Ministério da Agricultura, Pecuária e Abastecimento. Qualquer empresa que produz produtos de origem animal, principalmente para consumo humano e comercializados no mercado interno ou externo, tem a obrigatoriedade de adquirir essa licença.

A renovação de registro concedida pelo Mapa autoriza o estabelecimento fabricante, fracionador ou importador a continuar

desenvolvendo suas atividades, tendo em vista o prazo de validade de cinco anos do registro, conforme determinado no artigo 6º do Decreto nº 6.296, de 12 de dezembro de 2007.

- **Registro de produtos químicos controlados**

Polícia Federal. Atividades relacionadas por produtos controlados (a lista é composta por 146 produtos).

A renovação do Certificado de Licença de Funcionamento deve ser realizada anualmente de acordo com o que determina o artigo 5º da Lei Federal 10.357/01.

2.4 Variações de Alvará de Funcionamento

Existem quatro variações de alvará de funcionamento. São elas:

- **Auto de Licença de Funcionamento (ALF)**

Vale para imóveis não residenciais, com instalação de atividades comerciais, industriais ou de serviços;

O documento terá validade de 1 ano após a data de sua expedição.

- **Auto de Licença de Funcionamento Condicionado (ALF-C)**

Em casos de edificações ainda irregulares ou inscritas no Cadastro Informativo Municipal. Vale para imóveis com área total construída de 1500 m2 a 5000 m2;

O documento terá validade de 2 anos.

- **Alvará de Funcionamento de Local de Reunião:**

Solicitado para todos os locais que tenham reunião de público. Exemplos: bares, restaurantes, cinemas e similares, que tenham capacidade de público igual ou superior a 250 pessoas;

É o documento que regulariza o funcionamento das edificações com lotação superior a 250 (duzentos e cinquenta) pessoas, tais como: clubes, ginásios, estádios, auditórios, teatros, casas noturnas, boates, salões de bailes ou danças, templos religiosos, restaurantes e similares. Sua validação visa

assegurar a validade do Alvará de Funcionamento de Local de Reunião por mais um único período igual de 12 (doze) meses.

- **Alvará de Autorização para Eventos Públicos e Temporários**

Para locais onde haja eventos públicos e temporários para mais de 250 pessoas, seja em imóveis públicos ou privados.

O Alvará de Autorização para Eventos Temporários é a licença para funcionamento de atividades geradoras de público para um determinado local, com duração limitada de no máximo 6 (seis) meses.

- **Alvará para MEI**

Aos que se enquadram como Microempreendedor, um alvará provisório é emitido logo após o término do cadastro para MEI no Portal do Empreendedor. Esse alvará deve estar regularizado em até 180 dias após a formalização. Caso passado o prazo estipulado e não houver manifestação de algum órgão da prefeitura questionando a validade do alvará ou apontando alguma irregularidade, o documento se torna definitivo.

Esse documento pode ser considerado (de acordo com natureza desenvolvida pela empresa) como provisório ou definitivo o que dependerá da legislação do município competente.

Em geral, a licença provisória tem prazo de validade máximo de até 180 dias. Já a definitiva não tem prazo determinado desde que as características da atividade e da edificação permaneçam inalteradas.

Figura 3: Alvará para microempreendedor

28/5/2014 Certificado da Condição de Microempreendedor Individual - Impressão

Certificado da Condição de Microempreendedor Individual

Identificação

Nome Empresarial
ELIZABETH APARECIDA BEZERRA 31117742873

Nome do Empresário
ELIZABETH APARECIDA BEZERRA

Nome Fantasia
EDITORA BEZZ

Capital Social
100,00

Nº da Identidade **Órgão Emissor** **UF Emissor** **CPF**
4140033954 SSP SP 311.177.428-73

Condição de Microempreendedor Individual

Situação Cadastral Vigente **Data de Início da Situação Cadastral Vigente**
ATIVO 12/05/2014

Números de Registro

CNPJ **NIRE**
20.233.962/0001-06 35-8-1115906-9

Endereço Comercial

CEP **Logradouro** **Número**
07417-160 RUA MILTON FERREIRA FRANCO (PRQ R BARRETO) 101

Bairro
PIRITUBA

Município **UF**
ARUJA SP

Atividades

Data de Início de Atividades
12/05/2014

Código da Atividade Principal **Descrição da Atividade Principal**
58.11-5/00 Edição de livros

Termo de Ciência e Responsabilidade com Efeito de Alvará de Licença e Funcionamento Provisório - declaração prestada no momento da inscrição:

Declaro, sob as penas da Lei, que conheço e atendo os requisitos legais exigidos pelo Estado e pela Prefeitura do Município para emissão do Alvará de Licença e Funcionamento, compreendidos os aspectos sanitários, ambientais, tributários, de segurança pública, uso e ocupação do solo, atividades domiciliares e restrições ao uso de espaços públicos. O não-atendimento a esses requisitos acarretará o cancelamento deste Alvará de Licença e Funcionamento Provisório.

Este Certificado comprova as inscrições, alvará, licenças e a situação de enquadramento do empresário na condição de Microempreendedor Individual. A sua aceitação está condicionada à verificação de sua autenticidade na internet, no endereço: <http://www.portaldoempreendedor.gov.br/>

Certificado emitido com base na Resolução nº 16, de 17 de dezembro de 2009, do Comitê para Gestão da Rede Nacional para a Simplificação do Registro e da Legalização de Empresas e Negócios - REDESIM.

ATENÇÃO: qualquer rasura ou emenda invalidará este documento.

Para pesquisar a inscrição estadual e/ou municipal (quando convenientes do cadastro sincronizado nacional), informe os elementos abaixo no endereço eletrônico <http://www.receita.fazenda.gov.br/PessoaJuridica/CNPJ/fcpj/consulta.asp>

Número do Recibo: ME69642016

1/2

Fonte: <https://pt.slideshare.net/EffyLizzyBeth/certificado-da-condicao-de-microempreendedor-individual-impresso>

A seguir uma tabela com os alvarás e o prazo de validade de cada uma para uma melhor visualização:

Tabela 1: Prazo de Validade dos Alvarás

TIPO DE LICENÇA	VALIDADE
AUTO DE VISTORIA DO CORPO DE BOMBEIROS(AVCB)	3 anos
LICENÇA AMBIENTAL	
LICENÇA PRÉVIA (LP)	5 anos
LICENÇA DE INSTALAÇÃO (LI)	6 anos
LICENÇA DE OPERAÇÃO (LO)	4 anos
LICENÇA SANITÁRIA	1 ano
PRODUTOS DE ORIGEM ANIMAL	5 anos
REGISTRO DE PRODUTOS QUÍMICOS CONTROLADOS	1 ano
AUTO DE LICENÇA DE FUNCIONAMENTO ALF	1 ano
AUTO DE LICENÇA DE FUNCIONAMENTO CONDICIONADO ALF-C	2 anos
FUNCIONAMENTO DE LOCAL DE REUNIÃO	1 ano
AUTORIZAÇÃO PARA EVENTOS PÚBLICOS E TEMPORÁRIOS	6 meses
Alvará para MEI: a licença definitiva não tem prazo definido desde que as características da edificação e da atividade permaneçam inalteradas, a provisória tem o prazo de 180 dias.	

Fonte: Desenvolvida pelos autores.

3. Fundamentação Técnica

Para transformar dados em informações, alguns processos precisam ser realizados, para conseguirmos enxergar essas informações de maneira eficiente, utilizaremos um conjunto de técnicas, metodologias e ferramentas. Em resumo, nossa solução consiste em ler um conjunto de arquivos com a extensão PDF, capturar os dados necessários, incluir esses dados em um banco de dados, visualizar e analisar os dados capturados. Iremos explorar cada um dos pontos citados da forma correta.

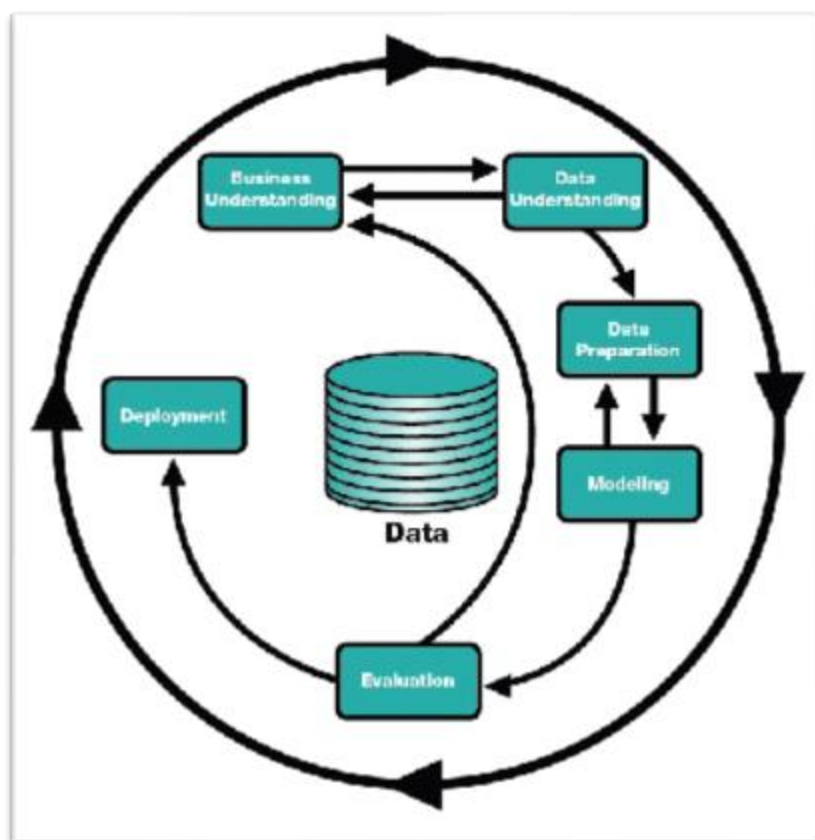
3.1 CRISP-DM

Decidimos adotar uma das metodologias mais conhecidas e disseminadas descritas por Chapman (2000) para o desenvolvimento do projeto,

devido a sua facilidade de aplicação e por mais se adequar as necessidades e desafios para obtermos um melhor resultado neste projeto.

Em 1996 um consórcio de empresas fornecedoras e potenciais consumidores de Data Mining criaram uma metodologia não proprietária, chamada CRISP-DM (*Cross-Industry Standard Process for Data Mining*). A metodologia é composta por seis fases interdependentes, conforme vemos na figura abaixo:

Figura 4: Modelo CRISP-DM



Fonte: <https://paulovasconcellos.com.br/>

1 - Entendimento do negócio: Esta fase busca a compreensão dos objetivos do projeto e suas necessidades do ponto de vista dos negócios, de forma a transformar esse conhecimento em uma definição de problema e um plano preliminar. Possui como tarefas: determinar os objetivos do negócio,

avaliação da situação, determinação dos objetivos da mineração de dados e a produção de um plano de projeto.

2 - Entendimento dos Dados: É a fase de coleta dos dados e de investigação preliminar, destinada a uma familiarização maior com os mesmos e avaliação da qualidade dos dados. Em consequência dessas atividades, é comum a descoberta de padrões interessantes já nessa fase. Possui como tarefas: coletar dados iniciais, descrever os dados, explorar os dados e verificar a qualidade dos dados.

3 - Preparação dos Dados: Fase da construção final do banco de dados que será submetido à ferramenta de análise. Este banco, retirado dos dados originais, passará por um processo de limpeza e transformação. Possui como produtos principais de saída o arquivo para mineração e sua descrição. Possui como tarefas: selecionar dados, limpeza dos dados, construção dos dados, integrar dados e formatar dados.

4 – Modelagem: Nesta fase é escolhida a técnica de modelagem dos dados. Dependendo da técnica utilizada, pode ser necessário o retorno à fase de preparação dos dados para o ajuste às exigências de determinadas técnicas. Possui como tarefas: selecionar a técnica de modelagem, gerar teste de projeto (*test design*), construir e validar o modelo.

5 – Avaliação: A avaliação é realizada visando identificar alguma questão do negócio que não foi explorada suficientemente. Este passo avalia o grau com o qual o modelo produzido alcançou os objetivos. Possui como tarefas: análise de resultados, revisão dos processos e determinação dos próximos passos.

6 – Distribuição: Fase em que o modelo é distribuído aos clientes de forma que esse possa extrair conhecimento. Pode ser um processo único, com resultado único, ou ainda um processo iterativo. Possui como tarefas: elaboração de plano de distribuição, plano de monitoramento e manutenção, elaboração do relatório final e revisão do projeto.

3.2 Extensão PDF

Todo arquivo virtual tem uma extensão, é um sufixo que informa o seu formato e informa o padrão que o arquivo foi compilado para que os programas saibam como ler, executar ou gravar. Sua sintaxe é descrita como:

[nome_do_arquivo].[extensão]

Ex: foto.jpg, foto_sem_fundo.png, documento_word.doc, planilha_excel.xls, arquivo_de_sistema.dll, jogo.exe, etc.

