Pontifícia Universidade Católica
do Rio de Janeiro

Marcus Vinicius Giollo Cesar

Classificação de falhas de equipamentos de unidade de intervenção em construção de poços marítimos por meio de mineração textual

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientadora: Profa. Marley Maria Bernardes Rebuzzi Vellasco

Co-orientadora: Profa. Karla Tereza Figueiredo Leite

Rio de Janeiro
Abril de 2017

Pontifícia Universidade Católica
do Rio de Janeiro

Marcus Vinicius Giollo Cesar

Classificação de falhas de equipamentos de unidade de intervenção em construção de poços marítimos por meio de mineração textual.

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Marley Maria Bernardes Rebuzzi Vellasco
Orientadora
Departamento de Engenharia Elétrica – PUC-Rio

Profa. Karla Tereza Figueiredo Leite
Co-Orientadora
UERJ

Prof. Ricardo de Melo e Silva Accioly
UERJ

Prof. Rafael de Olivaes Valle dos Santos
Petróleo Brasileiro - Rio de Janeiro - Matriz

Prof. Márcio da Silveira Carvalho
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 28 de abril de 2017

Marcus Vinicius Giollo Cesar

Graduou-se em Engenharia Aeronáutica pela Universidade de São Paulo, Campus São Carlos, São Paulo – Brasil em 2009. Atualmente exerce o cargo de Engenheiro de Petróleo na Petrobras S.A., atuando na área de Segurança das Operações de Poços Marítimos.

Ficha Catalográfica

Cesar, Marcus Vinicius Giollo

Classificação de falhas de equipamentos de unidade de intervenção em construção de poços marítimos por meio de mineração textual / Marcus Vinicius Giollo Cesar; orientadora: Marley Maria Bernardes Rebuzzi Vellasco ; co-orientadora: Karla Tereza Figueiredo Leite. – 2017.
171 f. : il. color. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2017.
Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Falhas de equipamentos. 3. Unidade de intervenção. 4. Poços marítimos. 5. Classificação textual. 6. Support Vector Machines. I. Vellasco, Marley Maria Bernardes Rebuzzi. II. Leite, Karla Tereza Figueiredo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Agradecimentos

Primeiramente a Deus, por permitir que este trabalho se concretizasse.

À minha querida esposa Rosinha, cujo meu empenho não teria sido igual neste trabalho se não tivesse recebido apoio e compreensão de sua parte.

Aos meus pais, Rosemary e Luis Antonio, pela educação, atenção e carinho.

Às minhas orientadoras Marley Vellasco e Karla Figueiredo, pelo excelente apoio desde o início até a conclusão deste trabalho, apesar da distância
.

À Petrobras, por patrocinar e por prover este excelente ambiente de engenharia de ponta e de grande relevância mundial.

Ao Ricardo Accioly, que me orientou a seguir o caminho da mineração de dados, exemplo de excelente profissional.

À gerência geral de Engenharia de Poço sob a liderança de Luiz Felipe Rego e à gerência de Eficiência e Otimização de Engenharia de Poço que, sob a liderança da gerente Helena Lobato, patrocinou e acreditou neste trabalho.

A todos os familiares, amigos e colegas de trabalho que de uma forma ou de outra me estimularam e me ajudaram.

Resumo

Cesar, Marcus Vinicius Giollo; Vellasco, Marley Maria Bernardes Rebuzzi. **Classificação de falhas de equipamentos de unidade de intervenção em construção de poços marítimos por meio de mineração textual**. Rio de Janeiro, 2017. 171p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A construção de poços marítimos tem se mostrado uma atividade complexa e de alto risco. Para efetuar esta atividade as empresas se valem principalmente das unidades de intervenção de poços, também conhecidas como sondas. Estas possuem altos valores de taxas diárias de uso devido à manutenção preventiva da unidade em si, mas também por falhas as quais seus equipamentos estão sujeitos.
No cenário específico da Petrobras, em junho de 2011, foi implantado no banco de dados da empresa um maior detalhamento na classificação das falhas de equipamentos de sonda. Com isso gerou-se uma descontinuidade nos registros da empresa e a demanda para adequar estes casos menos detalhados à classificação atual, mais completa. Os registros são compostos basicamente de informação textual. Para um passivo de 3384 registros, seria inviável alocar uma pessoa para classificá-los. Com isso vislumbrou-se uma ferramenta que pudesse efetuar esta classificação da forma mais automatizada possível, utilizando os registros feitos após junho de 2011 como base. O objetivo principal deste trabalho é de sanar esta descontinuidade nos registros de falha de equipamentos de sonda. Os dados foram tratados e transformados por meio de ferramentas de mineração textual bem como processados pelo algoritmo de aprendizado supervisionado SVM (*Support Vector Machines*). Ao final, após obter a melhor configuração do modelo, este foi aplicado às informações textuais do passivo de anormalidades, atribuindo suas classes de acordo com o novo sistema de classificação.

Palavras-chave

Falhas de equipamentos de Unidade de intervenção em poços marítimos; Classificação Textual; Máquinas de Vetores de Suporte.

Abstract

Cesar, Marcus Vinicius Giollo; Vellasco, Marley Maria Bernardes Rebuzzi (Advisor). **Text classification of offshore rig equipment failure.** Rio de Janeiro, 2017. 171p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Off-shore well construction has shown to be a complex and risky activity. In order to build off-shore wells, operators rely mainly on off-shore rigs. These rigs have an expensive day rate, related to their rental and maintenance, but also due to their equipment failure. At off-shore Petrobras scenario, on June of 2011, was implemented at the company database a better detailing on the classification of rig equipment failure. That brought a discontinuity to the database records and created a demand for adequacy of the former classification to the new classification structure. Basically, rig equipment failure records are based on textual information. For a liability of 3384 records, it was unable for one person to manage the task. Therefore, an urge came for a tool that could classify these records automatically, using database records already classified under the new labels. The main purpose of this work is to overcome this database discontinuity. Data was treated and transformed through text mining tools and then processed by supervised learning algorithm SVM (Support Vector Machines). After obtaining the best model configuration, the old records were submitted under this model and were classified according to the new classification structure.

Keywords

Offshore rig equipment failure; Text classification; Support Vector Machines.

Sumário

Lista de figuras

Lista de tabelas

Lista de equações

Now, don't hang on, nothing lasts forever but the earth and sky
It slips away, and all your money won't another minute buy

Dust in the wind
All we are is dust in the Wind...

Kansas

1.
Introdução

1.1
Motivação

A variável custo possui destaque na indústria de construção de poços marítimos. Grande parcela deste custo (além dos materiais utilizados na construção do poço) está relacionada às taxas diárias das unidades de intervenção, também conhecidas como plataformas de perfuração ou simplesmente como sondas. A busca por petróleo, iniciada em poços terrestres, se deslocou para águas rasas e seguiu para águas cada vez mais profundas.

Devido à alta complexidade de construir com segurança e em tempo hábil poços submarinos que estão geralmente a 2000 m de profundidade abaixo do nível do mar (cenário típico de poços de extração de óleo da camada pré-sal), as unidades de intervenção devem possuir equipamentos complexos e serem operadas por pessoal altamente capacitado. Isso reflete em altos custos nas diárias cobradas pelas empresas para seu uso (podendo chegar a US$ 500000,00 por dia), conforme gráfico abaixo:

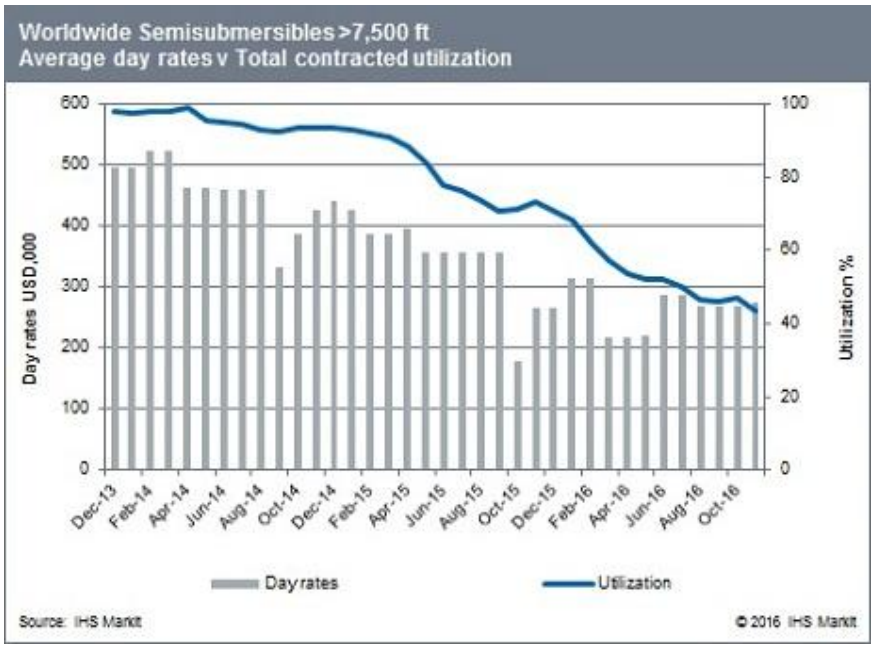

Figura 1: Valor médio de taxa diária de unidades de intervenção no período 2013 a 2016 (IHS, 2016).

Atualmente observa-se um declínio nos custos devido à redução na demanda deste equipamento, causado principalmente pela queda do preço médio do barril de petróleo. Mesmo assim, estes valores ainda causam alto impacto no custo da construção de poços marítimos.

O tempo dispendido na construção de um poço de petróleo pode ser separado de acordo com a Figura 2. Anormalidades operacionais na construção de poços marítimos (também conhecidas na indústria internacional como *NPT – non-productive time*, ou tempo “não produtivo”, ou ainda tempo perdido) trazem significativo impacto tanto na segurança quanto no custo das operações, chegando a 30% do tempo de construção de um poço marítimo (Ashraf, 2016) e (Hsieh, 2010). Ou seja, em um poço construído em 100 dias, 30 dias não correspondem a tempo de construção efetivo de poço, mas a tempo perdido pela ocorrência destas anormalidades.

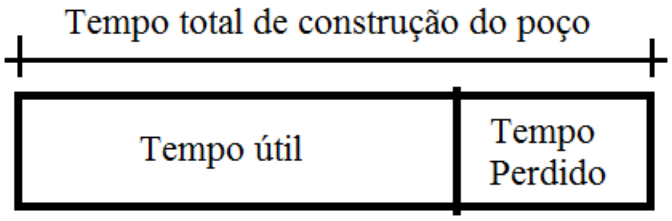

Figura 2: Divisão no tempo de construção do poço.

Este tempo perdido estará sempre presente no processo de construção de poços. Porém, com o intuito de aprimorar as operações, esta parcela de tempo é analisada constantemente pelo corpo gerencial, por meio de relatórios e indicadores. Esta análise possui foco tanto na redução do tempo quanto no aspecto de segurança das operações, que envolvem equipamentos e, principalmente, pessoas, trazendo riscos aos envolvidos no processo. Portanto, as ações provenientes da análise das anormalidades contribuem tanto na redução do tempo de construção do poço quanto, principalmente, na manutenção de altos índices de segurança das operações.

A anormalidade se caracteriza quando há um desvio não planejado durante a construção de um poço. Os motivos são diversos: desde falha relacionada à ação humana e falha de equipamentos, passando por condições meteorológicas adversas, além de imprevisões geológicas (grandes falhas na formação perfurada, influxos rasos e etc.). No caso especial da Petrobras, todos estes desvios são avaliados e categorizados manualmente, por um operador experiente (o engenheiro fiscal), de acordo com um sistema de classificação de anormalidades elaborado internamente à empresa. Em seguida, o mesmo é preenchido em um *software* de relatório de operações de campo (denominado *Open Wells* (Landmark, 2017), da empresa Landmark, subsidiária Halliburton).

Neste sistema, a classificação da anormalidade possui uma estrutura específica, baseada em histórico de ocorrências bem como na experiência dos operadores e especialistas. Esta estrutura é atualizada ocasionalmente, com intuito de aperfeiçoar os registros das anormalidades. As classes cobrem todas as possíveis causas de desvios que venham a ocorrer no processo de construção de um poço.

Em junho de 2011, a estrutura de classificação utilizada na Petrobras foi alterada substancialmente, aumentando de forma significativa o detalhamento das falhas de equipamentos empregados na construção de poços.

Como a análise dos relatórios e indicadores gerenciais é baseada em dados históricos, observou-se uma descontinuidade após esta alteração na estrutura de classificação. Com isso, para não se perder este histórico, propôs-se uma classificação das anormalidades ocorridas antes desta alteração da estrutura de classificação, ou seja, propôs-se realizar uma atualização das classificações segundo as novas classes criadas.

Esta tarefa de atualizar a base de dados demanda muito tempo para ser executada, pois as anormalidades são caracterizadas essencialmente por dados textuais, sendo necessária a leitura de cada uma delas para que seja feita a atualização adequada.

Como consequência desta demanda, também se propôs uma análise das anormalidades sob a nova estrutura de classificação, com o intuito de verificar inconsistências no banco de dados, falhas ou dificuldades de preenchimento dos operadores.

1.1.1 Objetivos

Em função das demandas apontadas na seção anterior, o principal objetivo desta dissertação é o desenvolvimento de um classificador de anormalidades para a Petrobras, com as seguintes características principais:

- Atualizar as anormalidades relativas às falhas de equipamentos, preenchidas até junho de 2011, com base na nova classificação existente. Tal atualização deve possuir um nível mínimo de precisão e deve ser feita da forma mais automatizada possível.
- Demandar o mínimo possível de um operador para cumprir os objetivos, dado que o mesmo é um recurso humano crítico na empresa.
- Avaliar o preenchimento das falhas que já foram apontadas com base na estrutura atualizada (ou seja, preenchidas após 2011), corrigir possíveis apontamentos equivocados e se necessário, propor ajustes no processo de classificação de anormalidades.

Vale destacar que há vários trabalhos desenvolvidos relacionados à mineração textual na área de construção de poços marítimos (para o idioma português), conforme observado em (Miura, 1992), (Miura et al, 2003) e (Guilherme et al, 2006). Porém o objetivo destes trabalhos consistiu em avaliar relatórios de operações realizadas durante a construção de poços (principalmente com foco no tempo útil das operações). Nesse trabalho o intuito é analisar e propor uma solução para a classificação de anormalidades operacionais relacionadas às falhas de equipamentos das unidades de intervenção.

1.2.
Organização da dissertação

Este trabalho está organizado em mais 5 capítulos. Ao final há as referências bibliográficas.

No capítulo 2 são apresentados fundamentos da mineração textual, com foco no processo de categorização de textos. Também são discutidos aspectos de aprendizado de máquina aplicados à categorização textual.

O capítulo 3 aborda a contextualização do problema de categorização de anormalidades no cenário de construção de poços marítimos da Petrobras, bem como a base de dados utilizada para a construção do modelo de classificação.

O capítulo 4 detalha como se desenvolveu o modelo de categorização textual, tratando passo a passo suas etapas: pré-processamento na coleta e tratamento dos dados, transformação dos dados, processamento via aprendizado de máquina e o pós-processamento por meio de métricas.

O capítulo 5 aborda a aplicação do modelo elaborado, através do estudo de casos visando obtenção do classificador com melhor desempenho. Neste capítulo também são destacadas as correções efetuadas na base de dados.

O capítulo 6 apresenta as conclusões finais e sugestões de trabalhos futuros.

2.
Mineração de texto - fundamentos

2.1.
Introdução

A mineração de texto é um termo abrangente que descreve tecnologias e soluções (também chamadas de aplicações) para analisar e processar dados textuais semiestruturados e não estruturados (Miner, 2012).

Para melhor entendimento da característica dos dados textuais (quanto à forma como estão estruturados), tem-se a seguinte taxonomia para os dados:

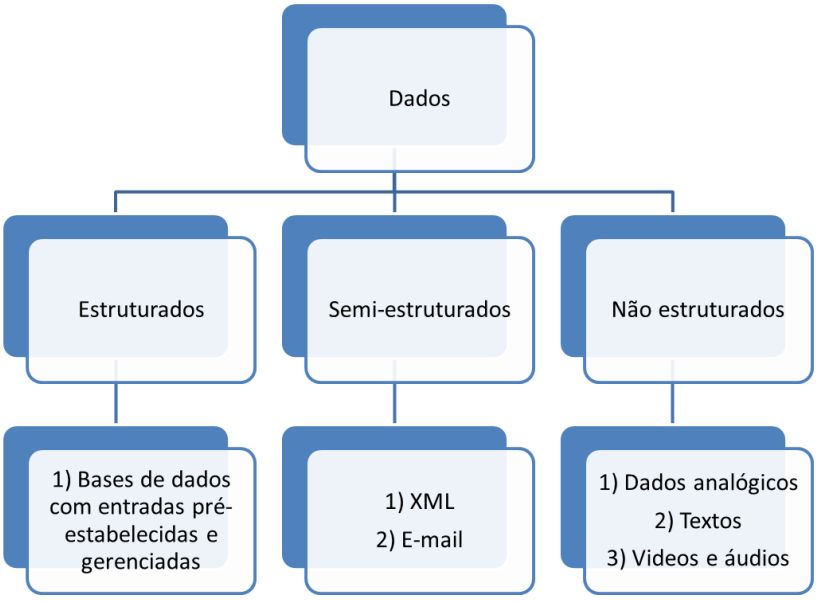

Figura 3: O dado quanto à sua estrutura. Imagem adaptada de (Kumar, 2017).

Em ambiente de dados estruturados há formas rígidas prefixadas para os dados, o que permite a manipulação desses por modelos de mineração de dados, normalmente com um mínimo de pré-processamento. No dado não estruturado, não se conhece antecipadamente a estrutura da informação que deve ser manipulada. Esses podem ser o caso de textos, imagens, etc. Os dados semiestruturados são uma forma de dado estruturado que não está de acordo com

a estrutura formal dos modelos de dados associados com bancos de dados relacionais ou outras formas previamente definidas, mas que contém *tags* ou outros marcadores para separar elementos semânticos e impor hierarquias de registros e campos dentro dos dados. Nesse caso, pode-se citar a linguagem XML, e-mails, etc.

Uma forma simples e eficaz de estruturar a informação textual é apontada em (Miner, 2012): transformar texto em números. A partir dessa transformação, podem-se aplicar diversos algoritmos tradicionais às grandes bases de dados textuais transformadas.

De acordo com (Hearst, 1999), mineração de textos é a descoberta, através de meios computacionais, de informações desconhecidas ou novas, através da utilização de ferramentas de extração automática de informação, a partir de documentos de textos não estruturados.

Esta definição, de maneira análoga à Mineração de Dados, evidencia a busca por extrair informação útil de bases de dados (*knowledge discovery*), através da identificação e aplicação de padrões relevantes para o problema em mãos. Conforme apontado na Figura 3 a grande diferença entre as três modalidades reside em quão estruturados os dados em análise estão. A mineração de dados geralmente parte de uma base de dados com estrutura altamente rígida. Ou seja, as variáveis em análise e que compõe a base de dados estão dispostas em uma estrutura pré-definida pelos responsáveis por esta base. Isto é feito com o intuito de padronizar a informação a ser inserida pelo usuário na base de dados, de modo que o dado possa ser empregado de forma consistente e eficiente (Soares, 2013).

Conforme apontado por (Date, 2005), esses dados estruturados podem ser manipulados e gerenciados por um Sistema Gerenciador de Bancos de Dados (SGBD). Esses sistemas exigem a integridade semântica e de domínio, ou seja, que sejam seguidas regras pré-estabelecidas para o preenchimento da base de dados de modo que não ocorram inconsistências futuras.

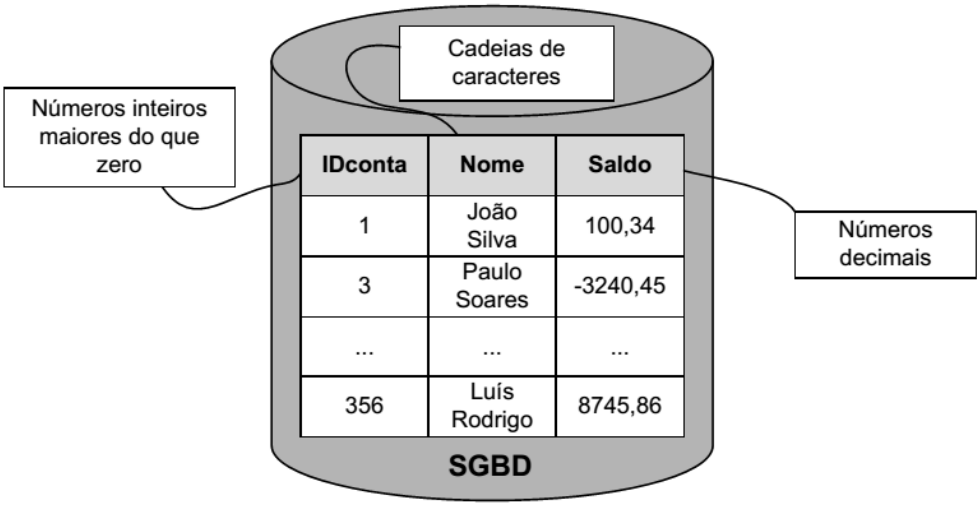

Figura 4: Integridade semântica e de domínio de um SGBD (Soares, 2013).

Pela Figura 4 observa-se que a primeira coluna só suporta números inteiros maiores que zero (“IDconta”). Já na terceira coluna, a variável “Saldo” suporta valores numéricos com duas casas decimais (valor para moeda), mas que podem ser tanto positivos como negativos. Por outro lado, ao se observar a coluna “Nome”, observa-se que a única restrição é que a mesma seja composta por uma cadeia de caracteres (Soares, 2013).

Com isso, nota-se que para a primeira e terceira colunas há atributos classificados como estruturados e com baixa probabilidade de serem inconsistentes, bastando a limpeza e integração de seus dados para promover o processo de mineração de dados (Zhu & Davidson, 2007). Obviamente pode ocorrer preenchimento incorreto, mas como há regras para o preenchimento, estas imprecisões são mais fáceis de rastrear e tratar.

Por outro lado, para uma base de dados não estruturados (caso da coluna 2 da Figura 4) há a necessidade de um pré-processamento complexo (Gomes, 2013), devido à grande imprevisibilidade dos dados inseridos. Pode se observar que a regra de preenchimento restringe a entrada a somente caracteres. Por exemplo, pode-se ter um nome inserido como “Aaaaa da Silva” ou ainda “Jo7ão”. Em termos de cadeia de caracteres estes dados estariam adequadamente preenchidos, mas claramente inconsistentes e com razoável grau de dificuldade de correção.

2.2.
Processo geral de mineração textual

2.2.1. Elementos básicos

A unidade básica de análise do processo de mineração textual é o documento. Este pode ser desde um *tweet* (publicação na rede social *Twitter)*, uma mensagem instantânea entre dispositivos móveis, uma página na internet ou um e-mail, até um trabalho científico complexo. Assim, há documentos de grandes ou pequenas dimensões, em termos de caracteres, além possuírem diferentes graus de complexidade semântica ou complexidade de conteúdo.

Independentemente do tipo de documento avaliado há a necessidade de transformar o dado não estruturado em dado estruturado, com a menor perda de conteúdo possível. De acordo com (Konchady, 2006), há vários níveis possíveis de fracionar o documento, sendo as frações mínimas denominadas grãos. Cada nível possui prós e contras e podem assim ser classificados (Soares, 2013):

- Caracteres: componentes individuais do documento que se agrupam com intuito de formar blocos com um nível semântico maior. Abordagens mais comuns utilizam um número predefinido de caracteres, como o bigrama (dois caracteres) e o trigrama (três caracteres). A dimensionalidade destas abordagens é altíssima, tornando impeditiva a utilização de diversas técnicas.
- Palavras: constituem a menor unidade capaz de representar algum valor semântico. Geralmente é o nível de desmembramento mais utilizado. Porém, conjuntos de palavras podem perder seu valor semântico quando separadas. Com isso, prefere-se a abordagem de termos, ou *tokens* (a ser abordado no próximo item).
- Termos ou *tokens*: nível mais elevado semanticamente, podendo ser representado por termos múltiplos (duas ou mais palavras cujo conjunto tenha sentido), bem como serem representados por uma só palavra reduzida ao seu radical (*stemmin*g) (Gomes, 2013).

Para melhor entendimento desta representação dos dados, segue um exemplo (Soares, 2013):

“**Luis XV vivia no palácio de Versalhes”.**

Tabela 1: Exemplo de representação de um documento por termos.

Palavras	**Termos**
Luís	Luís XV
XV	
vivia	viv- (radical do verbo viver)
no	no
palácio	palácio de Versalhes
De	
Versalhes	

Conforme se verifica no exemplo anterior, os termos (ou *tokens*) são elementos mais abrangentes que as palavras, e acabam por trazer significado a um grupo de palavras, que separadamente poderiam ter significados diferentes do real. Vale destacar, nesse exemplo, que o termo “no” não traz significado relevante e pode ser um candidato à lista de palavras a serem removidas (*stopwords*) (Dias, Malheiros, 2004). Conforme será detalhado na Seção 2.3, a transformação das palavras em termos é de suma importância, tanto no aumento da eficácia do modelo, quanto na redução de tempo de processamento computacional (geralmente proporcional à redução da dimensionalidade).

2.2.2. Etapas

As etapas do macroprocesso de mineração textual podem ser visualizadas de acordo com (Aranha, 2007):

Figura 5: Etapas do processo de mineração textual.

De forma mais abrangente, o processo de mineração textual consiste em primeiramente se criar a base de dados a ser analisada, formando uma coleção de documentos, denominada corpus. Em seguida, manipulam-se internamente os documentos por meio da execução do *stemming* (redução dos termos à raiz das palavras, ou ao seu radical) e remoção de *stopwords* (lista de termos não relevantes a serem removidos).

Na etapa três, há indexação dos termos, onde se efetua a transformação dos dados, sendo este o campo de atuação denominado Recuperação da Informação (RI). Esta é uma área da computação que lida com o armazenamento de documentos, geralmente textuais, e a recuperação automática de informação associada a eles (Baeza-Yates & Bertier, 1999). Nela há elaboração do *thesaurus* (dicionário de sinônimos e de termos compostos), bem como a transformação do texto em número, via matriz de termos por documentos. Finalizada a transformação dos dados não estruturados em estruturados, efetua-se o processamento da base disponível, executando a mineração de dados em si (processamento dos dados). Cabe destacar que, a depender do escopo da aplicação, este processo pode ser revisto e ter suas etapas mais detalhadas e aprimoradas, bem como suprimidas. Finalmente na análise (ou pós-processamento), avalia-se o conhecimento descoberto pela aplicação do modelo.