Em nossa solução, utilizaremos os arquivos com a extensão .pdf disponibilizados pelos portais do governo.

Os arquivos com extensão .pdf são a abreviatura de *Portable Document Format*, ele foi criado pela empresa Adobe, em 1991. O cofundador da Adobe, o Dr. John Warnock lançou uma revolução digital que foi chamado de “*The Camelot Project*”, com a finalidade de usar e compartilhar arquivos de maneira compatível, independentemente do software, hardware ou sistema operacional, mantendo sua aparência original, independente se for texto ou imagem.

O PDF se tornou um padrão de documento aberto e é mantido pelo *International Organization for Standardization* (ISO). Eles podem conter links, botões, campos de formulário, áudio, vídeo e lógica de negócios, eles podem ser assinados eletronicamente, atendendo aos padrões ISO 32000.

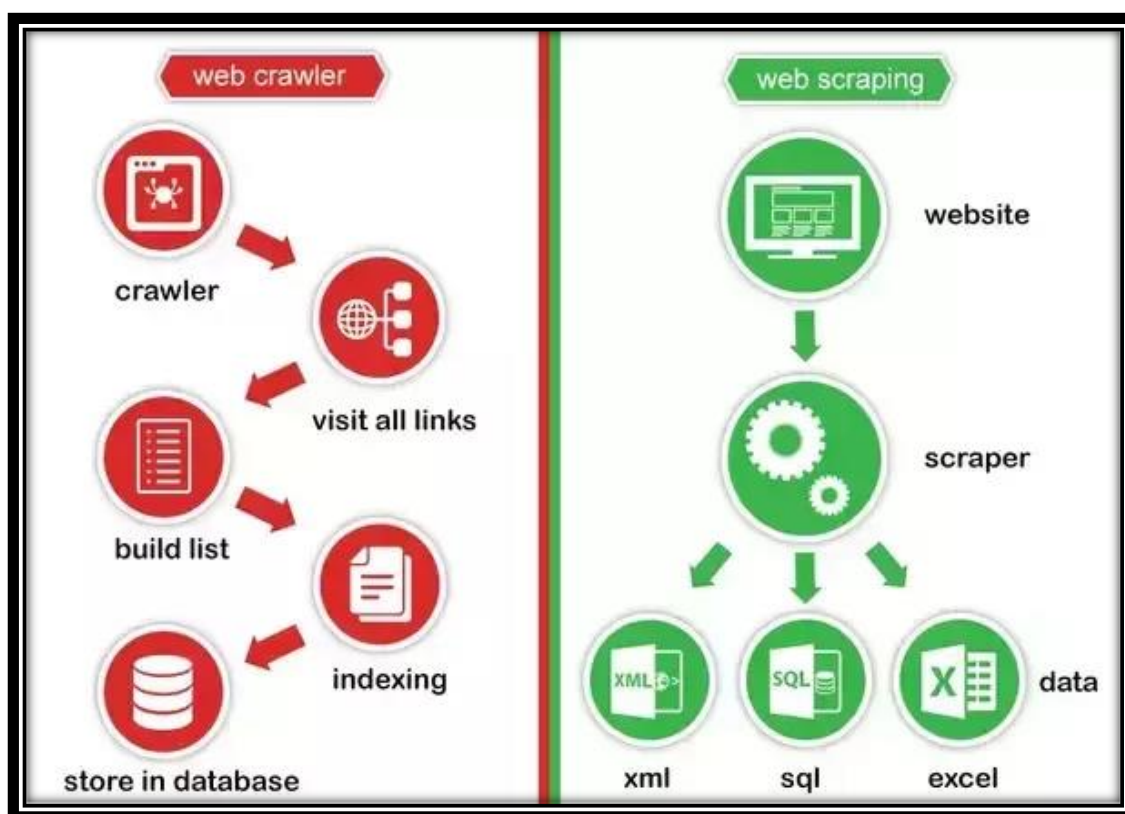
3.3 Web Scraping e Web Crawling

A maior fonte de informação do mundo é a internet, as páginas da web carregam milhares de informações e quando conseguimos acesso a essa informação temos em mãos dados importantíssimos para a tomada de decisão. Para conseguirmos automatizar o processo de captura na web é necessário utilizar as técnicas de *Web Scraping* e *Web Crawling*.

Web Crawling, segundo Mitchell (2015), é o processo de localizar e capturar informações da internet, rastreando páginas web e realizando o download de todo o conteúdo. Essa técnica é muito utilizada pelos mecanismos de busca do Google, Bing e agências governamentais, eles leem o código fonte da página inteira.

O *Web Scraping*, segundo Mitchell (2015), é um processo parecido, porém, ele é muito mais específico na busca de dados. Ao invés de capturar todo o código fonte, capturamos apenas os trechos que desejamos. Empresas de marketing e geradores de leads capturam apenas o que necessitam, por exemplo, capturar informações de produtos específicos no mercado livre.

Figura 5: Diferença entre Web Crawler e Web Scraping



Fonte: <http://proweb scraping.com/>

3.4 Banco de dados

Quando realizamos o processo de carga no banco de dados, precisamos de um lugar para armazenar os dados transformados no processo de ETL. Para essa finalidade, existem os bancos de dados que são sistemas criados e evoluídos para essa função, praticamente todos os sistemas utilizam banco de dados para armazenamento de dados.

Antigamente as empresas armazenavam seus dados em papéis que eram armazenados em pastas, arquivos, gavetas, etc. Para ter acesso a alguma informação, por mais simples que fosse, era uma tarefa muito custosa. Com o

tempo, esse processo se tornou totalmente digital. As primeiras versões de banco de dados eram softwares que realizavam manipulações de arquivos digitais, eles realizavam consultas, inserções, exclusões e alterações, facilitando muito essa tarefa na época.

Com o tempo, a necessidade dos dados se relacionarem se tornaram muito mais complexas, pois, pelo uso dos softwares, aumentou a quantidade de dados, então esses dados precisavam conversar entre si. Perguntas como qual produto veio de qual fornecedor ou qual cliente solicitou qual serviço ficaram cada vez mais difíceis.

Na década de 60, a empresa IBM investiu fortemente na solução desse problema de relacionamento, eles criam diversos modelos, como o modelo hierárquico e o modelo de rede. Em junho de 1970, Edgar Frank "Ted" Codd pesquisador da IBM apresentou o modelo relacional em um artigo chamado de "*A Relational Model of Data for Large Shared Data Banks*", onde descrevia como pessoas sem muito aprofundamento técnico conseguiria relacionar os dados, a ideia foi adiante, e foi criado o modelo relacional que é usado até hoje. A primeira versão foi o "Sistema R" que utilizava uma linguagem única chamada SQL (*Structured Query Language*) a qual se tornou a linguagem padrão para bancos de dados relacionais. Comercialmente, o "Sistema R" não funcionou, porém, os outros sistemas de banco de dados foram baseados nele.

Na década de 80 surgiram novos bancos de dados, como o Oracle 2 da empresa Oracle e o SQL/DS que anos mais tarde se tornou DB2, da empresa IBM. Após algum tempo surgiram outros bancos de dados como o SQL Server (MySQL, DBase III, Paradox, entre outros). Hoje existem diversos bancos de dados relacionais e não relacionais que desconstróem toda a visão de entidades relacionais e são orientados à documentos.

3.5 Algoritmos Supervisionado e Não Supervisionado

Figura 6: Tipos de algoritmos



Fonte: <https://king.host/blog/2018/03/machine-learning-uma-introducao/>

Quando falamos em algoritmos de *machine learning*, onde o intuito é utilizar algoritmos matemáticos para que a máquina encontre padrões e com isso nos auxiliando em soluções robustas. O ser humano, não consegue por si só, analisar, entender e encontrar padrões quando existe muitos dados e uma infinidade de variáveis. Quando utilizamos algoritmos de *machine learning* é possível resolver esse problema de forma eficiente e eficaz

Nesse universo, encontra-se diversos algoritmos como K-Mean, DbSCAN, regressão linear, regressão logística, árvore de decisão, entre outros. Estão separados em 3 grandes grupos, o supervisionado, não supervisionado e por reforço.

Quando dizemos que o algoritmo é supervisionado, ele funciona com dados rotulados, isso quer dizer que existe um alvo, onde, em uma base de

dados, existem todas as variáveis que vão ser utilizadas para a criação do modelo de dados e possuir também uma variável que é o resultado esperado. Os algoritmos que estão nesse grupo, aprendem padrões que irão gerar o resultado esperado. Então, ao aplicar o modelo em uma base real sem ter o resultado, o algoritmo irá utilizar o padrão aprendido com as variáveis, e apresentará o resultado.

É muito utilizado em análise preditiva, onde já existem dados com os resultados, e é aplicado para saber quando um mesmo padrão acontecer, e qual será o resultado. Algoritmos de regressão, por exemplo, podem identificar padrões químicos de diversos vinhos e entender porque um *sommelier* deu uma nota x, assim, quando um novo vinho é adicionado a base, com sua respectiva composição química, e passa pelo modelo gerado do algoritmo, ele consegue prever qual a nota que um *sommelier* poderia dar.

No grupo dos não supervisionados, não existe um resultado esperado. Podemos simplificar o método de funcionamento desse grupo como: “existem dados, ache um padrão”. Quando existem uma série de dados e sabemos o que queremos identificar, porém não sabemos como procurar um padrão, utilizamos algoritmos.

Então, ao usar algoritmos como K-means e o DbSCAN, será processado e gerado um modelo que funciona por meio de clusters, onde existirá agrupamentos de dados, indicando que em cada grupo existe algo encontrado que torna aqueles dados semelhantes.

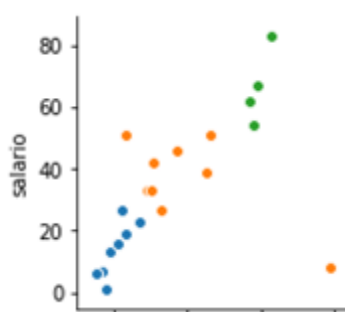
Algoritmos não supervisionados, são muito utilizados em sistemas de recomendações de sites, onde não existe uma variável alvo. Por exemplo: cada hora nós como usuários, estamos procurando diversos conteúdos, uma hora podemos procurar sobre ferramentas, cursos, literatura ou vídeos, não existe muito bem um padrão. Em nosso caso, podemos procurar sobre ferramentas de informática, cadeiras de escritório, notebooks, livros sobre Python, vídeos tutoriais de Javascript, cursos de faculdade de tecnologia. Se esses dados forem analisados para identificar um grupo de usuário, é possível que um algoritmo identifique que sejamos do grupo X, o mesmo não se pode fazer de alguém que faça pesquisas relacionadas a design ou marketing, nesse caso, estariam no grupo Y, e assim por diante. Após identificar esses grupos, devida as pesquisas

de diversos usuários naquele grupo específico, é feita recomendações de produtos e serviços que pessoas daquele mesmo grupo fizeram.

3.5.1 Algoritmo K-Means

O algoritmo não supervisionado Kmeans, segundo FORGY (1965), analisa todos os dados presentes no *dataset* e cria classificações, através dela ele inicia um cluster e define quais pontos pertencem aquela classe. No KMeans, o número de clusters deve ser informado através do parâmetro *n_cluster*, esse número de clusters que temos que definir é chamado de *k*, daí o *k* de K-means. Para gerar os clusters, o algoritmo compara os valores de cada linha por meio da distância euclidiana, outros tipos de distância podem ser utilizados. Após o cálculo da distância, o algoritmo calcula centroides para cada um dos clusters gerados. O processo iterativo do KMeans faz o mesmo processo repetindo até a convergência.

Figura 7: Gráfico onde o Kmeans foi utilizado



Fonte: Elaborado pelos autores

Vantagens

- Ser relativamente escalável e eficiente para grandes conjuntos de dados;
- É sensível a ruídos, visto que pequeno número de tais dados pode influenciar, substancialmente, o valor médio das coordenadas;
- O método frequentemente termina num local ótimo.

Desvantagens

- Esse método só pode ser aplicado quando a média (centroide) de um cluster pode ser definido. Isto pode não ser o caso em algumas aplicações que utilizam dados com atributos categóricos (nominais) envolvidos.

A abordagem por k-means é sensível a partição inicial, gerada pela escolha aleatória dos centroides. A técnica k-means necessita que o número k de clusters seja informado com antecedência.

O método k-means não é adequado para descobrir clusters com formas não convexas, pelo fato desse método gerar clusters com figuras circulares esse problema é conhecido como problema da superposição de classes. Esse fenômeno ocorre quando a base de dados analisada possui muitos pontos de confusão.