Conforme destacado em (Rezende, 2005), é importante estabelecer um objetivo do processo de mineração de dados, que basicamente consiste na realização de uma tarefa. Este objetivo definirá a abordagem a ser empregada, bem como o tratamento, transformação e processamento dos documentos.

2.2.3.
Áreas de aplicação

As áreas de aplicação de mineração textual são as mais diversas possíveis. O fluxograma da figura, extraído de (Miner, 2012) traz uma abordagem prática das principais áreas, baseada justamente na tarefa a ser desempenhada pelo modelo:

Figura 6: Fluxograma de aplicações da mineração textual.

As caixas em azul referem-se às áreas disponíveis em mineração textual. Elas podem ser resumidas conforme (Miner, 2012):

- **Busca e recuperação de informação (*Information Retrieval - IR*):** armazenar e recuperar documentos, incluindo *search engines* (buscadores) como Google, Yahoo e Bing, bem como buscar por palavras chave.
- **Agrupamento de documentos (*Clustering*):** agrupamento de documentos via técnicas de agrupamento de mineração de dados.
- **Categorização de documentos:** agrupamento e categorização de documentos, utilizando as ferramentas de classificação disponíveis na

mineração de dados usual, baseando-se em modelos treinados a partir de dados rotulados.

- ***Web Mining*:** mineração de dados e textos na Internet, com foco específico na escala e interconectividade da rede (*web*).
- **Extração de informação (*Information Extraction - IE*):** identificação e extração, tanto de fatos relevantes quanto de conexões, em textos não estruturados.
- **Processamento de linguagem natural (*Natural Language Processing - NPL*):** processamento de linguagem de “baixo nível”, ou seja, análise sintática dos textos.
- **Extração de conceitos:** agrupamento de palavras/frases em grupos com mesmo significado (similaridade semântica).

Analisando o fluxograma apresentado acima, observa-se que para o tema específico desta Dissertação (classificação de anormalidades relacionadas a falhas de equipamentos em construção de poços) há interesse em resultados relacionados a documentos que já possuem categorias pré-selecionadas (11 para este trabalho, conforme será detalhado no Capítulo 3), e tais documentos são considerados sem conexões ou *hyperlinks*. Isso leva à aplicação de categorização textual, que será detalhada na próxima seção.

2.3. Etapas da categorização textual

Conforme apontado na seção anterior, o processo de mineração textual possui as seguintes etapas (ver Figura 5):

- Coleta dos documentos;
- Pré-processamento;
- Indexação;
- Mineração (ou processamento);
- Análise (ou pós-processamento).

O detalhamento de cada uma dessas etapas é fornecido nas subseções a seguir.

2.3.1.
Coleta de documentos

Os documentos a serem coletados (e que irão compor a base de dados) devem ser relevantes para o cenário de aplicação da tarefa (Soares, 2013).

2.3.2.
Pré-processamento

A etapa que pode ser considerada a mais trabalhosa na análise de mineração textual é a preparação dos dados. Ela é composta pela elaboração do dicionário de termos múltiplos, *case folding*, criação da lista de elementos não desejáveis e finalmente a radicalização dos termos. O intuito é reduzir a grande quantidade de termos presente na coleção de documentos.

2.3.2.1.
Dicionário de termos múltiplos (multi-termos)

A ideia deste dicionário consiste em concatenar termos que, separados, possuiriam outro sentido, ou sentido algum. Por exemplo: o termo “casa da moeda” possui sentido diferente de seus termos separadamente (“casa”, “da”, “moeda”). Em uma análise que desconsidera este dicionário, a palavra “da” seria retirada (preposições geralmente são termos que agregam pouco significado), e as demais palavras seriam tratadas de forma separada. De certa forma, este dicionário auxilia a recuperar o significado de termos mais elaborados.

2.3.2.2.
Case folding

Consiste em reduzir todas as letras para minúsculas ou maiúsculas, descaracterizando nomes próprios (Lopes, 2004).

2.3.2.3.
Stop words

Baseia-se na remoção de palavras e termos que não agregam valor ao processo (denominadas *stop words*), como preposições, conjunções, símbolos, números e pontuação dos documentos (Dias, Malheiros, 2004). A seleção de *stop words* é uma tarefa altamente dependente do banco de dados e de seu conteúdo (Makrehchi & Kamel, 2008). Algumas ferramentas de mineração textual já oferecem listas de *stop words*. Entretanto, é comum o uso do conhecimento de um especialista para acrescentar à lista inicial palavras menos relevantes para a tarefa específica.

2.3.2.4.
Radicalização

Nesta etapa efetua-se a normalização morfológica dos termos. A ideia consiste em reduzir os termos ao seu radical (*stemming*), removendo os sufixos (Gomes, 2013). Com isso, palavras como “carro” e “carros”, se reduzem a “carr-” somente. O mesmo se aplica às variações verbais: “escrevo”, “escreveis”, “escrevem”, que resultam no termo “escrev-”. Tal tarefa reduz consideravelmente o número de termos de uma matriz de termos por documentos.

O revés da radicalização reside no fato de que palavras com significados diferentes podem apresentar o mesmo radical, como em “caminhão” e “caminhar”: “caminh-“. Isso gera uma perda na informação disponível inicialmente.

A radicalização é uma tarefa bem específica, e é desempenhada por algoritmos (denominados *stemmers*). Conforme apontado em (Gomes, 2013), dos algoritmos de radicalização para língua portuguesa destacam-se:

- Algoritmo de Porter: desenvolvido em 1980, é uma adaptação de um algoritmo desenvolvido para língua inglesa (Dias, Malheiros, 2004). Este apresenta cinco etapas, com uma tabela sendo aplicada a cada uma dessas etapas para remoção de sufixos (Paice, 1983).

- Algoritmo RSLP: desenvolvido em 2001, o RLSP é um acrônimo para Removedor de Sufixos para Língua Portuguesa, publicado como *StemmerPortuguese* (Orengo, Huyck, 2001). O mesmo foi elaborado com o intuito de aprimorar o processo de radicalização utilizando oito passos, cada um com um conjunto específico de regras.

Em Flores (2009) são apontados testes que constataram que o algoritmo RSLP tende a apresentar uma taxa de acerto menor que o de Porter, fato também relatado em (Xavier, Silva, Gomes, 2013), além de exigir um maior esforço computacional (por ser mais complexo). No entanto, o RSLP é considerado um algoritmo mais completo que o de Porter, tendendo a acertar casos mais específicos, porém menos frequentes.

2.3.3. Indexação textual

Após este tratamento inicial dado ao banco de dados textual, inicia-se a etapa de indexação textual. Este processo foi iniciado e bastante desenvolvido pela área de sistemas de Recuperação da Informação (RI), que surgiu na segunda metade do século XX (Rijsbergen, 1979).

Vale destacar que o processo deste trabalho é aplicado à categorização textual, porém é necessário o uso da aplicação de RI, de forma a prover um alicerce para que o processo de categorização textual possa ser construído. As ferramentas de RI utilizadas na categorização se restringem às técnicas de representação e identificação (por meio de índices) de documentos. A Figura 7 ilustra o processo de RI como um todo (Soares, 2013):

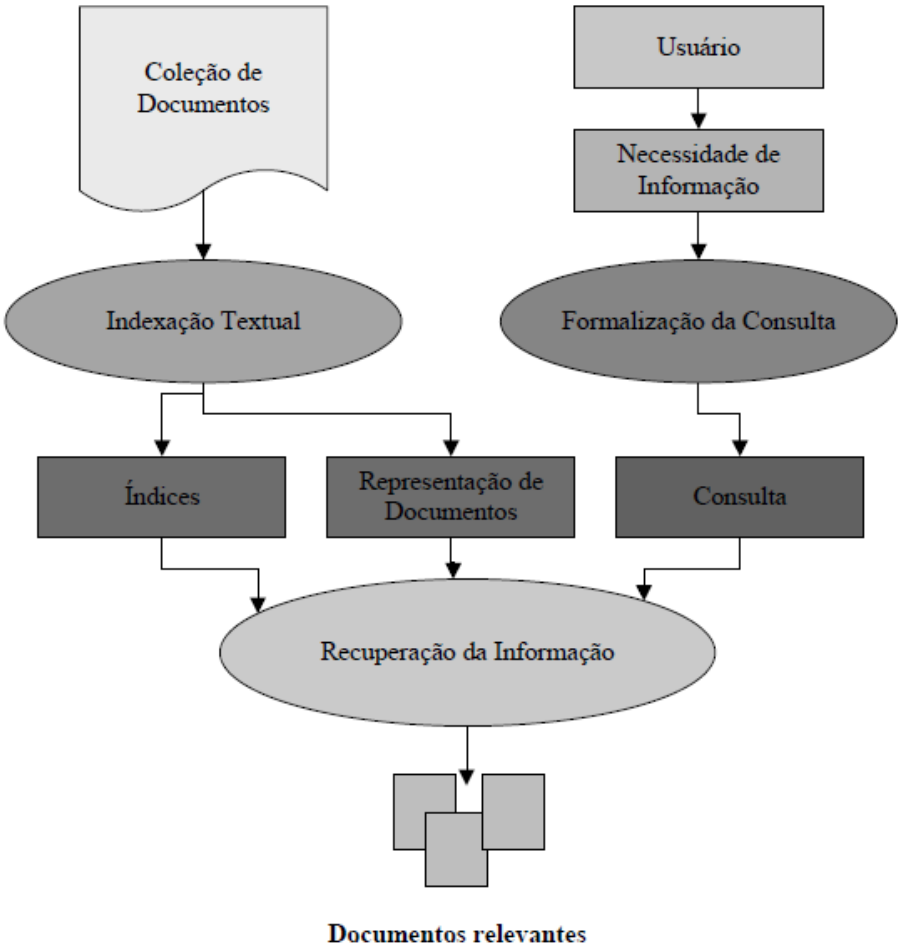

Figura 7: Fluxograma do processo de RI.

No processo de RI uma coleção de documentos sofre uma indexação textual, que irá gerar os índices e também a representação dos documentos (lado esquerdo do fluxograma). Um usuário demanda uma informação específica, que se formaliza em uma consulta. A conjunção da localização do documento (índice) com o seu conteúdo, será confrontada com a consulta, de modo a se atender a necessidade do usuário (objetivo final da RI).

Conforme já mencionado, a categorização textual se utiliza da indexação textual (lado esquerdo do fluxograma) para localizar e representar os documentos. Porém, diferente do processo RI, não há o usuário com uma demanda específica, mas sim uma tarefa mais ampla, que consiste em separar (categorizar) os documentos da coleção.

2.3.3.1.
Criação de índices

A primeira etapa de criação dos índices se justifica pelo fato de que, para que se obtivesse a consulta desejada, se fazia necessário percorrer toda a base de dados, documento a documento, com alto esforço computacional (Soares, 2013). Em uma analogia, é como se consultar um assunto específico em um livro que não possui um índice. Seria necessário procurar o assunto em todo livro, até sua localização.

No intuito de aprimorar esse processo de busca, foi desenvolvida uma indexação dos assuntos, que como em um índice remissivo de um livro, traz atalhos para o acesso às informações requisitadas. Com estes atalhos, bastaria o modelo acessar as informações por este índice de modo a tornar mais rápida sua busca pelo dado ou informação necessária (Maming, Raghavan, Schutze, 2007).

De acordo com (Soares, 2013), a forma mais indicada de indexação consiste em efetuar um tratamento prévio dos dados (pré-processamento), com o intuito de aperfeiçoar o processo de indexação. Há outra forma menos eficiente que seria de utilizar o dado bruto, sem tratamento prévio. O pré-processamento geralmente é feito como abordado na Seção 2.3.2: uso do dicionário de múltiplos termos, case folding, remoção de palavras pouco relevantes, bem como a radicalização dos termos.

Após efetuar o pré-processamento, dar-se-á início à indexação temática, que faz o uso de um dicionário externo de termos, denominado *thesaurus*, conforme será descrito abaixo.

2.3.3.2.
Dicionário de sinônimos (*thesaurus*)

Este processo simplifica o apontamento do termo no respectivo documento, pois estabelece uma correlação prévia por meio de um dicionário de sinônimos. O dicionário possui um termo preferido e os demais termos não preferidos, que se relacionam ao primeiro. Todos os sinônimos do termo preferido

são apontados pelo termo preferido. A Figura 8 mostra um exemplo de como ocorre este apontamento (Soares, 2013):

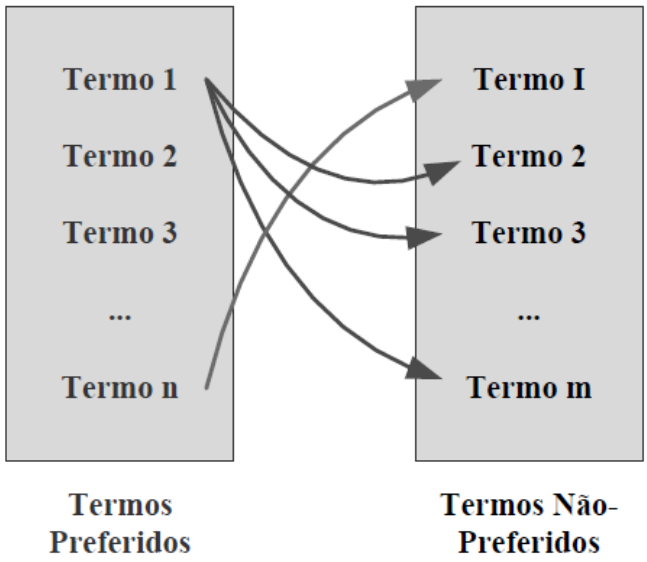

Figura 8: Esquema simplificado de um *thesaurus*.

É importante destacar que este dicionário é altamente dependente do assunto tratado pelos documentos. Tal fato é ainda mais acentuado neste trabalho, que possui um conjunto de termos técnicos altamente variados e especializados.

Prosseguindo a etapa de indexação, dado que os documentos já possuem os devidos apontamentos, será abordada a forma como estes documentos podem ser representados para posterior mineração dos dados.

2.3.3.3. Modelo de espaço vetorial

A maneira mais usual de efetuar a representação de documentos textuais se dá pelo modelo de espaço vetorial (Silva, 2007). A ideia é transformar o documento em um vetor, onde as coordenadas representam os termos presentes no documento.

Toda a coleção de documentos é representada pelo uso da matriz de termos por documentos - MTD (em inglês: *term-by-document matrix - TDM*). Conforme Figura 9, para a representação da magnitude de cada variável (no caso cada termo

ou *token*), foi utilizado o método de frequência do termo no referido documento. Há diversas formas de fazer esta contabilização, a serem tratadas mais adiante. A ideia do exemplo é somente ilustrar de forma simples a MTD.

COLEÇÃO (CORPUS)	
Documento 1	deposite o dinheiro e o cheque no banco
Documento 2	o barco está preso no banco de areia
Documento 3	sentei no banco da praça
Documento 4	foi para o banco dos réus

MTD: MATRIZ DE TERMOS POR DOCUMENTOS

	depositar	o	dinheiro	e	cheque	no	banco	barco	estar	preso	de	areia	sentar	praça	ir	para	dos	réus
d1	1	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
d2	0	1	0	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0
d3	0	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0
d4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1

Figura 9: Exemplo de criação de uma MTD.

Retornando à representação via modelo de espaço vetorial, observa-se que ao assumir a forte premissa de transformar a coleção de textos em uma matriz, com a mínima perda de informação, têm-se as seguintes consequências (Miner, 2012):

1- “*Bag of words*”, ou bolsa de palavras:

- Desconsidera-se o significado em meio ao contexto (semântica), ou seja, os termos são desvinculados e a sua ordem no documento é desconsiderada. Entretanto, conforme apontado em (Lopes, 2004), é difícil obter uma informação exata das relações semânticas apenas a partir de informações textuais, de maneira automática.
- Palavras possuirão somente um significado. Como exemplo, a palavra “banco” pode ser tanto o móvel banco (assento) como também pode ser relativo à instituição financeira, ou ainda a um acúmulo de areia em um trecho de um rio (banco de areia). Uma forma de mitigar tal problema reside na elaboração de um dicionário de termos compostos (quando aplicável).
- Demanda a escolha de uma métrica para quantificar o “valor” dos termos na MTD. No exemplo acima se utilizou a frequência dos termos no documento. Estas métricas serão detalhadas mais adiante e impactam o processo de mineração textual.

2- Alta dimensionalidade:

- Matriz esparsa (com muitos “zeros”), pois a maioria das palavras não consta em todos os documentos. Isso acarreta a uma exigência maior de processamento computacional.
- Demanda um critério para seleção/redução de variáveis (termos, no caso), para mitigar este efeito. Alguns métodos serão tratados na próxima seção.

De forma complementar, mesmo com a relativa perda observada de informação com uso da MTD, para o caso em que se deseja agrupar e/ou classificar documentos desconhecidos, o impacto é pouco sentido (Miner, 2012).

Antes de finalizar esta etapa, onde se utilizaram as ferramentas de recuperação da informação para atribuir índices e para representar de forma estruturada os documentos, se faz necessária a remoção de algumas características que não agregam valor, de modo a selecionar somente os atributos mais relevantes, conforme apontado na Figura 10, extraída de (Soares, 2013):

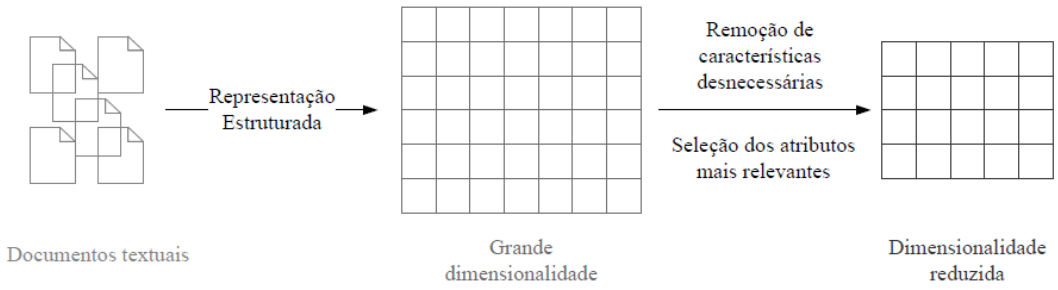

Figura 10: Processo de representação estruturada de uma coleção de textos.

A ideia da redução de dimensionalidade é tornar o processamento menos oneroso, dada à particularidade da alta dimensionalidade acarretada por esta abordagem representativa. Esta redução de termos já foi em partes contemplada no processo de remoção de termos (*stop words*). O uso do dicionário multi-termos, *case folding*, *stemming* e o *thesaurus* também contribuem neste sentido. Porém, mesmo após todo este tratamento, a grande dimensionalidade ainda pode diminuir a efetividade do modelo de categorização textual, tanto reduzindo o

desempenho para execução da tarefa, quanto em termos de aumento do custo computacional para executar tal tarefa.

2.3.3.4.
Redução de termos por relevância

Conforme mencionado, após todo o tratamento visando redução de variáveis com foco nas características internas aos documentos, executa-se a redução da quantidade de termos avaliando sua relevância em toda a coleção de documentos.

Uma proposta se baseia na lei de Zipf (Zipf, 1949) e no corte de Luhn (Luhn, 1958). A lei de Zipf pontua que se ***f*** é a frequência de ocorrência de qualquer termo do texto e ***r*** é a posição da ordenação (*rank*) com relação aos outros termos, então o produto ***f x r*** é aproximadamente constante. Luhn propôs que em um gráfico ***f*** por ***r***, pode-se definir um limite superior e um limite inferior de corte. As palavras que estiverem fora do intervalo são excluídas da análise, conforme a Figura 11:

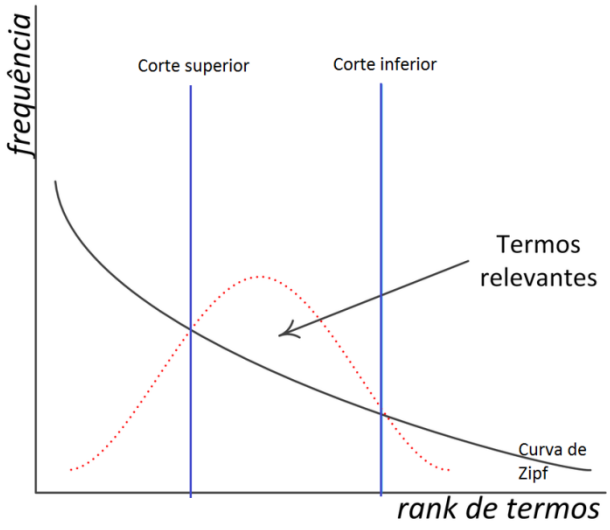

Figura 11: Lei de Zipf e o corte de Luhn. Adaptado de (Soares, Moura, 2015).

Uma forma possível de aplicar este corte inferior (retirando termos menos representativos) é efetuar a redução de termos esparsos conforme consta no pacote

em linguagem R, para mineração textual, desenvolvido por (Feinerer, Hornik, 2015). Esta ferramenta parte do princípio de que quanto menos presente um termo é em toda a coleção, menos relevante este seria para o êxito do processo de classificação. Pode-se definir a esparcidade de um termo como o percentual de documentos onde não há presença do termo.

Considere uma coleção de cinco documentos, que possuam quatro termos (A, B, C e D) mapeados, representados na MTD da Figura 12:

		termos			
	documento	A	B	C	D
	1	2	1	1	3
	2	0	0	1	4
	3	0	3	1	1
	4	0	0	0	2
	5	0	0	0	1
Núm. De zeros		4	3	2	0
Esparcidade do termo		80%	60%	40%	0%

Figura 12: Exemplo do cálculo de esparcidade dos termos de uma matriz de termos por documentos.

Se for arbitrado o corte inferior dos termos menos representativos em, por exemplo, 75%, o termo A seria retirado da análise. Ou seja, termos com esparcidade maior que 75% seriam excluídos por corresponderem a termos raros e que agregariam pouco ao processo. Este corte deve ser feito com cautela, pois pode haver termos importantes para determinar certa classe, mas que podem ser raros para os demais documentos da coleção. A ideia é excluir termos raríssimos para melhorar o processamento da tarefa.

2.3.3.5. Métrica definidora de importância

Após extensa redução dos termos da MTD, na etapa de indexação há questão relativa ao peso que será dado aos termos, ou seja: como quantificar a presença dos termos nos documentos utilizando a MTD? Há, segundo (Gomes,

2013), várias maneiras de atribuir os pesos aos termos, que serão discutidas nas próximas quatro seções.

2.3.3.5.1.
TF – Term Frequency

No acrônimo em inglês TF para frequência dos termos (*term frequency*), o peso é atribuído simplesmente como a contagem (frequência) do referido termo em um documento. Com isso, na matriz de termos por documentos, supondo as linhas serem os documentos e as colunas serem os termos, tem-se para cada documento, uma célula correspondente a quantas vezes cada termo foi contabilizado no referido documento.

2.3.3.5.2.
Booleano

Baseado no operador booleano, este tipo de peso se propõe a somente atestar que o termo está presente em um documento, sem levar em consideração sua frequência dentro do documento. Ou seja, se estiver presente será atribuído o peso 1, caso contrário será utilizado o peso 0.

2.3.3.5.3.
IDF

Outros pesos mais elaborados surgiram, sendo um deles a contagem do inverso da frequência dos termos nos documentos (*inverse document frequency*).

Diferente dos demais pesos, que possuem seu valor dependente de um termo e de um documento específicos, este peso é uma característica somente do termo (sem relação com um documento específico). Ela é utilizada para medir se

o termo é comum na coleção de documentos ou se é raro. Sua formulação é dada pela Equação 1:

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Equação 1

Onde t é o termo considerado, d é um documento específico, D é toda a coleção de documentos e N é a frequência do termo em toda a coleção.

2.3.3.5.4. TF-IDF

Esta modalidade de peso refere-se a uma fusão da modalidade TF com a modalidade IDF (*term frequency - inverse document frequency*). É apontada como uma medida de peso global, pois considera a frequência do termo com relação a um documento específico, porém ponderado pela presença deste mesmo termo em toda a coleção. É calculado de acordo com a Equação 2:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Equação 2

Onde t é o termo considerado, d é um documento específico e D é toda a coleção de documentos.

Se um peso alto de TF-IDF é obtido, isso implica em uma frequência alta em certo documento (alto TF), porém uma baixa frequência do mesmo nos demais documentos da coleção (alto IDF). Portanto este peso tende a ser baixo para termos comuns e que são citados em vários documentos.

2.3.3.6.
Normalização dos dados

Conforme apontado por (Hsu, Chang, Lin, 2010), o impacto da normalização dos dados em algoritmos de aprendizado de máquina deve ser levado em consideração. Primeiramente, a ideia em normalizar os dados consiste distribuí-los de forma mais adequada para o algoritmo que irá processar o dado. Com isso, minimizam-se os problemas oriundos do uso de unidades e amplitudes distintas entre as variáveis.

Em (Hsu, Chang, Lin, 2010) há a sugestão pelo método de normalização por extremos. Consiste em submeter o valor da variável X à expressão da Equação 3:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Equação 3

Para o caso da mineração textual, X seria o valor do peso obtido de um termo para um dado documento. X_{min} seria o menor valor em toda a matriz de termos por documentos e X_{max} o valor máximo. Finalmente X_{norm} seria o valor do peso já normalizado.

Por meio dos valores máximos e mínimos de todas as variáveis no banco de dados, se obtém o valor da mesma já normalizado em relação a estes extremos.

2.3.4.
Mineração dos dados (processamento)

Nesta etapa, após possuir uma matriz tratada de termos por documentos, utilizam-se as ferramentas usuais de mineração de dados para realizar a tarefa de categorização de textos aplicada à área de aprendizado de máquina.

2.3.4.1.
Aprendizado de máquina aplicado à categorização

O problema de categorização de documentos no âmbito do aprendizado de máquina consiste em um processo que gera um classificador que irá separar os documentos em diversas classes de acordo com suas características representativas. Considerando a existência de documentos já previamente categorizados por um operador, o processo de aprendizado é regido pelos exemplos fornecidos para seu treinamento, confirmando sua efetividade através de validação e teste do algoritmo.

Esta etapa de treinamento, validação e teste possui várias abordagens disponíveis. De acordo com (Soares, 2013), as abordagens mais utilizadas são:

- *Holdout*: consiste em separar de antemão um conjunto de teste e outro de treinamento (geralmente 1/3 da base para teste). Este método recebe críticas, pois a separação não permite utilizar de forma completa todo o conjunto de observações. Ou seja, o classificador poderia ser favorecido pela base de treinamento e teste escolhida, pela própria aleatoriedade dos dados. Com isso é alta a probabilidade de ter-se uma impressão errônea do desempenho do modelo para a tarefa a qual foi concebido.
- Validação cruzada (ou *K-fold Cross Validation*): a base de dados é dividida em K conjuntos. Em seguida, separa-se um destes conjuntos para validação e os demais para treinamento do modelo. Ou seja, o classificador é treinado com (K-1) frações restantes e é efetuado um teste específico (validação), com a fração restante (fração de validação). Isto é feito por K iterações, com intuito de reduzir a influência da aleatoriedade dos dados sobre o processo.