3.5.2 Algoritmo DBSCAN

O algoritmo do tipo não supervisionado DBSCAN (*Density Based Spatial Clustering of Application with Noise*) é um método de clusterização não paramétrico baseado em densidade, proposto por ESTER et al (1996) e funciona da seguinte forma: verifica a vizinhança de cada ponto para encontrar os clusters, partindo de um ponto arbitrário p. Se p for um ponto central, um novo clusters com p com um centro é criado. Quando p for um ponto de fronteira, e nenhum outro ponto poder ser alcançado por densidade a partir de p o processo se inicia novamente e o DBSCAN parte para o próximo ponto da base.

Ele coleta os objetos que são alcançáveis por densidade direto de pontos centrais cujo podem vir a envolver alguns clusters alcançáveis por densidade. Quando nenhum novo ponto poder ser adicionado a qualquer cluster o processo termina.

Todos os pontos com distância menor ou igual são colocados no mesmo cluster. Pontos de borda perto de um ponto central é colocado no mesmo cluster do ponto central e os pontos que não são alcançáveis por algum ponto central são classificados como ruído.

Resumindo, o algoritmo escolhe alguns pontos aleatórios. Usando o parâmetro da distância ele traça um raio e verifica todos os elementos que estão dentro desse raio.

Requer apenas dois parâmetros de entrada:

eps: float, opcional.

A distância máxima entre duas amostras para serem consideradas como na mesma vizinhança:

min_samples: int, opcional.

O número de amostras (ou peso total) em uma vizinhança para que um ponto seja considerado como um ponto central. Isso inclui o próprio ponto.

Vantagens:

- Diferente do KMeans, você não precisa informar o número de clusters;
- Pode encontrar clusters com formas geométricas arbitrárias;
- O DBSCAN requer apenas dois parâmetros e não é suscetível a ordem em que os pontos estão localizados no banco de dados (no entanto, os pontos de borda de dois clusters diferentes podem pertencer a um cluster ou outro se a ordem dos pontos no banco de dados for alterada e a atribuição de cluster for exclusiva apenas em isomorfismos);
- Encontrar padrões não lineares;
- Robusto contra *outliers*;
- O resultado pode ser mais consistente que o k-means porque a inicialização dos centroides não afeta tanto o algoritmo.

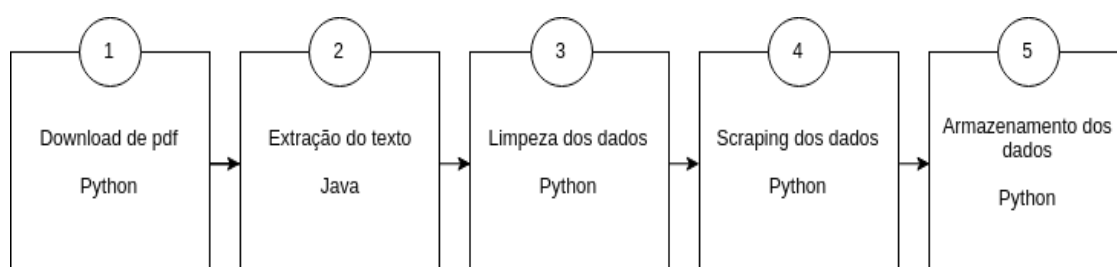
Desvantagens

- Dependendo da inicialização, um ponto pode pertencer ao cluster dependendo da ordem;
- Pode ser difícil encontrar um bom valor para o parâmetro da distância.

4. DESENVOLVIMENTO

Após todo o entendimento do negócio, foi então realizado o processo de desenvolvimento que constitui em obter, limpar e persistir os dados. Para realizarmos nossa análise, necessitamos dos dados, que estão em sua maioria em PDFs do Diário Oficial em formato não estruturado. O seguinte diagrama expressa de forma lógica como funciona desde o download do arquivo bruto até o armazenamento da informação em um banco de dados.

Figura 8: Ordem lógica extração de dados



Fonte: Elaborado pelos autores.

No diagrama (Figura 11), está desenhada a ordem lógica de extração dos dados que são:

4.1 Download de PDF

Na primeira etapa foi verificado como os arquivos são armazenados no site da imprensa oficial do estado de São Paulo. Após o entendimento do padrão, foi constatado que cada página do PDF é uma URL diferente, contendo o caderno do diário oficial e a data de sua publicação:

O Download dos PDFs são realizados por meio de um loop, onde existe uma URL no formato:

```
http://diariooficial.imprensaoficial.com.br/doflash/prototipo/<YYYY>/<MesString>/<dd>/<caderno>/pdf/pg_<0001>.pdf
```

A cada iteração do loop, a URL é modificada e acessada via Script Python, o mesmo realiza o download e envia para uma pasta chamada Download, onde, para facilitar a organização, existem subpastas referentes a data de criação do PDF no formato:

/Download/<YYYY>/<MM>/<dd>/<caderno>/pg_<0001>.pdf

Para iniciar todo o processo de download, é definida uma data inicial onde é utilizado como parâmetro para o loop, executando assim o download e a criação do diretório.

Não existe um padrão de páginas para cada caderno, então a lógica utilizada para saber se já chegou na última página, antes do download e criação do diretório, é realizada uma checagem para saber se não é detectada uma *exception* de *PageNotFound*. Caso isso aconteça, o loop já passou por todas as páginas e então ele para de realizar o download e avança para o próximo dia, reiniciando todo o loop.

Após finalizar os downloads, é iniciada a etapa de extração do texto.

4.2 Extração do texto

Após todos os downloads concluídos, realizamos a extração de cada PDF e para isso utilizamos a linguagem Java com a biblioteca Apache PDFBox. Essa biblioteca é muito completa e muito utilizada para manipulação de arquivos PDFs.

A biblioteca Apache PDFBox tem a capacidade de converter arquivos PDF em textos, sem a necessidade de utilizar OCR, fazendo com que seja rápida a conversão.

Os arquivos de PDFs do diário oficial tem uma particularidade, os textos são escritos em colunas (como em um artigo de jornal) e o maior problema enfrentado é como identificar onde começa e termina o texto em cada coluna, pois se apenas extrairmos o texto, não é possível identificar qual coluna está cada texto, onde ele se inicia e onde ele termina.

O PDFBox permite criar áreas de corte com o método *stripperArea*. Ele funciona com retângulos, ou seja, manualmente deve-se definir um objeto do tipo retângulo, contendo a largura e altura em pixels. Com isso, conseguimos definir o tamanho de cada coluna. Quando utilizamos esse tamanho dentro do método *stripperArea*, informamos também qual a posição inicial do retângulo em pixels. Após definir o tamanho de cada coluna e onde a mesma se inicia, executamos o método e adicionamos cada linha lida em um objeto do tipo *List<String>*.

Utilizando um loop, realizamos esse processo página a página, por caderno e dia. Ao loop ser encerrado, temos uma lista contendo cada linha de todos os PDFs gerados naquele dia e naquele caderno. Com essa lista, geramos um arquivo texto em um diretório separado com o seguinte padrão:

```
/EXTRAIDO/<YYYY>_<Mes-String-Capitalizado>_<DD>_<caderno>.txt
```

4.3 Limpeza de dados

O próximo passo é extrair os dados dos arquivos txt e para isso temos que pré processá-los em memória utilizando um loop. Processamos arquivo por arquivo utilizando algumas premissas, como o PDF é dividido em colunas, é comum uma palavra estar separada por hífen, para iniciar em uma linha e terminar em outra. Ao ler linha a linha, utilizamos o objeto *List<String>* para identificar se o último caractere é igual a “-”, se isso for verdade, removemos o carácter “-”, copiamos a linha abaixo e concatenamos na linha com o “-” removido, e por fim removemos a linha copiada.

O objetivo principal do projeto é detectar possíveis locais perigosos, e como base nisso, utilizamos o Alvará AVCB do corpo de bombeiros. Percorrendo linha a linha, procuramos por palavras chaves como “GRUPAMENTO DE BOMBEIROS” ou “BATALHÃO DA POLÍCIA MILITAR”, identificamos também linhas que contem “1.1 - PROJETO”, “1.2 - ENDEREÇO”, etc. Criamos um objeto chamado AVCB e geramos uma lista com todos os objetos encontrados, porém, quando fomos verificar, em 6 meses, apenas 90 casos foram achados, estranhamos a quantidade e ao verificar, descobrimos que nem todos os processos de vistoria dos bombeiros é realizado pelos bombeiros ou pela polícia militar. Por motivos de desburocratização, o governo liberou um sistema para engenheiros contratarem bombeiros particulares e realizar as vistorias, não passando pela solicitação do município, e não aparecendo no diário oficial.

Com essa nova informação em mãos, tivemos que entender outros modos de como podemos identificar possíveis riscos. Ao analisar diversos casos de problemas relacionados com alvarás, visualizamos que os problemas tendem a ocorrer com Alvarás de licenças condicionadas. Esses alvarás não têm toda

documentação completa e existe um prazo de 2 anos para ajustar o local e realizar novamente as vistorias. Partindo desse princípio, mudamos nosso foco para localizar alvarás que podem ser condicionados.

Dentro do arquivo PDF não existe um padrão muito bem definido, pois eles são na maioria das vezes escritos à mão por algum sistema ou planilha e depois importado para um programa que gera o PDF do diário oficial. Então, para capturar os dados necessários, utilizamos primeiramente a palavra-chave “SQL/INCRA”, a maioria das linhas que contém essa chave informa o início dos dados que gostaríamos de buscar.

A partir dessa linha verificamos se entre as primeiras 3 linhas existe a palavra “deferido” ou “indeferido” e alguma palavra chave como “ALVARÁ” ou “RECONSIDERAÇÃO” ou “APOSTILAMENTO”, se existir, sabemos que detectamos a informação que procuramos. Para facilitar o processo, lemos todas as linhas, e cada vez que conseguimos o padrão informado, adicionamos o index desse padrão dentro de uma *list*.

Após esse pré-processamento, percorremos a lista e verificamos a distância de um index até o outro, se a distância do index que contém “SQL/INCRA” for maior que 2, sabemos então que o tipo de alvará está separado em mais de uma linha. Então realizamos um processo parecido de copiar a linha abaixo para a de cima e excluimos a linha copiada para que a distância da chave “SQL/INCRA” se de no máximo 2. Com esse processo, já temos várias informações à disposição.

A última parte dessa extração é conseguir extrair a descrição do motivo deferido ou indeferido. O problema dessa informação é que com o index das palavras chaves “DEFERIDO” e “INDEFERIDO” sabemos onde se inicia a descrição e onde ela termina. Se houver outro “SQL/INCRA”, caso contrário, fica muito difícil saber onde a descrição termina. Para tentar prever esse padrão, identificamos que na maioria dos casos, quando a descrição termina, o último caractere é sempre o “.”, porém, a descrição pode conter diversas linhas utilizando a mesma lógica de copiar, colar e excluir, realizamos esse pré-processamento.

No fim, o esperado é ter 4 linhas para cada conjunto que buscamos. Criamos então um objeto chamado “Alvará”, e nele contém cada variável em formato de *String* que queremos extrair.

Na primeira linha, onde existe a palavras chave “SQL/INCRA” o mesmo existe o código do sql/Incra, o dígito do mesmo, o solicitante que pode ser uma pessoa ou um nome fantasia, e o número do processo.

Separamos então com diversos métodos de *substrings* e *splits* cada uma dessas informações e alocamos nas suas respectivas variáveis do objeto “Alvara”. Próxima etapa é a segunda linha onde está o tipo de alvará, nela alocamos a linha inteira em sua respectiva variável dentro do objeto “Alvara”. Na terceira linha temos o status, que pode ser deferido ou indeferido, removemos o carácter “:” e adicionamos a linha dentro de uma lista. Por fim, a última linha é a de descrição, que é adicionado dentro da lista.

Para melhorar o processo e velocidade de extração, utilizamos Regex, substituindo o anterior “SQL/INCRA”, e procurando pelo protocolo no início da linha (^\\d{4}\\-\\d{7}\\-\\d{1}), criando uma lista contendo todas as linhas bases, e assim, realizando o processo de contagem de linhas para cima e para baixo como feito anteriormente, extraindo as seguintes informações do arquivo .TXT gerado pela etapa 2:

- Nome do caderno;
- Data de emissão do PDF;
- Protocolo;
- Número do SQL/Incra;
- Responsável pessoa Física ou Jurídica;
- Status do processo;
- Tipo do processo;
- Descrição do processo.

Ao extrair essas informações, é gerado um arquivo em formato .CSV no diretório /Processado/<YYYYmmDDcaderno>.csv. O formato do arquivo foi escolhido para facilitar a importação por qualquer banco de dados relacional.

4.4 *Scraping* dos dados

Apenas essas informações ainda não são suficientes para realizar uma boa análise, então utilizamos a informação do protocolo extraídas do texto, e coletamos dados do site <<http://processos.prefeitura.sp.gov.br/Forms/consultarProcessos.aspx>>, nele é necessário fazer o seguinte processo:

Nesse site existe uma proteção contra robôs (o conhecido *Captcha*), optamos por utilizar scripts Python para poder decifrar o *Captcha* e buscar as informações. Com o Python e um conjunto de bibliotecas contendo o Selenium Web Driver, Beautysoup, Urllib e SpeechRecognition.