Para cada iteração, uma nova fração do conjunto de treinamento é separada para atuar como fração de validação e as demais K-1 frações são utilizadas para treinar o classificador. Atentar que a fração utilizada para validação nunca pode se repetir nas iterações. A Figura 13 ilustra o processo para K = 10:

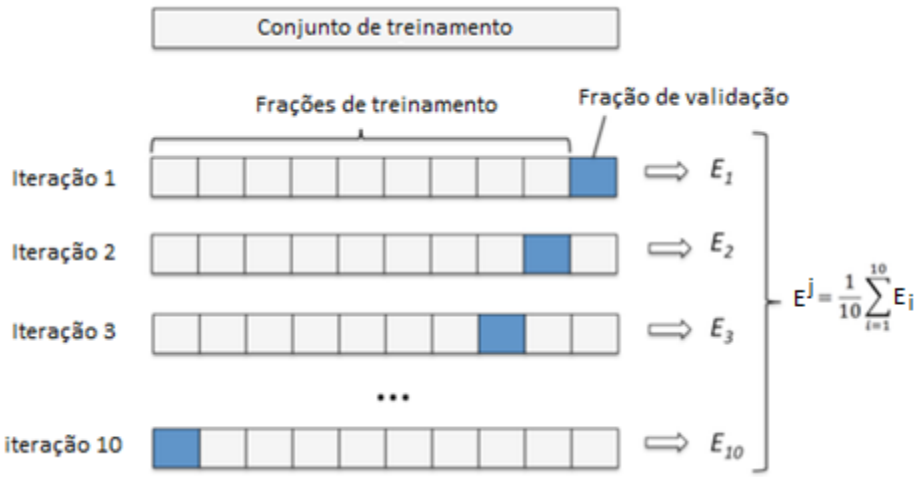

Figura 13: Validação cruzada (*K-fold Cross Validation*). Adaptado de (Raschka, 2017).

Cada iteração resultará em um desempenho associado E_k. Ao final de todas as iterações, o modelo 1 é avaliado efetuando a média do desempenho de todas estas iterações, obtendo o valor E^j (j=1).

Vale destacar que a validação cruzada toma considerável tempo computacional, mas acaba por trazer resultados mais robustos.

2.3.4.2. Classificadores de texto

2.3.4.2.1. Algoritmos existentes

O processo de mineração dos dados textuais estruturados para classificação de documentos pode ser desempenhado pelos seguintes classificadores, conforme (Gomes, 2013):

- Bayesiano
- K-vizinhos mais próximos (*Nearest Neighbors*)
- Árvore de decisão
- Regra de decisão
- Regressão
- Rocchio
- Probabilísticos

- Redes neurais
- *Deep Learning*
- SVM (*Support Vector Machines*)

Para cenários de classificação de documentos por meio do método “*bag of words*”, devido à sua alta dimensionalidade, muitos algoritmos tornam sua aplicação inviável devido ao alto tempo de processamento requerido. (Joachims, 2005) faz uma comparação entre os vários algoritmos e os SVMs e evidencia o desempenho superior deste algoritmo neste cenário de categorização de textos. Isso também é apontado em (Hsu, Chang, Lin, 2010). No trabalho de (Zang, Tang, Yoshida, 2007) comparam-se os SVMs com redes neurais (*back propagation*) para vários casos e também, no cenário de categorização textual, o SVM prevalece como algoritmo mais robusto para a análise.

Finalmente, conforme apontado em (Aggawar, Zhai, 2012), para a categorização textual foi evidenciado, após uma análise de vários classificadores de texto disponíveis, que o classificador SVM desempenha acima da média para os vários cenários analisados.

Portanto para este trabalho decidiu-se utilizar o algoritmo SVM para efetuar a tarefa de classificação textual. Na próxima seção será abordado o algoritmo SVM.

2.3.4.2.2. SVM

Consistindo em uma técnica de aprendizado de máquina supervisionado, os *Support Vector Machines* (SVM) ou máquinas de vetores de suporte tiveram sua aplicação inicialmente para problemas de classificação. Posteriormente os SVMs foram aplicados para problemas de regressão (Vapnik, Golowich, Smola, 1997). Com isso, (Gunn, 1998) aponta que para evitar conflito de nomenclatura, quando aplicados os SVMs a problemas de classificação, sua atribuição seria SVC (*Support Vector Classification*) e quando aplicados a problemas de regressão, o mesmo seria considerado como SVR (*Support Vector Regression*).

As bases do SVC foram lançadas por (Cortes, Vapnik, 1995) e a ideia principal de um SVC é construir um hiperplano como superfície de separação (ou decisão), de tal forma que dadas duas classes e um conjunto de pontos destas referidas classes, este hiperplano separa o maior número de pontos possíveis (cada classe de um lado do hiperplano) e ao mesmo tempo maximiza a distância de cada classe a esse hiperplano (esta máxima distância é conhecida como margem máxima).

A distância dos pontos mais próximos de uma classe até o hiperplano de separação é denominada margem de separação. A Figura 14 traz um exemplo de hiperplanos de separação a serem avaliados pelo SVC. As distâncias dos pontos mais próximos aos hiperplanos são denominadas d_1 e d_2 (respectivamente para o hiperplano 1 e 2). O hiperplano 1 é o que possui a maior distância entre os pontos mais próximos de ambas as classes e, portanto, distância d1 pode ser considerada a margem máxima de separação (quando comparada com a distância d_2 do hiperplano 2). Para gerar o hiperplano com margem máxima são necessários alguns pontos de ambas as classes (os mais próximos deste hiperplano). Estes pontos são denominados vetores de suporte. O hiperplano 1 seria o escolhido pelo SVC (distância máxima das classes ao plano) em detrimento ao plano 2, conforme a Figura 14:

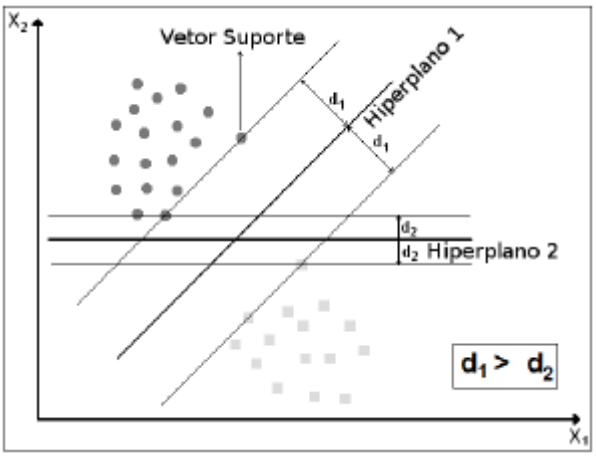

Figura 14: Hiperplanos de separação (SVC). O hiperplano 1 possui margem máxima.

Analisando a ferramenta com enfoque na teoria do aprendizado estatístico, a proposta do SVC se baseia no princípio de Minimização do Risco Estrutural (SRM - *Structural Risk Minimization*). Este, quando comparado à abordagem de

Minimização do Risco Empírico (ERM - *Empirical Risk Minimization*) apresenta desempenho de generalização (ou desempenho na etapa de teste) bem robusto quando comparado aos demais algoritmos, para determinadas aplicações (Gunn, 1998). A ERM é a abordagem utilizada, por exemplo, para as redes neurais *back-propagation*.

É notório que no processo do aprendizado de máquina há as etapas de treinamento e de teste. Neste cenário, (Vapnik, 1999) propôs a dimensão VC, ou Vapnik-Chervonenkis, que apesar do nome dimensão não possui de fato o carácter geométrico de dimensão, mas sim uma medida de magnitude ou poder de atuação de um conjunto de funções. O princípio de SRM se baseia no fato de que, após treinamento e teste de um algoritmo, a taxa de erro de generalização (ou erro de teste) se limita à soma da taxa de erro de treinamento por um termo que é dependente da capacidade deste conjunto de funções (Chaves, 2006). O conceito de capacidade mais utilizado é a dimensão VC e pode ser visualizado na Figura 15. São apresentados dois casos de dimensão VC: um conjunto de três retas e uma elipse. Cada caso seria um conjunto específico de funções, os quais possuiriam uma dimensão VC:

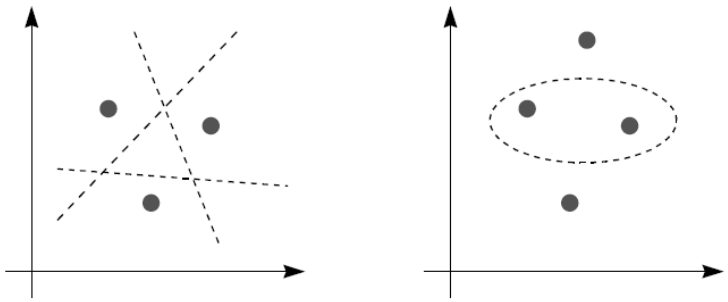

Figura 15: Diferentes conjuntos de funções que podem ser avaliadas pela dimensão VC quanto à sua competência em separar classes (Gunn, 1998).

Se os pontos de um problema binário (duas classes) são separáveis, ou seja, se o problema for linearmente separável, o SVC consegue diminuir sensivelmente a taxa de erro de treinamento bem como minimizar a dimensão VC. Isso implica em minimizar a complexidade dos hiperplanos (reduzindo sua capacidade) e por consequência se obter um bom desempenho na fase de teste dentro dos cenários de classificação (baixa taxa de erro de generalização)

(Chaves, 2006). A Figura 16 ilustra o impacto da dimensão VC (denominada na figura como *capacity term*) sobre as taxas de erro do modelo.

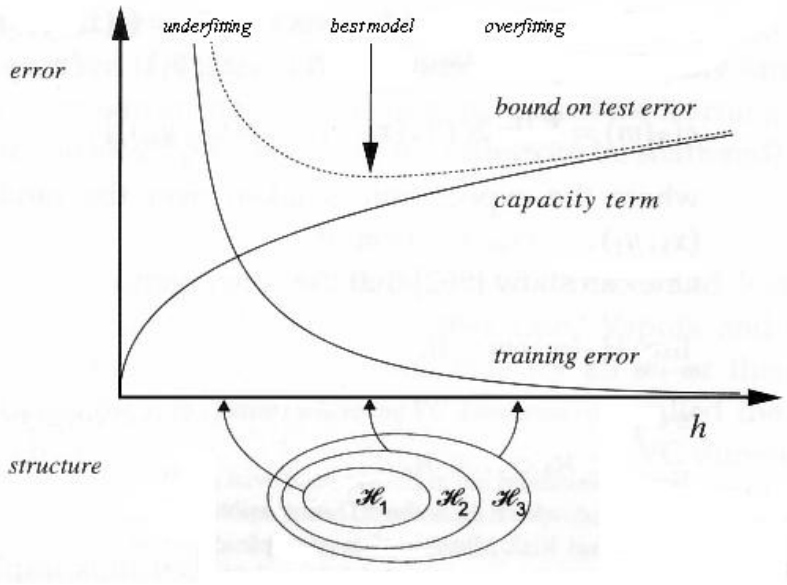

Figura 16: Exemplo de SRM (Sewell, 2008).

Como pode se observar, nem sempre diminuir o erro de treinamento acarreta em melhorar o desempenho na fase de teste. Ou seja, quanto mais se aumenta a dimensão VC maior o efeito de *overfitting* do modelo. Portanto o SVC alcança um ponto ótimo de modo a manter o bom desempenho de teste sem aumentar a dimensão VC consideravelmente, diminuindo este efeito de *overfitting*.

A melhor relação custo benefício da dimensão VC, que depende do risco empírico e da capacidade da classe de funções especificada, leva ao princípio de minimização do risco estrutural (SRM) (Almeida, Braga, 2001), que está intimamente relacionado à topologia do SVC. O SVC calcula um conjunto de hiperplanos e por meio do SRM identifica a separação ótima, de modo a maximizar a distância (margem) dos exemplos mais próximos. Com isso se minimiza a dimensão VC, bem como se minimiza o erro de teste.

Até então a abordagem do SVC esteve restrita a problemas linearmente separáveis. Porém há problemas mais complexos, não separáveis linearmente, que podem ser resolvidos por dois meios (Gunn, 1998): o primeiro foi proposto por (Cortes, Vapnik, 1995), consistindo na introdução de variáveis não negativas

denominadas variáveis de suavização (ou variáveis soltas – *slack variable*) em conjunto com o uso de funções de penalização. Com isso uma gama de problemas não linearmente separáveis pôde ser tratada. O segundo modo de atuar nos problemas não separáveis linearmente é por meio de funções *Kernel* (será abordado mais adiante).