Com o Selenium Web Driver, para navegar de forma autônoma pelo site, sem carregar a parte visual, apenas em memória, identificamos que quando clicamos no botão “ouvir *captcha*”, é acionado uma função Javascript “TocarAudio()”, em que o navegador baixa um arquivo no formato “.mp3” chamado “pgAudio.mp3”.

Então, ao carregar o site, enviando um comando de javascript “TocarAudio()”, assim, conseguindo realizar o download do arquivo.

Para converter o arquivo .mp3 para texto, utilizamos o *SpeechRecognition*, que consegue linkar o Python com serviços externos, que reconhecem as palavras em áudio e retorna em *String*. O serviço externo que escolhemos foi o *recognize_google*, uma API gratuita do Google que tem um dos programas mais bem treinados do mundo.

Ao realizar esse processo, temos então em forma de texto, o *captcha* para enviar via formulário e conseguir acessar os dados. E gravar em formato “.csv”, onde adicionamos mais informações:

- *Assunto*;
- *Subassunto*;

- *MotivoAutuacao;*
- *DtAutuacao;*
- *Interessado;*
- *Sigla;*
- *NomeCompleto;*
- *OrgaoSupervisor;*
- *DtRecebimento;*
- *funcao;*
- *responsavelLegal;*
- *Cargo;*
- *Logradouro;*
- *-Complemento;*
- *Bairro;*
- *Cep;*
- *Telefone;*
- *Ramal;*
- *HrAtendimento;*
- *EndEletronico;*

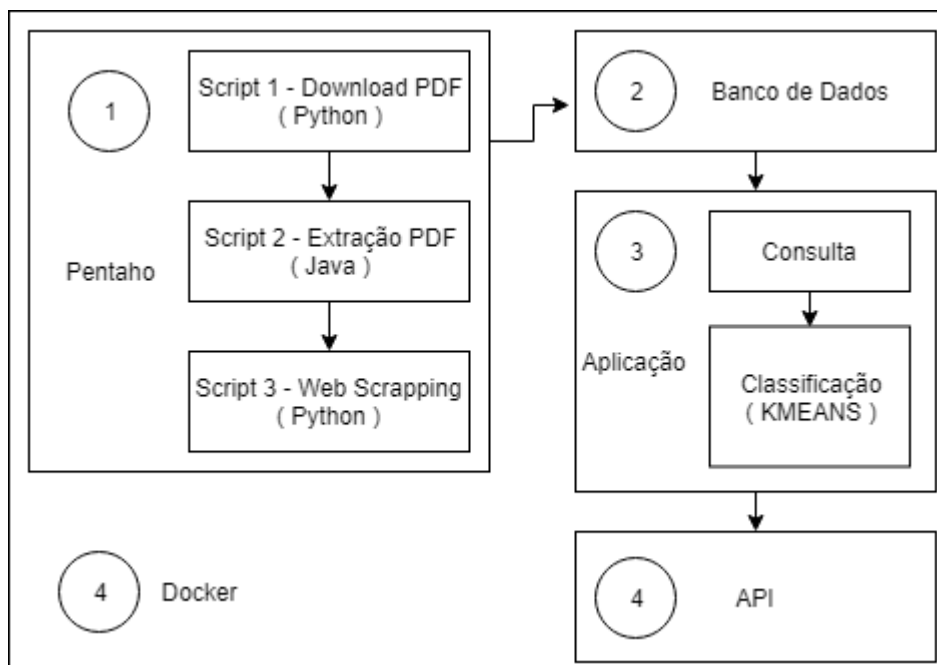
4.5 Armazenamento

Após conseguirmos todos dados necessários, importamos os arquivos .csv e armazenamos em um banco de dados, e assim, finalizamos a etapa de download dos dados.

4.6 Arquitetura

Nossa arquitetura consiste em:

Figura 9: Arquitetura



Fonte: Desenvolvida pelos autores:

No diagrama (Figura 12), está desenhada a arquitetura proposta da solução, onde:

4.6.1. Pentaho

Utilizando o Pentaho PDI (Spoon), todos os Scripts informados na etapa Lógica, serão acionados por ele, assim conseguimos monitorar o fluxo de dados de forma automática.

4.6.2. Banco de dados

Utilizando o banco de dados SQL Server, armazenaremos todos os dados encontrados para a solução.

4.6.3. Aplicação

Uma aplicação ficará responsável por realizar consultas no banco de dados e aplicará o modelo gerado pelo algoritmo, resultando em uma consulta com sua devida classificação.

4.6.4. API

Disponibilizará APIS *Rest* para que sistemas externos a solução possam acessar e consultar a aplicação, facilitando a integração do sistema.

4.6.5. Docker

Para facilitar a configuração, manutenção e uma possível réplica desse projeto, optamos por utilizar o *Docker*, onde, em qualquer máquina que tenha a instalação do *Docker* seja possível ter todo o sistema pronto para o uso por utilização de containers.

4.7 Iniciando os testes com os algoritmos

Neste capítulo será apresentado os testes realizados para utilização dos algoritmos K-Means e DBscan.

4.7.1 K-Means

Antes de carregar os dados para o python para poder rodar o algoritmo Kmeans, os dados foram pré-processados no sql server onde dados nulos foram removidos e os tipos de empresa foram transformados em números. Essa foi a lógica utilizada para a conversão dos tipos de empresa: S/S =1 ,S/A=2,LTDA=3, EIRELI=4 ,EMPRESARIO OU MICROEMPRESARIO INDIVIDUAL =5. Tínhamos também uma coluna chamada “título” onde constava o decreto e a lei a qual aquele alvará pertencia, pesquisamos o que significava cada lei e cada decreto e chegamos ao seguinte resultado: essa coluna será chamada de `categ_alvara`.

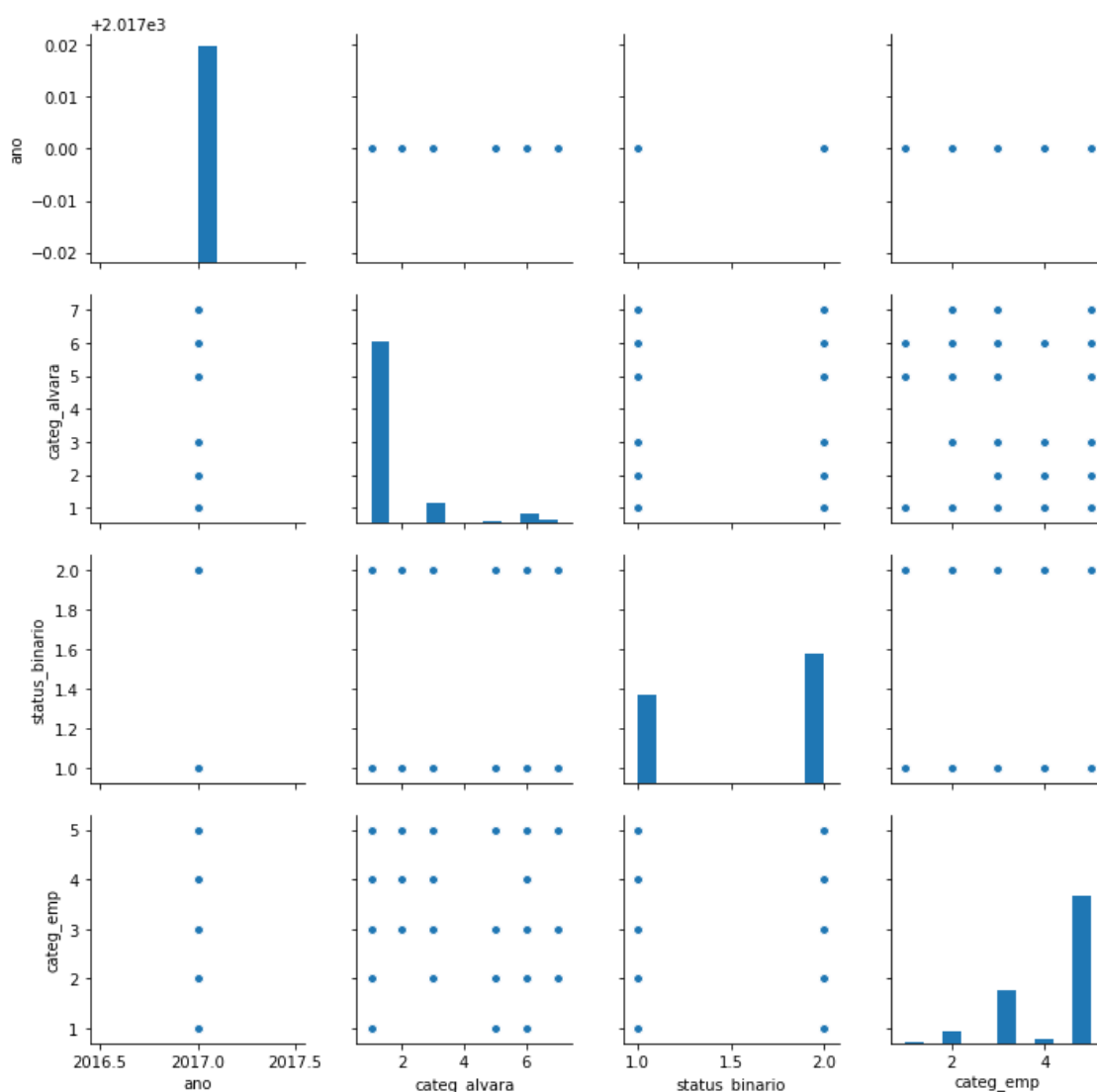
- Alvará de Funcionamento, Autorização para Eventos, Local de Reunião = 1
- Atividade não Residências Conforme Categoria de Uso = 2

- Código de Obras e Edificações e dá Outras Providências = 3
- Expedição Via Eletrônica do Alvará de Aprovação de Edificação Nova ou Reforma = 4
- Alvará Condicionado = 5
- ATENDIMENTO DO GOVERNO = 6
- Código de Obras e Edificações do Município de São Paulo = 7

Temos mais três colunas em nosso *dataset*, são elas: data, mês e status binário. Onde, 1 é deferido e 2 indeferido. A variável mês foi removida por não ter relevância em nossa base.

Carregamos a base para o Python e fizemos uma visualização dos dados pura sem nenhuma tratativa.

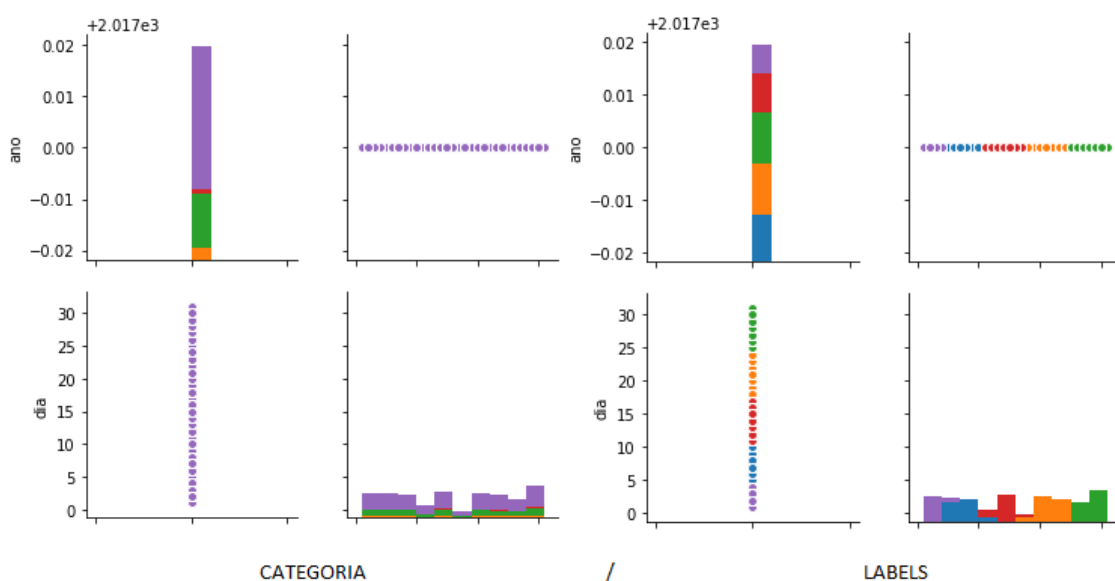
Figura 10: Exemplo de visualização de dados.



Fonte: Elaborado pelos autores

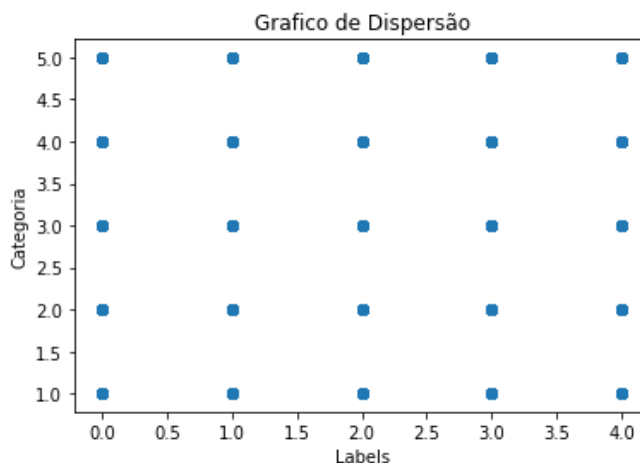
Esses são os dados antes do algoritmo Kmeans ser rodado. O Kmeans é um dos algoritmos não supervisionados mais simples de ser utilizado. A base bruta tem 5 colunas e 498476 registros. Por problemas de processamento iremos utilizar apenas 1000 registros selecionados de forma aleatória tanto no kmeans quanto no DBSCAN para os estudos.

Figura 11: Exemplo de visualização de dados:



Fonte: Elaborado pelos autores

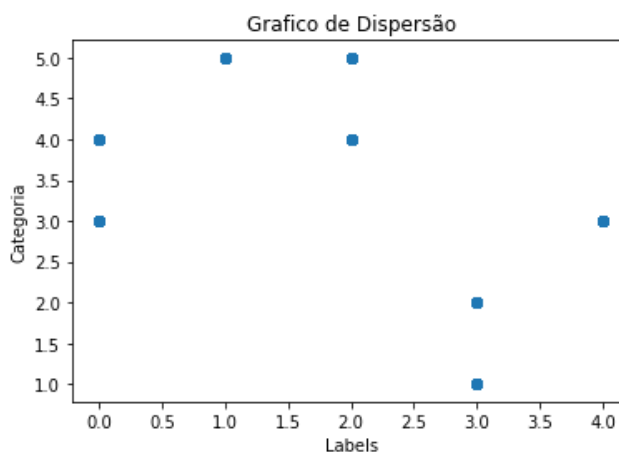
Em algum momento durante a execução do algoritmo, o Kmeans não conseguiu definir bem os grupos devido a semelhança entre eles, isso deve ter influenciado no momento de gerar os clusters, como podemos ver na primeira linha do gráfico onde temos o eixo Y como ANO e X com Categ. Podemos ver que a maioria pertence a categoria 5 que é microempreendedor. Na comparação onde foi usado as *Labels* criadas pelo algoritmo a mesma categoria foi dividida 4 vezes.

Figura 12: Exemplo de visualização de dados:

Fonte: Elaborado pelos autores

Podemos ver que a categoria de empresa 5 está em todas as *labels*, o mesmo acontece com todas as outras categorias.

Como a variável DIA não tem um peso muito grande em nosso *dataset*, vamos removê-la e fazer o teste novamente, vamos ver como ficou o gráfico de dispersão sem a variável DIA.

Figura 13: Exemplo de visualização de dados:

Fonte: Elaborado pelos autores

Podemos ver agora após a alteração que há uma distribuição melhor das categorias entre as *labels*.

4.7.2 DBSCAN

Tivemos problemas ao rodar o algoritmo em cima das 4 colunas e 498476 registros e cortando o *dataset* pela metade. Então, realizamos 4 cortes no *dataset* para poder executar o algoritmo, onde o resultado foi bem diferente do gerado pelo KMeans. E por causa da forma de processamento do DBSCAN, onde ele vai fazendo se auto gerenciando até achar a quantidade de clusters para utilizar, ele acaba sendo muito mais custoso para a máquina.

No KMeans, a estratégia utilizada para gerar os clusters foi a de que temos 5 tipos de empresa onde o Kmeans conseguiu acertar uma boa quantidade mas errou bastante. Diferente do Kmeans, onde é necessário informar a quantidade de clusters, não temos essa necessidade no DBSCAN.

Carregamos o Dataset para o python e rodamos o DBSCAN.

Figura 14: Exemplo do código

```
In [16]: #rodando o DBSCAN
         model = DBSCAN(eps=0.10,min_samples=30)

In [25]: model.fit(dados)
Out[25]: DBSCAN(algorithm='auto', eps=0.1, leaf_size=30, metric='euclidean',
               metric_params=None, min_samples=30, n_jobs=1, p=None)

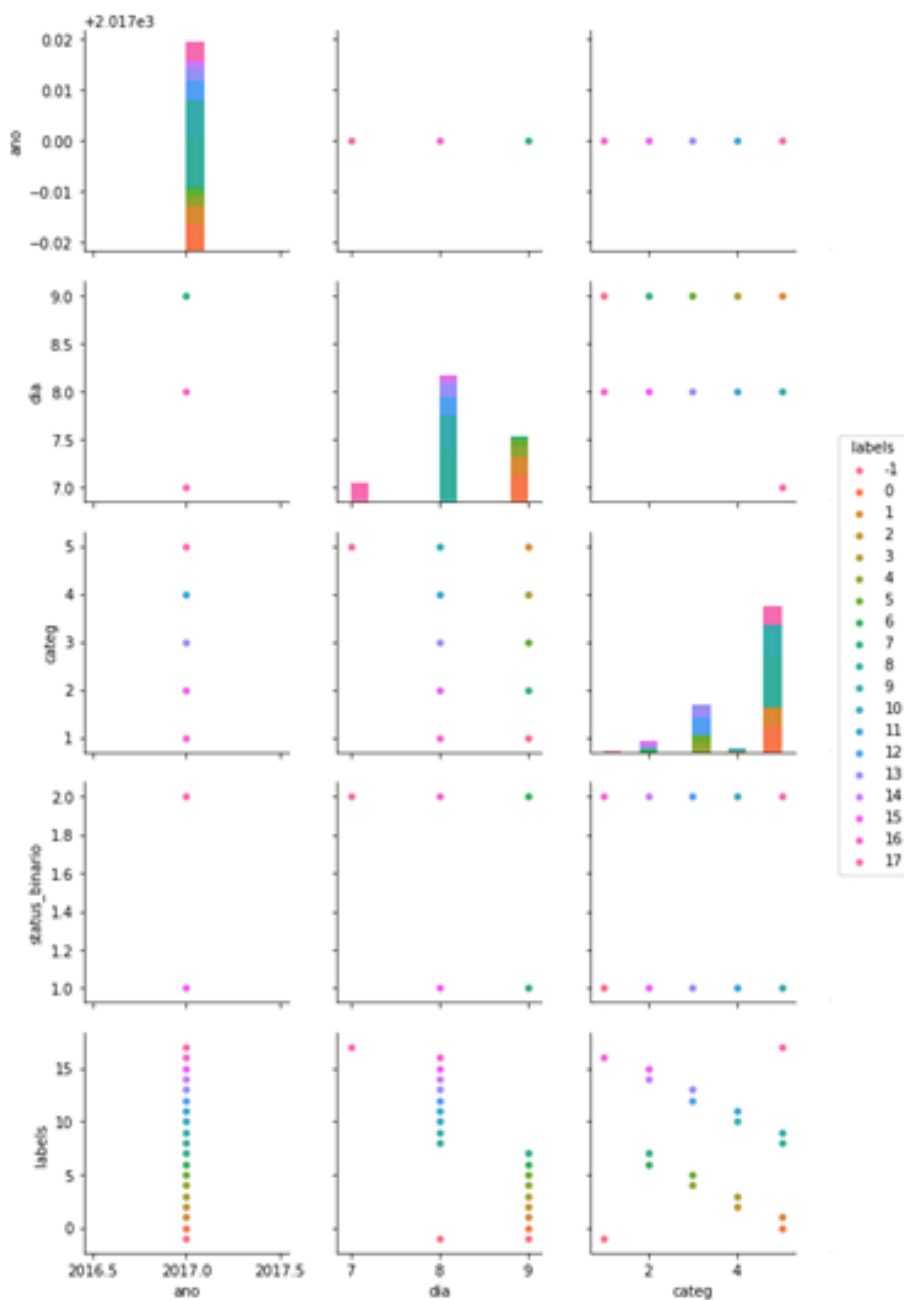
In [26]: model.labels_
Out[26]: array([ 0,  0,  0, ..., 17, 17, 17], dtype=int64)

In [21]: #Treinando o DBscan
         dados['labels'] = model.labels_
Out[21]: array([ 0,  0,  0, ..., 17, 17, 17], dtype=int64)

In [27]: dados.iloc[0:3]
Out[27]:
```

	ano	dia	categ	status_binario	labels
0	2017	9	5	2	0
1	2017	9	5	2	0
2	2017	9	5	2	0

Figura 15: Exemplo de visualização de dados:

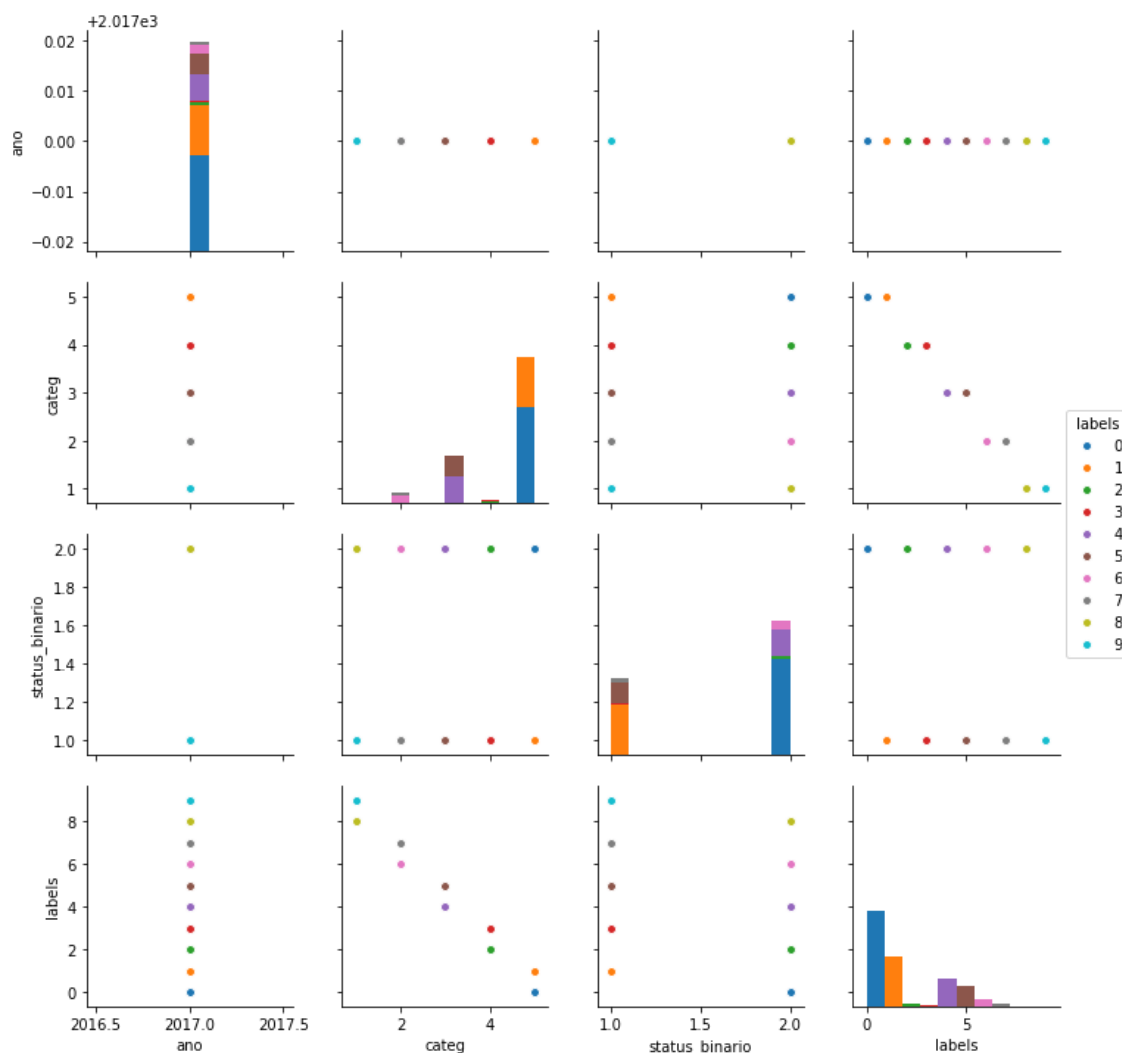


Fonte: Elaborado pelos autores

O número de *labels* gerados foi muito grande (19 *labels*). Vamos remover o DIA do *dataset* e rodar o algoritmo novamente. Passos realizados para alteração do *dataset*.