Basicamente, a variável solta faz com que o classificador tolere alguns pontos (considerados a princípio como erros) da classe azul dentro da classe cinza, e vice-versa, conforme a Figura 17:

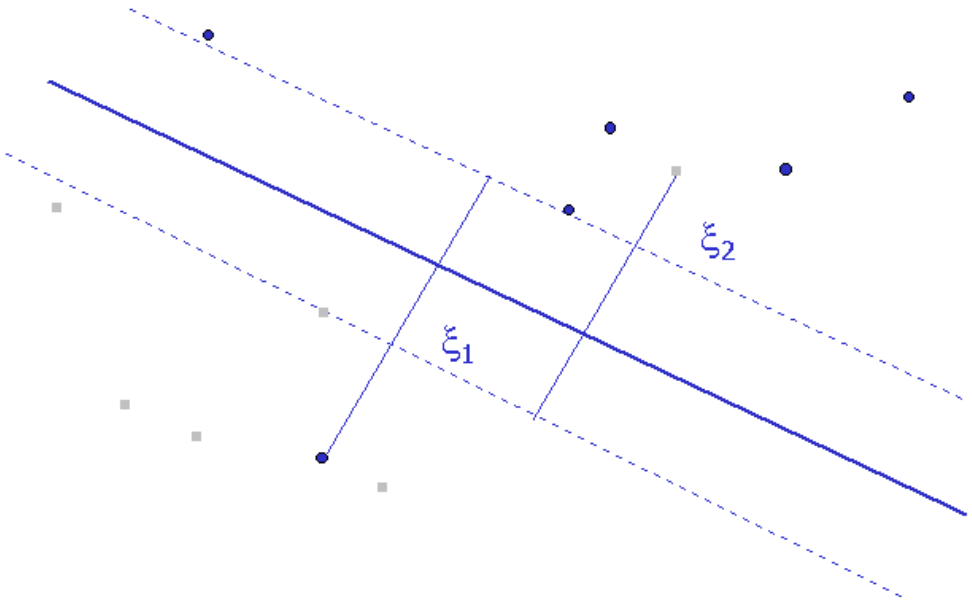

Figura 17: Exemplo de 2 variáveis soltas (Chaves, 2006).

Ao inserir as variáveis soltas no equacionamento, há de se considerar a função de penalização, que depende do parâmetro C, também conhecido como constante de regularização. Esta possui grande influência no compromisso entre a complexidade do modelo e o número de pontos não separáveis. Ou seja, o aumento no valor de C aumenta a taxa de penalização, de modo que se diminui a tolerância pelos pontos cobertos pela variável de suavização (apontados na Figura 17).

Por outro lado, quanto menor o valor de C, mais “suavizado” estaria o modelo e com isso mais pontos seriam desconsiderados como erros. Este fato

tende a aumentar o fenômeno de *overfitting*, diminuindo a generalização do modelo.

Com o intuito de encontrar o valor ideal para a constante de regularização (C), são utilizadas heurísticas como a proposta por (Hsu, Chang, Lin, 2010), de modo a se obter mínimo erro de teste. Na Figura 18 há um exemplo de heurística, onde cada curva é um SVC com suas respectivas constantes de regularização alteradas. O propósito desta heurística é de alterar os valores da constante e analisar o comportamento do desempenho do modelo, de forma a se obter os pontos de inflexão das curvas de desempenho:

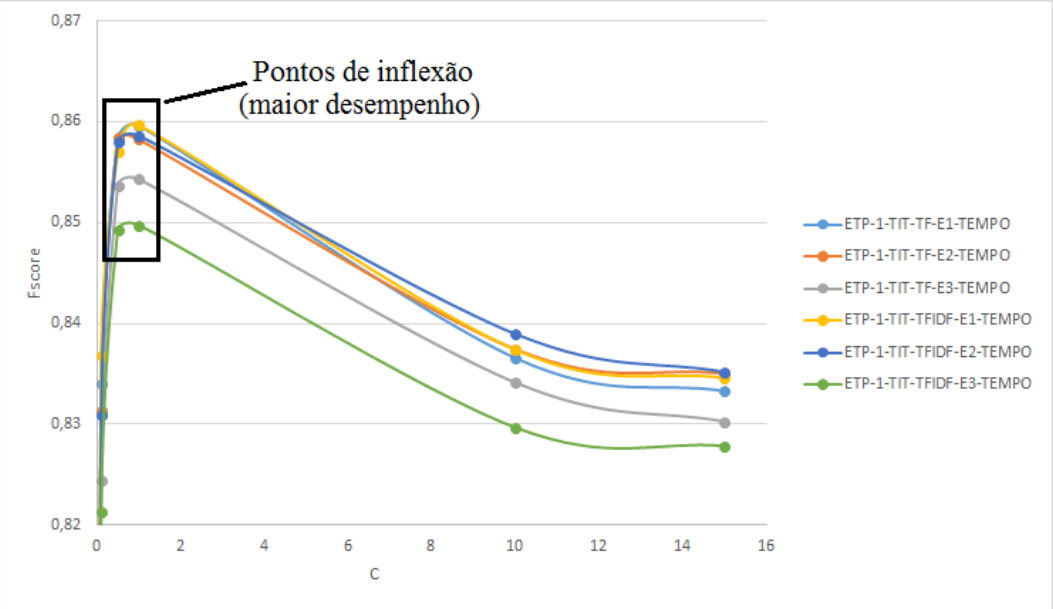

Figura 18: Exemplo de heurística para a constante de regularização (C).

Conforme apontado anteriormente, para tratar os problemas linearmente não separáveis, além do uso das constantes de penalização há também o uso das funções *Kernel*. Propostas inicialmente por (Aizerman, Braverman, Rozonoer, 1964) esta abordagem se baseia em mapear os dados de entrada em um espaço de características de alta dimensão. Após efetuar o mapeamento não linear com estas funções, aplica-se o SVC para construir um hiperplano ótimo de separação neste espaço de alta dimensão. Neste ambiente o problema seria linearmente separável, conforme mostrado na Figura 19:

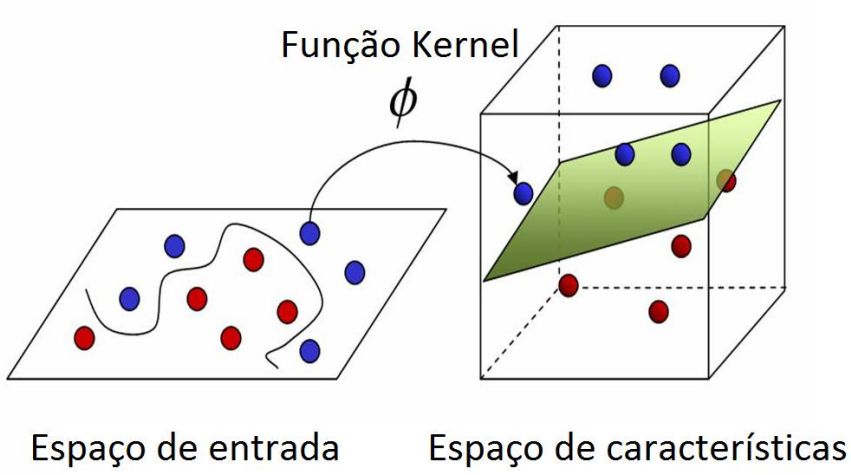

Figura 19: Aplicação das funções *Kernel* a problema não lineares. (Imtech, 2012)

As funções *Kernel* são funções simétricas obtidas pelo produto interno do mapeamento não linear do espaço de entrada no espaço de características (Chaves, 2006). A Tabela 2 apresenta alguns exemplos de funções *Kernel*:

Tabela 2: Exemplos de funções Kernel. (Hsu, Chang, Lin, 2010).

Kernel Name	$K(\vec{x}, \vec{y})$
Linear	$\vec{x} \cdot \vec{y}$
Polynomial	$(\vec{x} \cdot \vec{y} + c)^d$
RBF	$e^{-\gamma \|\vec{x} - \vec{y}\|^2}$
Sigmoid	$tanh(\vec{x} \cdot \vec{y} + c)$

Apesar da gama de funções *Kernel* disponíveis, conforme apontado em (Hsu, Chang, Lin, 2010), o *Kernel* linear apresenta bons resultados no cenário de classificação textual. Isto ocorre, pois, o problema de categorização textual abordado pelo método “*bag of words*” faz com que seja utilizada uma matriz de alta dimensionalidade (muitos atributos) e nesse caso, como os documentos já estão altamente dispersos no modelo de espaço vetorial, basta a simples abordagem linear para conseguir resultados tão bons quanto com os *Kernels* mais sofisticados. Esta escolha traz ganho de tempo de processamento além de simplificar a escolha dos parâmetros do SVC.

Ao final, conjugando o uso da margem suave com o Kernel linear, basta efetuar a calibração de somente um parâmetro: a constante de penalização C.

De modo a finalizar a análise dos SVCs, o último tópico a ser abordado é o uso de SVCs em problemas de múltiplas classes. É sabido que os SVCs foram desenvolvidos originalmente para classificação binária. Já para abordar o problema de várias classes, dois métodos básicos são propostos: “*one-against-all*”, ou um contra todos; e o “*one-against-one*”, ou um contra um (Hsu, Lin, 2012).

A ideia é reduzir um problema de várias classes em vários problemas binários. Para a classificação no método “um contra todos” se constrói N SVCs, onde N é o número de classes. Com isso há um SVC que separa a classe A das demais, outro que separa a classe B de todas as demais (incluindo a classe A) e assim sucessivamente, até o enésimo SVC. Em seguida se agrupam estes SVCs e pode ser feita tanto a combinação linear dos resultados de cada SVC quanto valer-se de métodos de votação (Chaves, 2006). A Figura 20 ilustra a abordagem “um contra todos”:

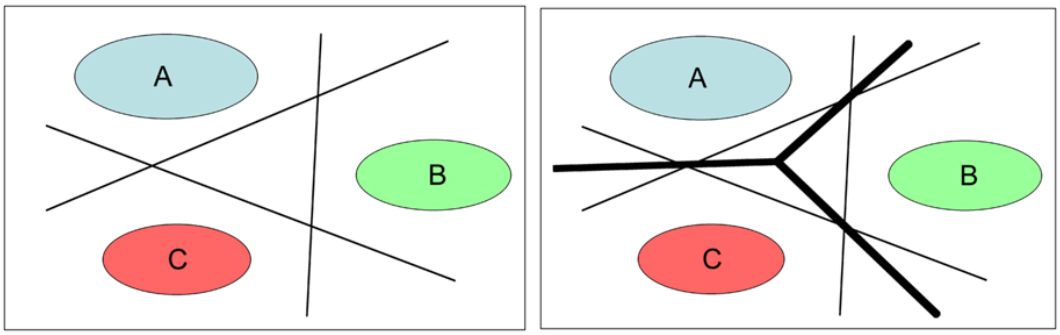

Figura 20: Resolução de um problema de múltiplas classes para SVCs binários: um contra todos.

Já o método “um contra um” separa as classes duas a duas, de modo que cada SVC é treinado com dados de somente 2 classes. Combinando-se todas as classes, constrói-se (N(N-1)/2) SVCs. Apesar de gerar mais SVCs, os problemas a serem resolvidos são de menor dimensão. Estes SVCs também têm seus resultados combinados, sendo o método mais usual a estratégia de votação. A classe que possuir mais votos dos SVCs vence (Hsu, Lin, 2012). A Figura 21 ilustra esta abordagem:

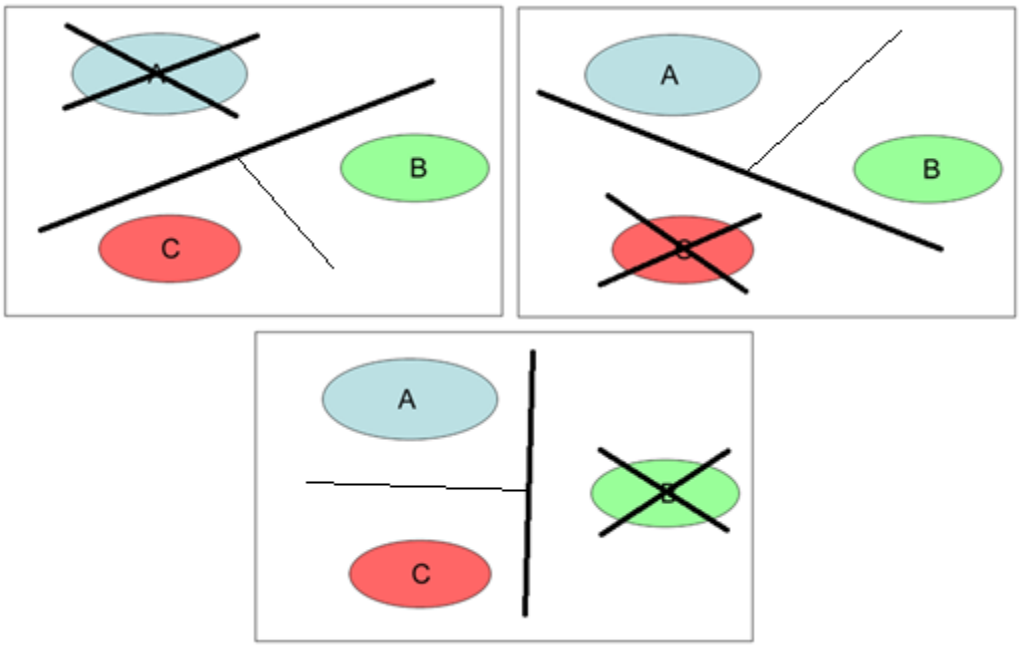

Figura 21: Resolução de um problema de múltiplas classes para SVCs binários: um contra um.

2.3.4.3. Classificação hierárquica ou por etapas

No processo de treinamento dos modelos de categorização, quando há um desbalanceamento significativo no número de observações das classes, estes começam a enfrentar dificuldades em categorizar as classes de menor tamanho. Isso ocorre, pois há menos exemplos disponíveis para estas classes (quando comparadas com as classes maiores) e o classificador fica menos exposto às essas classes menores.