Na imagem a seguir temos o mesmo gráfico sem a categoria dia:

Figura 16: Exemplo de visualização de dados:



Fonte: Elaborado pelos autores

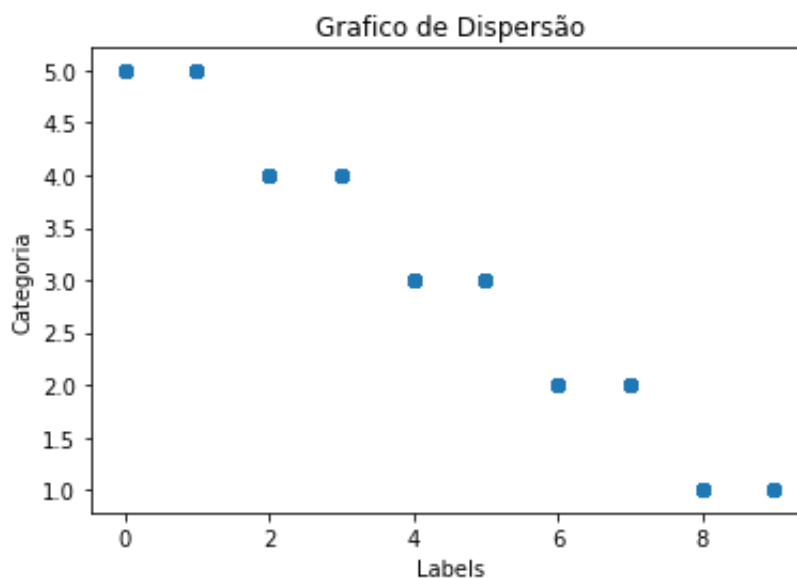
Após remover a categoria dia do *dataset* os clusters foram de 19 para 10, uma redução bem significativa.

Podemos ver que o gráfico mostra em maior quantidade as *labels* de 0 a 1 que correspondem respectivamente a categoria de empresa 5 que são os microempresários.

Comparando o resultado com o gráfico número 1 aquele em que selecionamos a Categoria das empresas como matriz de cor, esse é o que mais se assemelha.

Fizemos um gráfico de dispersão para poder entender a distribuição das *labels* em relação as categorias de empresa:

Figura 17: Exemplo de visualização de dados:



Fonte: Elaborado pelos autores

Podemos verificar que cada categoria foi separada em dois clusters, e que os grupos estão perto uns dos outros, no Kmeans os grupos ficaram mais distantes uns dos outros.

Temos em nosso *dataset* 4 variáveis são elas ANO, DIA, CATEG, STATUS_BINARIO, após as análises e testes realizados, podemos perceber que as variáveis que mais influenciam nos resultados são Categ e Dia. Com o STATUS_BINARIO sabemos quem está em dia com seus alvarás, e Ano podemos usar como medida para saber quem em maior quantidade deixa de cumprir com suas responsabilidades etc. A variável DIA no nosso *dataset* não tem muito peso, ela faz com que o algoritmo não consiga definir muito bem as fronteiras que são aqueles pontos que ficam no limite do raio da distância euclidiana após sua remoção, ambos os algoritmos conseguiram trabalhar de melhor forma, definindo melhor os clusters.

Porém, o algoritmo que conseguiu agrupar de melhor forma foi o DBSCAN.

5. Validação

Neste capítulo apresentamos os resultados e considerações de algoritmo utilizado.

5.1 Técnicas de validação para algoritmos não supervisionados

Diferente dos algoritmos de aprendizado supervisionado onde podemos fazer matriz de confusão, medir a quantidade de acertos e erros, a não supervisionada é diferente. Como não temos os rótulos com os alvos definidos esse tipo de análise não é possível. Temos que utilizar de outras técnicas de validação, nelas podemos ver se a quantidade de *labels* criadas pelo algoritmo e a quantidade ideal, assim podendo definir a melhor estratégia e os melhores parâmetros para os algoritmos que serão utilizados. Utilizaremos 2 tipos de validação: *silhouette* e *elbow* (ou cotovelo).

5.1.1 Técnica de validação *silhouette*

Essa técnica mede a proximidade de cada ponto em um cluster com os pontos em seus clusters vizinhos. No Kmeans ele nos ajuda a determinar a quantidade ideal de K durante o agrupamento. Os valores de *silhouette* variam de 1 a -1, onde quanto mais próximo de 1 melhor, o valor 1 indica que o exemplo está longe de seu cluster vizinho e perto dos clusters a qual está designado, o -1 indica que o ponto está próximo ao cluster vizinho e distante do seu cluster, portanto quanto maior o número e quanto mais perto de 1 melhor.

5.1.2 *Elbow* (cotovelo)

Essa técnica consiste basicamente em testar a variância dos dados em relação ao número de clusters, no instante em que o número de clusters aumenta e esse aumento não representa um valor significativo a técnica para.

Tem esse nome de cotovelo porque ao plotar esse gráfico o mesmo lembra a articulação humana.

A partir do valor indicado pelo “cotovelo” no gráfico, significa que não existe ganho em relação ao aumento de clusters. O que significa que a *elbow* encontrou o número ideal para o K.

5.2 Testes com algoritmo DBSCAN

Primeiro rodamos os dois tipos de algoritmos e analisamos os resultados, para depois escolher o melhor.

A imagem a seguir mostra o código onde o algoritmo DBSCAN é rodado. Em seguida é feito o treinamento no *dataset* com o método `fit()`, e logo após as *labels* criadas foram passadas para o *dataset*:

Figura 18: Exemplo do Código

```
dbscan = DBSCAN(eps=0.10,min_samples=0.30)
dbscan.fit(dados)
dbscan.labels_

array([ 0,  0,  0, ..., 19, 19, 19], dtype=int64)
```

```
#Treinando o DBscan
dados['labels'] = dbscan.labels_
```

```
dados.iloc[0:4]
```

	ano	mes	categ_alvara	status_binario	categ_emp	labels
0	2017	8	1	1	5	0
1	2017	8	1	1	5	0
2	2017	8	1	1	5	0
3	2017	8	1	1	3	1

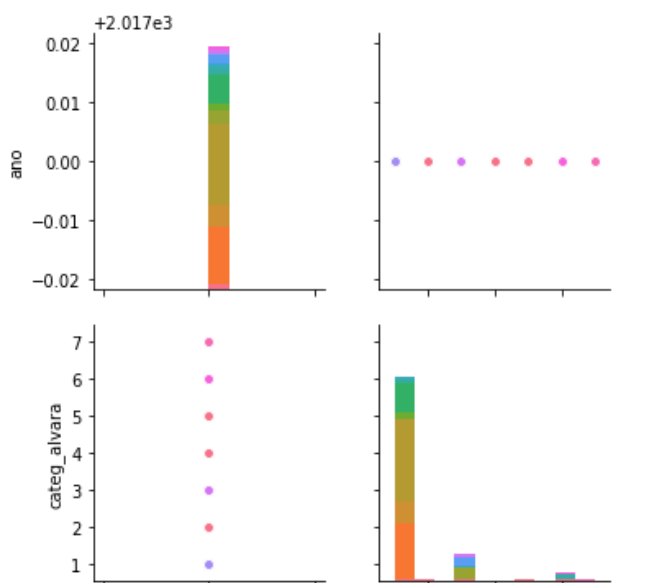
Fonte: Elaborado pelos autores

No DBSCAN após fazer o treinamento com o *dataset*, gerou um número de 28 *labels*, número muito grande que dificultaria a análise dos grupos. De dados importantes temos apenas tipo de empresa, tipo de alvará e o status do alvará. Após fazer a remoção do variável mês, o número de 28 foi para 16, que ainda sim é muito alto, isso pode ter ocorrido devido a estratégia utilizada pelo algoritmo que é estratégia de densidade, onde ele escolhe um ponto aleatório e por meio da distância euclidiana ele verifica todos os pontos que são alcançáveis

por ela. Caso algum ponto não possa ser alcançado por essa distância ele escolhe um novo ponto central e inicia o processo novamente.

Podemos ver com os gráficos gerados a densidade dos grupos:

Figura 19: Exemplo de visualização de dados.



Fonte: Elaborado pelos autores.

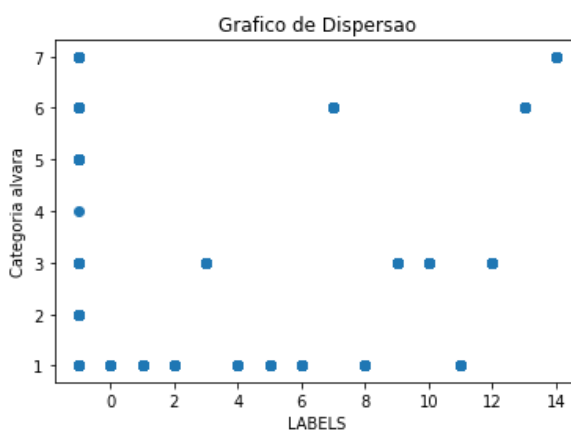
O DBSCAN com base na sua estratégia define qual seria a melhor quantidade de clusters que nesse caso foi de 16, ao rodar o *silhouette*, como o DBSCAN já define a quantidade de clusters. Todos os valores foram iguais. Segundo o entendimento da técnica e dos dados, o DBSCAN identificou que todos com os mesmos grupos, pois, não houve relação entre as variáveis, mas sim, ele identificou apenas os grupos de variáveis e realizou o agrupamento das mesmas.

- For n_clusters=2, The Silhouette Coefficient is 0.978283452985423
- For n_clusters=3, The Silhouette Coefficient is 0.978283452985423
- For n_clusters=4, The Silhouette Coefficient is 0.978283452985423
- For n_clusters=5, The Silhouette Coefficient is 0.978283452985423
- For n_clusters=6, The Silhouette Coefficient is 0.978283452985423

- For $n_clusters=7$, The Silhouette Coefficient is 0.978283452985423
- For $n_clusters=8$, The Silhouette Coefficient is 0.978283452985423
- For $n_clusters=9$, The Silhouette Coefficient is 0.978283452985423
- For $n_clusters=10$, The Silhouette Coefficient is 0.978283452985423

Na figura abaixo, temos uma amostra realizada com gráfico de dispersão:

Figura 20: Exemplo de visualização de dados



Fonte: Elaborado pelos autores.

No gráfico acima podemos ver a distribuição das categorias de empresa entre as *labels*.

5.3 Testes com algoritmo KMeans

No KMeans onde K é o número de clusters, colocamos k igual a 9 e vimos o resultado. Comparando ele com os resultados obtidos no DBSCAN, conseguimos identificar melhor onde estava cada grupo e o que eles poderiam nos dizer. Abaixo temos a imagem com o algoritmo sendo rodado, treinado, e suas *labels* implementadas no *dataset*.

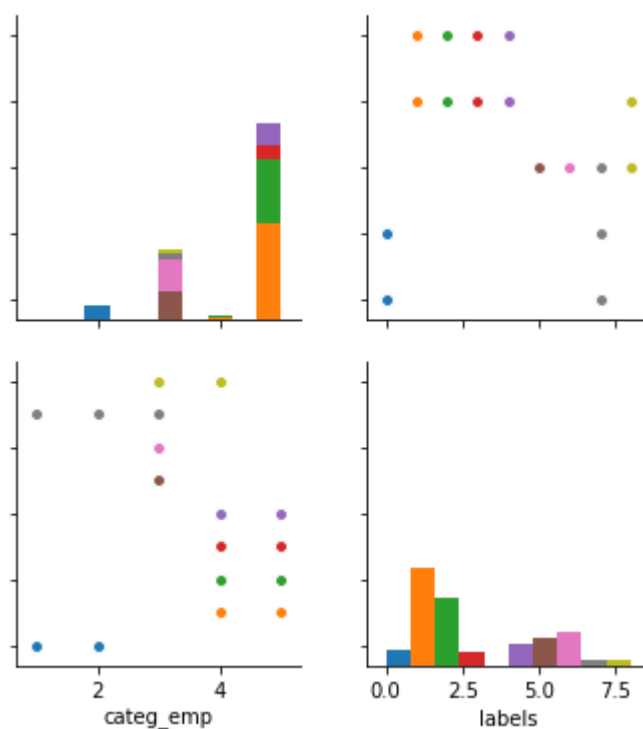
Figura 21: Exemplo do Código

```
: kmeans = KMeans(n_clusters = 9, random_state=0)
#treinando o algoritmo
kmeans.fit(dados)
#visualização dos grupos (Labels) criadas
kmeans.labels_
#passando os grupos criados para o DataSet
dados['labels'] = kmeans.labels_
```

Fonte: Elaborado pelos autores

Nos gráficos abaixo temos os grupos criados e as categorias de empresa, onde as *labels* criadas foram usadas como matriz de cor.

Figura 22: Exemplo de visualização de dados



Fonte: Elaborado pelos autores.

Podemos ver que o grupo de *labels*laranja é o grupo que possui mais empresas.

Uma das Desvantagens do Kmeans é definir a melhor quantidade de K grupos. Vamos usar o *silhouette* e ver qual seria a melhor quantidade de K grupos.

Código utilizado para validação:

Figura 23: Exemplo do código

```
X = dados
for n_cluster in range(2,11):
    kmeans = KMeans(n_clusters=n_cluster).fit(X)
    label = kmeans.labels_
    sil_coeff = silhouette_score(X, label, metric='euclidean')
    print("For n_clusters={}, The Silhouette Coefficient is {}".format(n_cluster, sil_coeff))
```

Fonte: Elaborado pelos autores.

Resultado do *silhouette*:

- For n_clusters=2, The Silhouette Coefficient is 0.5442973069751355
- For n_clusters=3, The Silhouette Coefficient is 0.6414068469749596
- For n_clusters=4, The Silhouette Coefficient is 0.7748767749382865
- For n_clusters=5, The Silhouette Coefficient is 0.8447413594540698
- For n_clusters=6, The Silhouette Coefficient is 0.8443131222268712
- For n_clusters=7, The Silhouette Coefficient is 0.8435426987409818
- For n_clusters=8, The Silhouette Coefficient is 0.8620799301757622
- For n_clusters=9, The Silhouette Coefficient is 0.900261363099737
- For n_clusters=10, The Silhouette Coefficient is 0.9054138113498571

Quanto mais próximo de 1, melhor, e quanto mais próximo de -1, pior. Podemos ver que n_clusters=5 saiu-se muito bem 0.8447413594540698, depois disso os ganhos não são tão significativos.

No KMeans conseguimos fazer um segundo tipo de validação, chamado ELBOW ou cotovelo, tem esse nome porque após o gráfico ser gerado lembra a aparência de um cotovelo.

Código da técnica de validação:

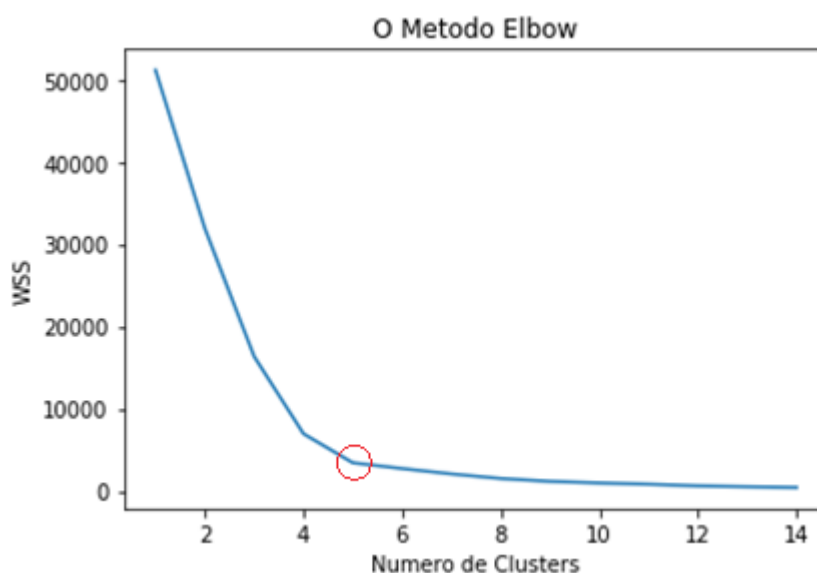
Figura 24: Exemplo do código

```
#Tecnica de validação cotovelo

from sklearn.cluster import KMeans
wcss = []

for i in range(1,15):
    kmeans = KMeans(n_clusters = i, random_state=0)
    kmeans.fit(X)
    print (i,kmeans.inertia_)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 15), wcss)
plt.title('O Metodo Elbow')
plt.xlabel('Numero de Clusters')
plt.ylabel('WSS') #within cluster sum of squares
plt.show()
```

Fonte: Elaborado pelos autores.

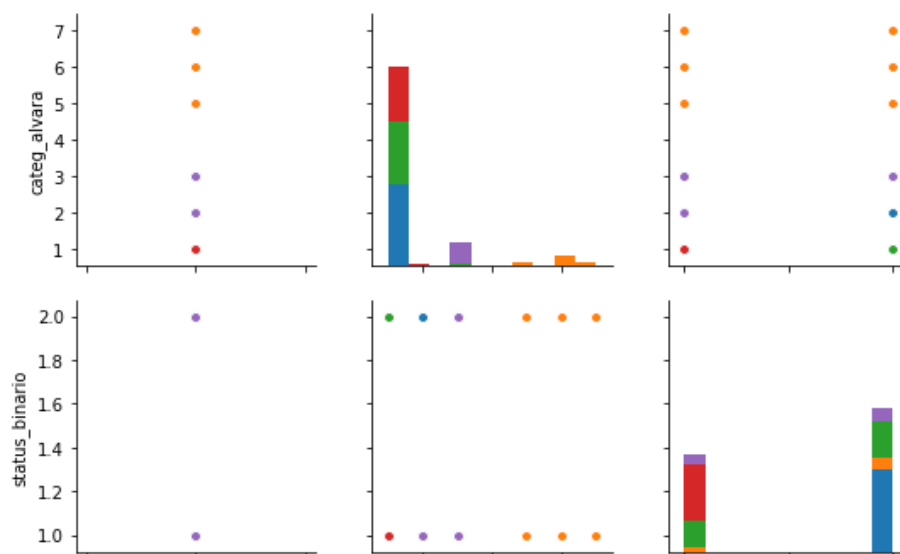
Figura 25: Exemplo de visualização de dados

Fonte: Elaborado pelos autores.

A partir do valor indicado pelo “cotovelo” no gráfico, significa que não existe ganho em relação ao aumento de clusters. Nesse caso, 5 foi o número ideal. Temos duas técnicas de validação onde ambas indicam 5 como o melhor número de clusters, então colocamos k igual a 5.

Nos gráficos a seguir, veremos os resultados obtidos:

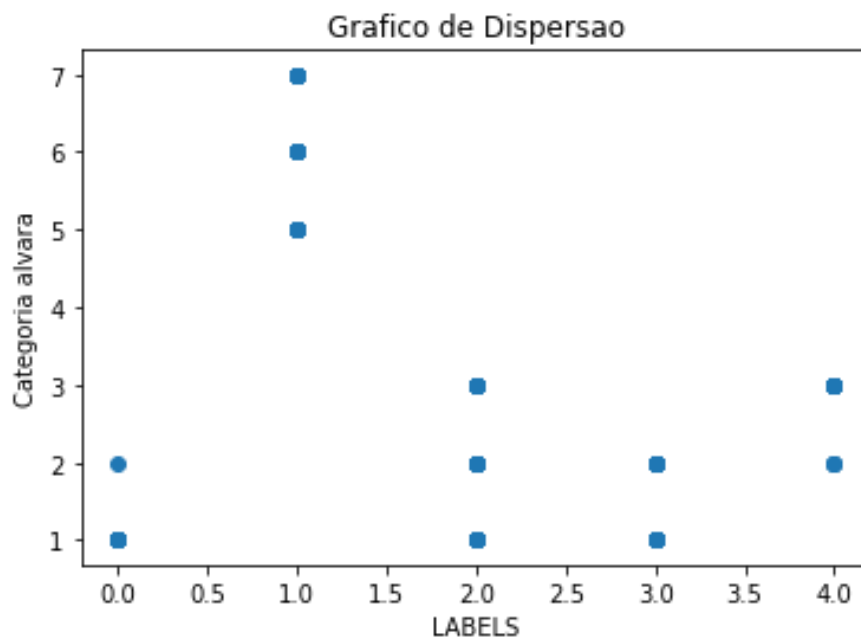
Figura 26: Exemplo de visualização de dados



Fonte: Elaborado pelos autores.

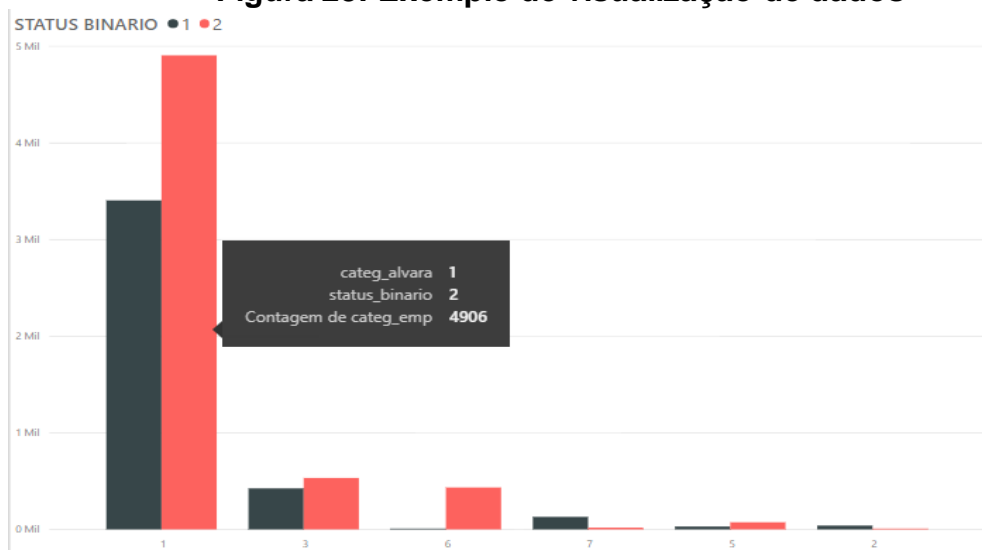
Podemos ver no gráfico abaixo a dispersão entre os alvarás e as *labels*:

Figura 27: Exemplo de visualização de dados



Fonte: Elaborado pelos autores.

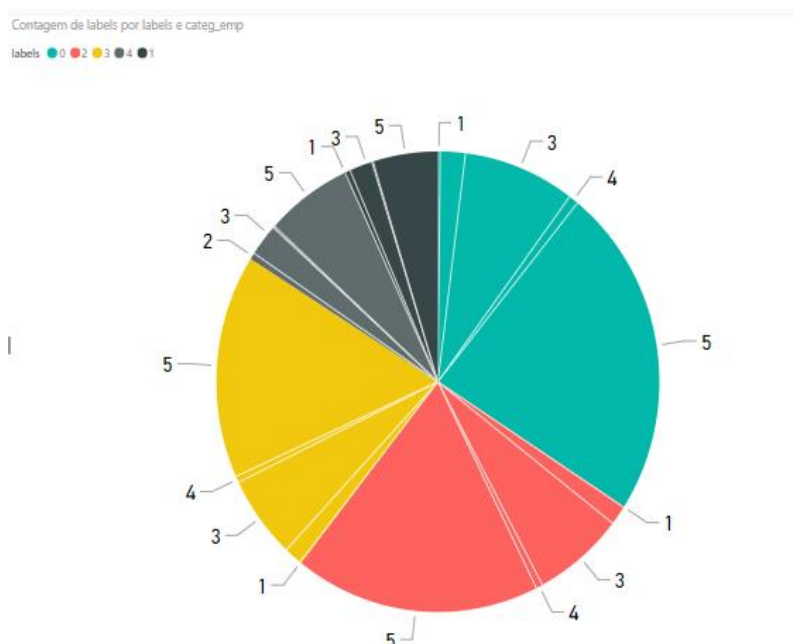
Abaixo teremos um exemplo de visualização em gráficos gerados pela ferramenta Power BI, onde podemos ter uma melhor visão sobre os resultados:

Figura 28: Exemplo de visualização de dados

Fonte: Elaborado pelos autores.

Podemos ver com o gráfico de barras anterior que 4906 empresas têm o alvará 1 que é o de funcionamento. A maioria dos acidentes que vemos em noticiários envolvem esse tipo de alvará, como podemos ver um número bem elevado está indeferido.

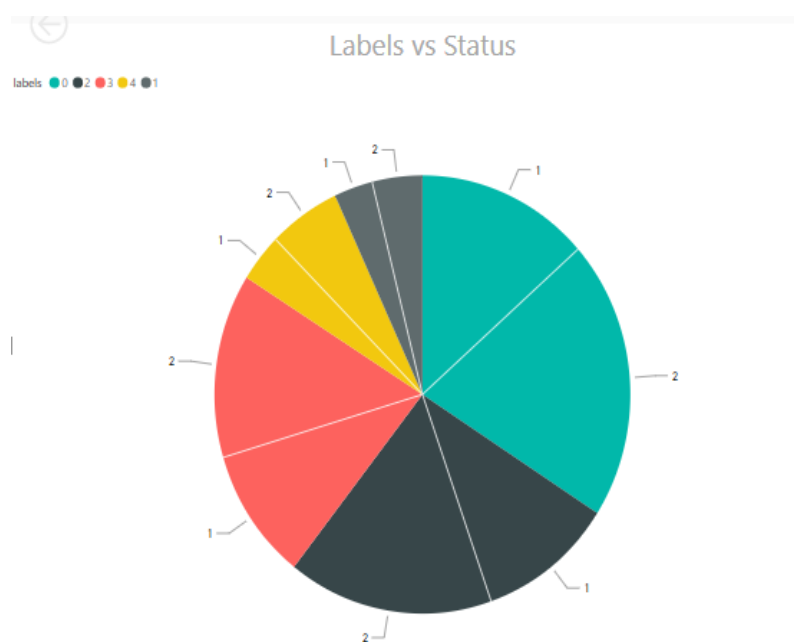
No gráfico abaixo podemos ver a distribuição das empresas dentro dos clusters criados:

Figura 29: Exemplo de visualização de dados

Fonte: Elaborado pelos autores.