Uma das formas de mitigar este problema é efetuar a classificação por etapas, ou hierárquica. Esta abordagem separa o problema de classificação em duas ou mais etapas, utilizando dois ou mais classificadores para efetuar a tarefa. Conforme ilustrado na Figura 22, esta abordagem consiste em classificar em uma primeira etapa as classes maiores (geralmente de desempenho de classificação maior). As “demais classes” seriam consolidadas em uma nova classe de forma a serem separadas das classes maiores nesta primeira etapa. Em seguida, todas as observações classificadas pelo primeiro classificador como “demais classes” seriam submetidas a um segundo classificador, que só foi treinado com os

exemplos destas classes de menor tamanho, e com isso o processo se conclui com a categorização de todas as classes:

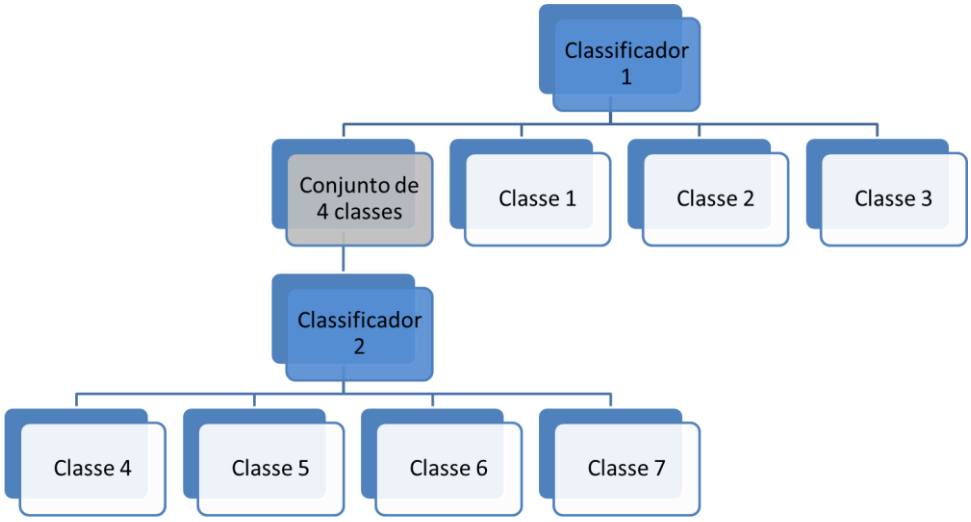

Figura 22: Exemplo de classificação hierárquica de duas etapas.

Para efetuar a divisão entre as classes que estariam na primeira etapa, das classes que estariam na segunda etapa, a seguinte heurística é aplicada:

- Efetua-se a abordagem usual (somente um classificador para todas as classes em etapa única).
- Analisa-se o resultado, de forma a separar as classes com maior desempenho (usualmente as maiores) das que tiveram menor desempenho das menores. Porém pode ocorrer de classes menores possuírem alto desempenho. Estas classes menores também seriam selecionadas para a primeira etapa, pois seu desempenho já é satisfatório. Isto é feito geralmente estipulando um corte mínimo de desempenho.
- As classes restantes seriam alocadas para um segundo classificador, mais especializado.

O ganho da classificação por etapas está em se utilizar um classificador mais especializado para grupos de classes distintos, ao invés de se utilizar somente um classificador para se adequar a todas as classes em somente uma etapa.

Por outro lado, os maiores reveses desta abordagem consistem em:

1) aumento do tempo de processamento, devido ao uso de dois classificadores ao invés de somente um. Isso ocorre, pois há mais parâmetros para serem calibrados (o dobro do usual).

2) as falhas de classificação do conjunto de “demais classes” (ou classes menores) do primeiro classificador impactarão diretamente no desempenho do segundo classificador. Ou seja, se o primeiro classificador acertar 70% da classificação das “demais classes”, o segundo classificador poderá acertar no máximo 70% das observações que ele recebeu da primeira etapa, pois 30% pertencem às classes maiores e foram classificados erroneamente como “demais classes” na primeira etapa.

2.3.4.4. Balanceamento das classes

Conforme abordado na seção anterior, em problemas de aprendizado de máquina é importante haver uma quantidade razoável de exemplos para que o algoritmo aprenda no treinamento as peculiaridades de cada classe. Conforme (He, Garcia, 2009) grande parte dos algoritmos espera que as bases possuam um balanceamento para que possam desempenhar sua classificação de forma mais eficaz, mas muitas vezes as bases apresentam elevado desbalanceamento, sendo necessárias técnicas para ao menos tentar equilibrar as classes menos numerosas.

Além da abordagem de classificação hierárquica, pode-se atuar diretamente no número das observações, aumentando classes menores e até reduzindo classes numerosas. De acordo com (Weiss, 2004), o balanceamento da base de dados pode ocorrer por amostragem (balanceamento aleatório de dados) ou por heurísticas diversas. Para os métodos de amostragem, de acordo com (Batista, Prati, Monard, 2004), observa-se em várias referências que diversos autores concordam que esta abordagem pode acarretar em distúrbios não desejados nos resultados. O mais proeminente destes distúrbios é o fenômeno de *overfiting* sobre os classificadores, devido à pura e simples replicação das observações das classes menores. Da mesma forma, a redução de classes majoritárias de forma aleatória pode remover observações importantes para o aprendizado do classificador.

Nas próximas seções serão discutidos métodos de balanceamento (tanto para aumento quanto para redução da base de dados), seja por métodos de amostragem, seja por outras heurísticas propostas.

2.3.4.4.1. Aumento da base de dados

- **Aleatório**

Consiste em aumentar a quantidade de classes menores escolhendo aleatoriamente documentos pertencentes a estas classes e replicando-os, até atingir a quantidade desejada. Conforme já apontado, pode acarretar em *overfitting* do modelo, pois a pura e simples repetição de documentos acaba por reforçar o algoritmo sobre esses exemplos numerosos, porém repetidos.

- **SMOTE**

(Chawla et al, 2002) propuseram um método para evitar o *overfitting* das classes minoritárias. O SMOTE (*Synthetic Minority Oversampling Technique*) cria novos dados por meio de interpolação de observações já existentes.

A figura 23 esclarece o funcionamento do algoritmo: no passo 1, para cada exemplo da classe minoritária k, computar os exemplos mais próximos e que pertençam à sua classe: elementos (i, j, l, m, n). No passo 2 se escolhe aleatoriamente um exemplo dos elementos obtidos no passo 1. Para o passo 3, gera-se sinteticamente um evento k_1, de modo que o mesmo esteja entre k e i. Finalmente, no passo 4 observa-se a replicação do método por mais duas vezes (gerando uma observação entre k e l e entre k e j):

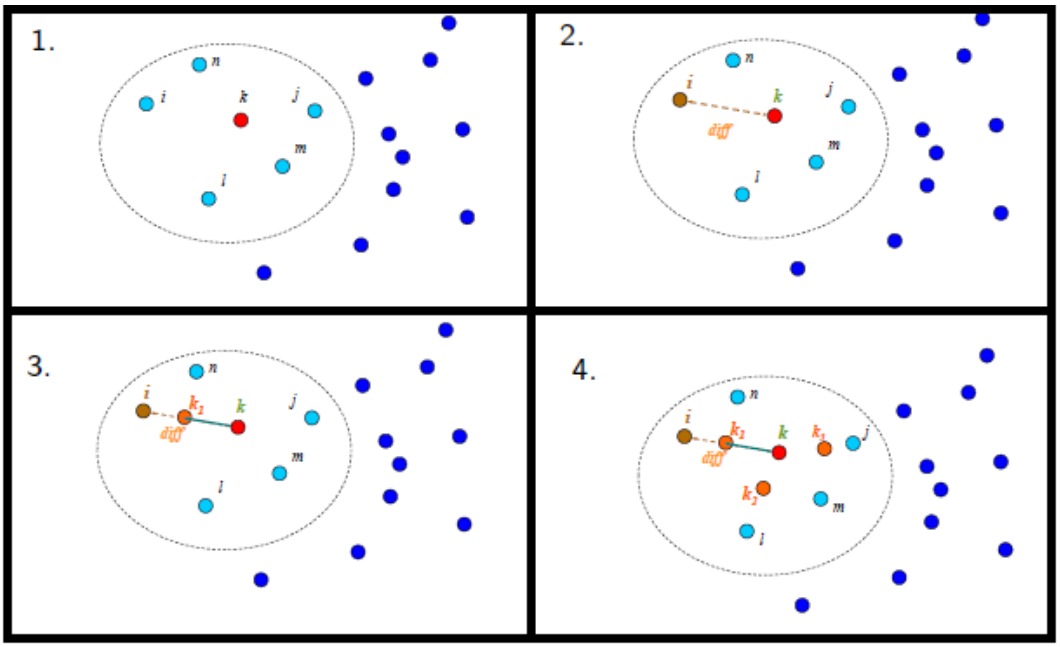

Figura 23: Funcionamento do algoritmo SMOTE. Adaptado de (Pozzolo et al, 2013).

2.3.4.4.2. Redução da Base de Dados

Conforme (Pozzolo et al, 2013) as técnicas de redução consistem em reduzir ruídos presentes nestas classes majoritárias bem como auxiliam na remoção de eventos que estejam na “fronteira” entre as classes. O objetivo é retirar possíveis dados de borda, que possam confundir o classificador durante o aprendizado.

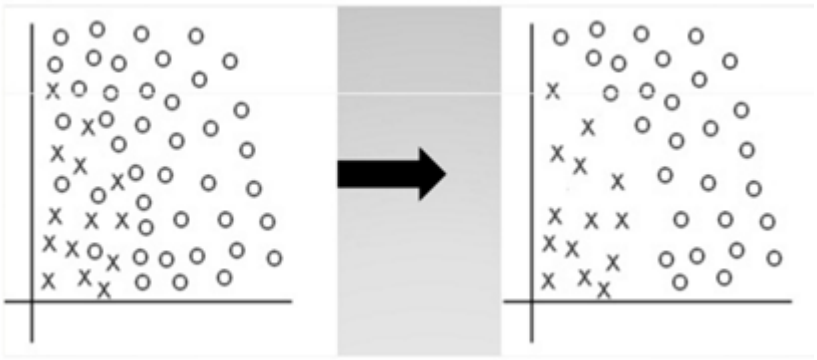

Figura 24: Aplicação da remoção de dados. Adaptado de (Pozzolo et al, 2013).

Nos próximos tópicos serão tratados os métodos de redução de observações para as classes de maior número.

- **Aleatória**

Consiste em retirada aleatória de elementos. Como não há uma lógica para retirada dos pontos, pode-se também excluir dados importantes para o aprendizado do algoritmo.

- ***Tomek Links***

O algoritmo *Tomek Links*, proposto por (Tomek, 1976), efetua a remoção dos exemplos da classe majoritária que estão próximos às demais classes. A Figura 25 ilustra a aplicação do algoritmo para duas classes, onde o círculo representa a classe majoritária e a classe menor é representada pelo formato de estrela. Pela Figura 25 a), são efetuados *links* entre os elementos da classe majoritária (representados por círculos) que estão mais próximos aos elementos das classes minoritárias (representadas por estrelas). Em seguida, estes elementos da classe majoritária são excluídos, mantendo somente os pontos das classes menores, de acordo com a Figura 25 b):

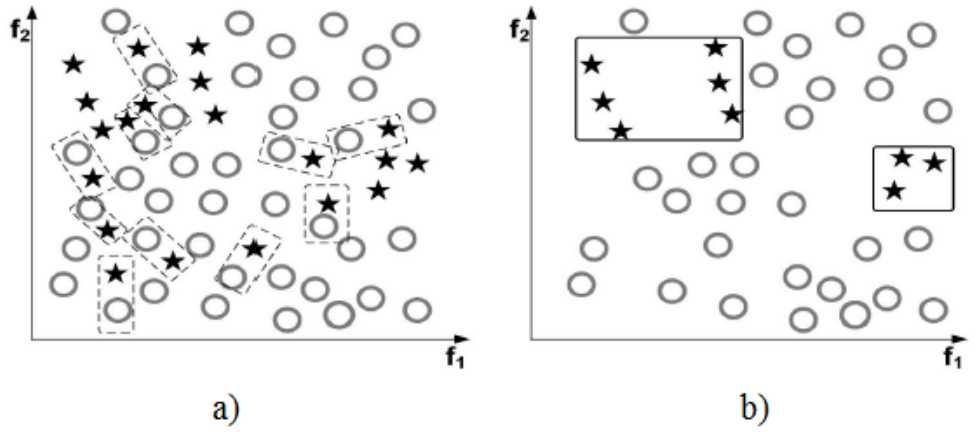

Figura 25: Aplicação do algoritmo Tomek Links. (He, Garcia, 2009).

- ***Edited Nearest Neighbor (ENN)***

Proposto por (Zhang, Mani, 2003), este algoritmo opera da seguinte maneira: se um elemento pertence à classe majoritária e a classificação dada pelos seus três vizinhos mais próximos contradiz a classe original do elemento, este elemento é removido. Com isso se retira elementos de borda, bem como ruídos presentes na classe majoritária.

- ***Neighborhood Cleaning Rule (NCL)***

Proposto por (Laurikkala, 2001), este algoritmo inicialmente aplica o algoritmo anterior (*ENN – Edited Nearest Neighbour*). Se um elemento pertencer à classe minoritária e seus 3 vizinhos mais próximos o classificarem de forma errada, então todos estes vizinhos que pertencem à classe majoritária são removidos. Ou seja, é semelhante ao algoritmo anterior, porém mais drástico na redução dos exemplos de fronteira.