No gráfico a seguir temos o status onde 1 é deferido e 2 indeferido. Podemos concluir que a quantidade de indeferidos é enorme:

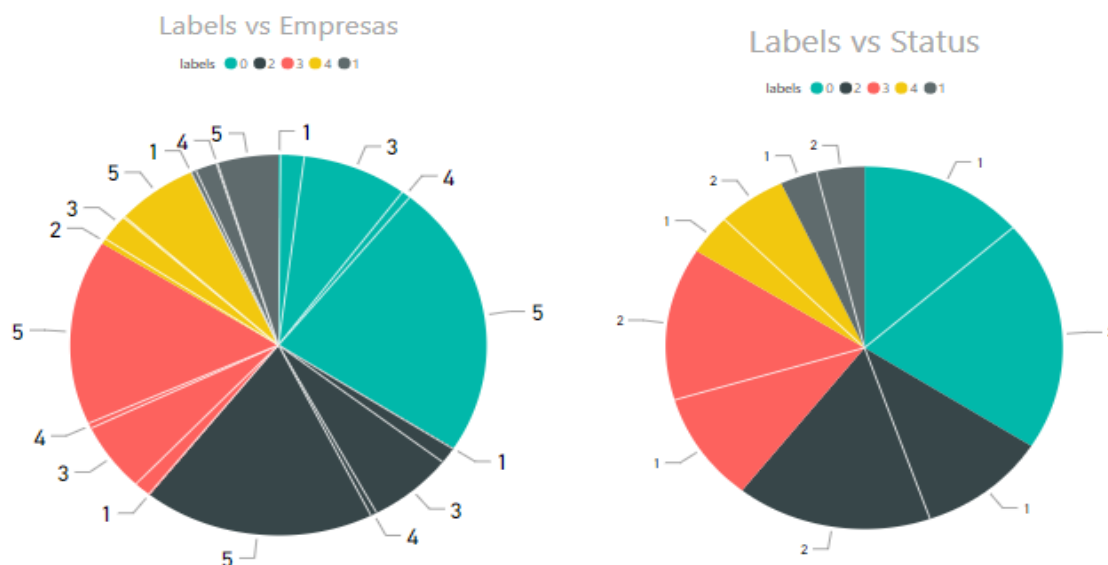
Figura 30: Exemplo de visualização de dados



Fonte: Elaborado pelos autores.

Com essas informações podemos criar uma escala de período, com alto, médio e baixo.

Figura 31: Exemplo de visualização de dados



Fonte: Elaborado pelos autores.

Os clusters criados conseguem nos dizer quais empresas podem representar um maior perigo, devido ao fato da mesma estar com seu alvará indeferido. Para acrescentar mais valor as análises, dados como, análise de laudos e incidentes seriam necessários.

5.4 Considerações finais

Depois de executar os dois algoritmos de aprendizado não supervisionado, vimos que o KMEANS foi o que se saiu melhor, nele podemos definir a quantidade de clusters.

No DBSCAN após fazer o treinamento com o *dataset*, ele gerou um número de 28 *labels*, número muito grande que dificultaria a análise dos grupos. De dados importantes, temos apenas tipo de empresa, tipo de alvará e o status do alvará. Após fazer a remoção do variável mês, o número de 28 foi para 16, que ainda assim é muito alto, isso pode ter ocorrido devido a estratégia utilizada pelo algoritmo que é estratégia de densidade. Onde, ele escolhe um ponto aleatório e traça um raio em volta dele todos os pontos que estiverem dentro desse raio, pertencem ao mesmo cluster. Caso contrário um novo ponto é escolhido e o processo inicia-se novamente.

Analisando os gráficos da figura 31, podemos dizer que o grupo dos grandes impactantes são as micro e pequenas empresas e que a maioria está

com seus respectivos alvarás indeferidos, representando assim um risco elevado. Essas micro e pequenas empresas tem um auto índice de envolver-se em acidentes. Nessas empresas, a vistoria deveria ser feita de uma forma mais intensiva para garantir que a lei seja cumprida, evitando assim que o local funcione de forma irregular com o intuito de evitar acidentes e que vidas sejam perdidas.

5. Conclusão

Em nossa hipótese, é possível prever problemas estruturais devido a irregularidade nos alvarás, como não temos um alvo para realizar uma análise mais precisa, foi escolhido um algoritmo não supervisionado (K-Means) para identificar uma possível relação entre os dados encontrados e um agrupamento que podemos identificar como um grupo de risco ou não.

Um grande problema nessa solução, foi justamente conseguir dados que possam auxiliar em nossa solução. Encontramos com muita dificuldade, dados sobre alvarás e infelizmente, todos são descritivos, não havendo dados como valores detalhados, número de CREA de engenheiros, cartório que foi regulamentado, etc.

Em nossa busca, encontramos alguns desses dados em textos descritivos dentro de outros cadernos do diário oficial, como o caderno jurídico onde houveram processos informando esses dados. Porém, para extrair essa informação torna-se muito difícil por conta de o texto não ter nenhuma estrutura.

Existe uma iniciativa do governo chamada portal brasileiro de dados abertos (dados.gov.br), onde é disponibilizado diversas bases de dados de forma gratuita, visto que, futuramente, se for disponibilizado bases de dados contendo informações detalhadas sobre alvarás, muito possivelmente, é possível utilizar a nossa solução para criar clusters mais assertivos, e assim, conseguindo um resultado muito mais preciso.

Como também não temos a informação de qual tipo de grupo houve acidente ou não, não temos um alvo, então foi escolhido um algoritmo não supervisionado para a solução.

Após testes com o KMEANS e o DBSCAN, optamos pelo KMEANS e identificamos diversos grupos, e que o algoritmo KMEANS conseguiu realizar muito bem a sua função, porém, os dados encontrados não geram uma confiança o suficiente para utilizar o resultado do algoritmo para uma tomada de decisão.

Analisando os dados de cada grupo, foi identificado um grande número de microempresas. Como não conseguimos contato com um especialista na área para confirmar se os clusters que nosso algoritmo gerou faz sentido ou não,

procuramos motivos para um excessivo número de microempresas, e encontramos diversos materiais mostrando que:

Com o forte aumento das demissões e a redução das vagas com carteira assinada, muitas pessoas parecem estar tentando a sorte como autônomos. E as MEIs podem auxiliar nesse processo de formalização do trabalho, segundo o especialista do G1.Globo.com (2016).

No Brasil, com o problema de economia e crises geradas através dos anos, houve um crescente problema de falta de trabalho no país. Nossa população está a cada ano procurando formas diferentes de inovar e garantir o seu próprio sustento, explicando a existência de tantos alvarás de microempresas.

Novamente, não temos como comprovar por falta de um especialista, mas supomos que, todas essas empresas realmente se encontram em um grau de alto risco, pois, o alvará garante que a empresa está devidamente alocada. Porém, não há uma forte fiscalização sobre este documento. Casos gravíssimos são reportados nas diversas mídias, como já citamos neste trabalho, e causaram prejuízos inestimáveis, mortes e problemas ambientais gravíssimos, e isso com empresas gigantescas. Imaginamos o risco que diversas pequenas empresas sem fiscalização podem oferecer também.

Infelizmente, nossa hipótese deve ser considerada inconclusiva, pois, mesmo analisando os gráficos e os dados, não conseguimos comprovar com exatidão se as empresas listadas estão ou não com um grau de risco alto ou baixo.

Referências Bibliográficas

CARDOSO, G. et al. **Clusterização K-Means: Uma Proposta de Melhoria.**

Disponível em:

<http://cursos.eldman.com.br/pluginfile.php/3848/mod_resource/content/10/Clusterizacao%20K%20means%20Uma%20Proposta%20de%20Melhoria.pdf>.

Acesso em: 15 jan. 2019.

CHAPMAN, Pete. **CRISP-DM methodology**. 1ª Edição, USA, SPSS, 2000.

Consulta previa do local: Aprenda a fazer para sua empresa. Disponível em

<<http://cultivandosuccesso.com/consulta-previa-do-local-aprenda-como-fazer-para-sua-empresa/>>. Acesso em: 16 set. 2018.

Consulta Prévia. Disponível em:

<https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/sp_mais_facil/slea/consulta_previa/index.php?p=9442>. Acesso em: 17 set. 2018.

DBSCAN. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>>. Acesso em: 10 jan. 2019.

Density-Based Clustering. Disponível em: <http://wiki.icmc.usp.br/images/c/ca/Density_Based_Clustering.pdf>. Acesso em 06 jan. 2019.

Devmedia. Data Mining na Prática: Algoritmo K-Means. Disponível em: <https://www.devmedia.com.br/data-mining-na-pratica-algoritmo-k-means/4584>. Acesso em: 15 jan. 2019.

ESTER et al. **A Density Based Spatial Clustering of Applications with Noise**, 1ª Edição, USA, AAAI, 1996.

FORGY, E.W. **Cluster Analysis of Multivariate data: efficiency vs. Interpretability of Classifications**. Vol.21, Biometrics, 1965.

História do Python. Disponível em: <<http://python-history-pt-br.blogspot.com/2009/06/historia-pessoal-parte-1-cwi.html>>. Acesso em: 17 set. 2018.

História do Python. Disponível em: <<https://wiki.python.org.br/HistoriaDoPython>>. Acesso em: 17 set. 2018.

Legislação. Disponível em: <https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/sp_mais_famil/legislacao/>. Acesso em: 19 set. 2018.

Licenciamento Ambiental: Prazos e Validade. Disponível em: <<https://www.cenedcursos.com.br/meio-ambiente/licenciamento-ambiental-prazos-e-validade/>>. Acesso em: 08 de outubro de 2018.

MITCHELL, Ryan. **Web Scraping com Python**, 1ª Edição, São Paulo, Novatec, 2015.

O Método DBSCAN. Disponível em: <https://www.maxwell.vrac.puc-rio.br/24787/24787_6.PDF>. Acesso em: 15 jan. 2019.

O que é a Consulta Prévia de endereço e atividade? Onde fazer a consulta prévia? Disponível em: <<http://www.portaldoempreendedor.gov.br/duvidas-frequentes/informe-se-antes-de-formalizar/2.1-o-que-e-a-consulta-previa-de-endereco-e-atividade-onde-fazer-a-consulta-previa>>. Acesso em: 16 set. 2018.

Quando e como renovar o Certificado de Licença de Funcionamento emitido pela Polícia Federal para controle de produtos químicos?

Disponível em: <<https://iusnatura.com.br/quando-e-como-renovar-o-certificado-de-licenca-de-funcionamento-emitido-pela-policia-federal-para-controle-de-produtos-quimicos/>>. Acesso em: 08 out. 2018.

Renovação de Auto de Vistoria do Corpo de Bombeiros (AVCB). Disponível em: <<http://www.bombeiros.mg.gov.br/component/content/article/475-renovacao-de-auto-de-vistoria-do-corpo-de-bombeiros-avcb.html>>. Acesso em: 08 out. 2018.

ROBERTO, Marcio. **Alvará e inscrições municipais e estaduais necessárias para abrir uma empresa.** Disponível em: <<https://blog.contaazul.com/tipos-de-alvara-e-inscricoes-municipais-e-estaduais-para-abrir-empresa/>>. Acesso em: 17 set. 2018.

SAIBA TUDO SOBRE ALVARÁ DE FUNCIONAMENTO! Disponível em: <<https://aberturasimples.com.br/alvara-de-funcionamento/>>. Acesso em: 18 set. 2018.

SCIKIT LEARN. Disponível em: <<https://scikit-learn.org/stable/modules/clustering.html>>. Acesso em: 09 mar. 2019.

Serviços para Empresa. Disponível em: <<http://www.capital.sp.gov.br/empresa/alvaras-certidoes-e-licencas/licenciamento-eletronico-de-atividades-slea/legislacao>>. Acesso em: 19 set. 2018.

SWEIGART, AL. **Automatize Tarefas Maçantes Com Python:** Programação Prática Para Verdadeiros Iniciantes. Edição 1, Brasil, Novatec, 2015.

Três letras que mudaram o mundo. Disponível em: <<https://acrobat.adobe.com/pt/pt/acrobat/about-adobe-pdf.html>>. Acesso em: 17 set. 2018.

ZIULKOSHI, L.C.C. **Coleta de Requisitos e Modelagem de Dados para Data Warehouse: um Estudo de Caso utilizando Técnicas de Aquisição de Conhecimento.** Porto Alegre: Projeto de Diplomação-Universidade Federal do Rio Grande do Sul Instituto de Informática, 2003.