2.3.5. Análise

Após todo o processo de construção da matriz de dados, bem como na elaboração da estrutura do modelo de classificação a ser empregado, inicia-se a etapa de análise. Esta etapa também pode ser chamada de pós-processamento, pois consiste em como atribuir uma métrica de avaliação com o intuito de selecionar os modelos mais aptos a cumprirem a tarefa de classificação textual. Serão abordadas as análises via matriz de confusão, bem como serão avaliadas métricas para seleção dos melhores modelos.

2.3.5.1. Matriz de confusão

A maneira mais usual de avaliar o desempenho de um classificador é por meio da matriz de confusão. Também conhecida como matriz de erro (Stehman,

1997), a matriz de confusão é uma tabela que possibilita visualizar de forma clara o desempenho de um algoritmo de classificação para todas as classes preditas.

Cada coluna da matriz representa as observações de uma classe predita pelo algoritmo enquanto as linhas representam as observações reais de uma classe. Na Figura 26 observa-se o exemplo clássico para duas classes (matriz de confusão binária), onde se considera como a classe “Positivo” a ocorrência de certo evento e a classe “Negativo” como sua não ocorrência:

		Predito pelo modelo	
		Positivo	Negativo
Classe real	Positivo	tp	fn
	Negativo	fp	tn

Figura 26: Matriz de confusão para 2 classes. tp = true positive, fp = false positive, fn = false negative, tn = true negative.

As células destacadas em cinza indicam a quantidade absoluta de observações preditas de forma correta pelo classificador. Já as demais células ilustram as quantidades absolutas de erros do classificador. Para este caso de duas classes, as atribuições abaixo são as mais usualmente empregadas:

- tp: *true positive* - verdadeiros positivos, ou as observações preditas como positivas pelo modelo e que de fato eram positivas;
- tn: *true negative* - verdadeiros negativos, ou as observações preditas como negativas pelo modelo e que de fato eram negativas;
- fp: *false positive* - falsos positivos, ou as observações preditas como positivas pelo modelo e que eram negativas;
- fn: *false negative* - falsos negativos, ou as observações preditas como negativas pelo modelo e que eram positivas;

Os valores correspondentes de tp, tn, fp e fn nas células das matrizes podem ser utilizados tanto em valores absolutos, quanto em valores percentuais relativos ao total de observações de sua respectiva classe real, conforme apontado na Figura 27:

Valor absoluto		Predito pelo modelo		
		Positivo	Negativo	Total
Classe real	Positivo	40	40	80
	Negativo	5	95	100

Valor percentual		Predito pelo modelo	
		Positivo	Negativo
Classe real	Positivo	50,00%	50,00%
	Negativo	5,00%	95,00%

Figura 27: Exemplo de representação da matriz de confusão.

Com isso pode-se ter uma visualização rápida do desempenho do algoritmo para cada classe, de forma a observar percentual das observações que o classificador apontou para cada classe. No exemplo acima se observa alto acerto na previsão da classe “Negativo” (95%), porém alto erro na classificação da classe “Positivo” (50%).

2.3.5.2. Métricas para avaliação

Como o processo de mineração de dados consiste em se testar inúmeros modelos, se faz necessária uma análise mais direta do desempenho de cada classificador, ou seja, seria inviável analisar as matrizes de todas as configurações uma a uma. Obviamente que, após encontrar os modelos de melhor desempenho, pode-se retornar à avaliação da matriz de confusão para analisar os resultados com maior detalhamento.

De modo a efetuar a avaliação dos resultados de forma direta, conforme indicado em (Sokolova, Lapalme, 2009), há as seguintes opções, para um classificador binário:

$$Acurácia = \frac{tp + tn}{tp + fn + fp + tn}$$

Equação 4

$$Abrangência = \frac{tp}{tp + fp}$$

Equação 5

$$Precisão = \frac{tp}{tp + fn}$$

Equação 6

No cenário de métricas de classificação, para o caso da abrangência busca se avaliar a taxa de acerto da classe “Positivo” (tp) quando comparada com todas as observações preditas como “Positivo” pelo modelo (tp+fp).

Já a precisão é a avaliação da taxa de acerto da classe “Positivo” (tp) quando comparada com todas as observações reais da classe “Positivo” (tp+fn). Pode ser considerado como o percentual de acerto de uma classe específica (no caso a classe “Positivo”).

Finalmente, a acurácia nada mais é que o percentual de acerto do modelo em relação a todas as observações avaliadas. Esta métrica tende a privilegiar modelos que possuem desempenho alto para as classes maiores, em detrimento às classes menores.

Outra métrica existente é o F-score. Ela consiste em efetuar a média entre as métricas de precisão e abrangência. A Equação 7 mostra como este cálculo é efetuado (Sokolova & Lapalme, 2009):

$$F - score = \frac{\beta^2 + 1\ tp}{\beta^2 + 1\ tp + \beta^2 fn + fp}$$

Equação 7

Utilizando β = 1 leva a um caso particular da métrica F-score, conhecida como F1-score, ou também como média harmônica entre precisão e abrangência. Esta aplicação é a mais usual.

Até este ponto toda a formulação das métricas considerou somente um problema binário (classes: positiva ou negativa). Em (Sokolova, Lapalme 2009) se estende a análise para o problema de mais de duas classes:

$$Precisão_{\mu} = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} tp_i + fp_i}$$

Equação 8

$$Abrangência_{\mu} = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} tp_i + fn_i}$$

Equação 9

$$F\ score_{\mu} = \frac{\beta^2 + 1\ \ Precisão_{\mu} Abrangência_{\mu}}{\beta^2\ \ Precisão_{\mu} + Abrangência_{\mu}}$$

Equação 10

$$Precisão_{M} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$$

Equação 11

$$Abrangência_{M} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$$

Equação 12

$$F\ score_{M} = \frac{\beta^2 + 1\ \ Precisão_{M} Abrangência_{M}}{\beta^2\ \ Precisão_{M} + Abrangência_{M}}$$

Equação 13

Onde temos que:

tpi: *true positive,* ou verdadeiro positivo relativo à classe i;

fpi: *false positive* ou falso positivo relativo à classe i;

fni: *false negative* ou falso negativo relativo à classe i;

tni: *true negative* ou verdadeiro negativo relativo à classe i;

Índice Macro (M): efetua a média de todas as medidas de precisão e abrangência, para todas as classes;

Índice Micro (μ): não leva em consideração a média das classes para as medidas de precisão e abrangência;

β = 1: média harmônica;

l: número de classes.

Conforme apontado em (Sokolova, Lapalme, 2009), a métrica Macro-F-score trata as classes ponderando o desempenho de cada uma delas de forma igual, enquanto a Micro-F-score favorece classes maiores.

Apesar de listado acima como uma simples fórmula, a aplicação prática das medidas F-score (tanto micro quanto macro) para um problema com três ou mais classes não é trivial e também não é abordada na referência. Foi proposta a inferência de uso conforme exemplo baseado na Figura 28:

		Predito pelo modelo			
		1	2	3	4
Classe real	1	tn 3	fn 3	fp3	fn 3
	2	fn 3	tn 3	fp3	fn 3
	3	fn 3	fn 3	tp 3	fn 3
	4	fn 3	fn 3	fp3	tn 3

Figura 28: Matriz de confusão genérica para 4 classes.

Ao se analisar a classe 3, por exemplo, seria considerado para o cálculo das métricas:

- tp, ou verdadeiros positivos: todas as observações preditas como classe 3 e que de fato pertenciam à classe 3 (tp3 na Figura 28).
- fp, ou falsos positivos: todas as observações preditas como classe 3, mas que pertenciam a outras classes (fp3 na Figura 28).
- tn, ou verdadeiros negativos: todas as observações preditas como classe diferente da 3 e que de fato pertenciam às demais classes (tn3 na Figura 28).
- fn, ou falsos negativos: todas as observações preditas como classe diferente da 3, mas que pertenciam a outras classes (fn3 na Figura 28).

Finalmente, outra forma de avaliar diretamente um problema de mais de 2 classes é efetuar a média aritmética do percentual de acerto das classes (sempre com relação ao total de observações em cada classe), conforme exemplo ilustrado na Figura 29:

		PREDITO			
		CLASSE 1	CLASSE 2	CLASSE 2	Total por classe
REAL	CLASSE 1	1000	500	100	1600
	CLASSE 2	100	300	12	412
	CLASSE 3	500	500	15	1015

		PREDITO			
		CLASSE 1	CLASSE 2	CLASSE 2	% por classe
REAL	CLASSE 1	62,50%	31,25%	6,25%	62,50%
	CLASSE 2	24,27%	72,82%	2,91%	72,82%
	CLASSE 3	49,26%	49,26%	1,48%	1,48%

Média entre classes:	**45,60%**

Figura 29: Matriz de confusão com média percentual entre classes.

O próximo capítulo traz o contexto onde o problema se insere, de modo a esclarecer a demanda pela classificação das anormalidades de sonda.

3.
Contextualização do problema

3.1.
Operações para construção de poços marítimos

A extração de hidrocarbonetos de formações rochosas depende da construção de vários poços de petróleo. No cenário específico do Brasil, com unidades *off-shore*, esta atividade é realizada por meio das operações de construção de poços marítimos. As operações se dividem em vários tipos de intervenções no poço, destacando-se: perfuração, completação (equipar o poço), avaliação (coleta de informações do reservatório), manutenção e abandono de poço (quando não há mais interesse no seu uso). O foco deste trabalho se dará somente sobre as intervenções de perfuração.

Cada intervenção possui um tempo total específico, o qual pode ser dividido em tempo útil e tempo perdido. Este critério de divisão se baseia principalmente no planejamento inicial das operações, cujo tempo é previsto de acordo com um histórico de intervenções. O tempo útil seria o intervalo de tempo esperado para se realizar a intervenção, assumindo que não ocorram anormalidades.

Porém, caso o escopo das operações fuja deste planejamento inicial, devido à ocorrência de anormalidades na intervenção, o tempo dispendido para retornar à situação inicial de planejamento é considerado como tempo perdido. Fatalmente podem ocorrer atrasos inerentes à operação, mas só serão considerados tempos perdidos os atrasos ocorridos por anormalidades (eventos não esperados).

O corpo gerencial analisa sob o mesmo enfoque as parcelas de tempo útil e perdido. O tempo útil é importante, pois corresponde à maior parcela dos tempos de intervenção, de modo que seu aperfeiçoamento impacta mais na redução do tempo total de intervenção de um poço. Já o tempo perdido, apesar de representar uma duração absoluta menor quando comparada à do tempo útil, geralmente está relacionado à ocorrência de desvios críticos. O tempo perdido muitas vezes é um

sinal da ocorrência de condições inseguras para a operação, tornando sua análise tão importante quanto à do tempo útil.

3.2.
Equipamentos necessários para perfuração de um poço de petróleo marítimo

A perfuração de um poço marítimo requer uma série de sistemas integrados e equipamentos complexos, além de uma robusta estrutura em terra para prover os insumos necessários para a execução das atividades.

Em todo o processo, a unidade de intervenção desempenha papel fundamental nas operações, sendo o local onde todas elas se sucedem. Esta unidade pode receber o nome de plataforma de perfuração, ou simplesmente de sonda. A Figura 30 ilustra os tipos de unidades existentes:

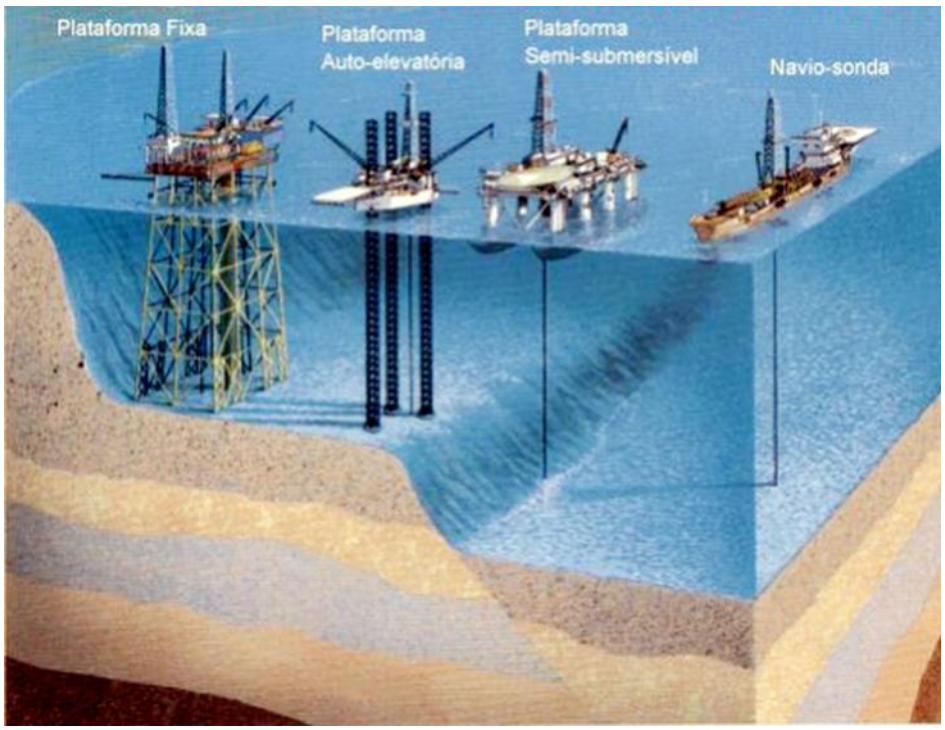

Figura 30: Exemplos de sondas utilizadas nas intervenções de poços marítimos.

O tipo de unidade utilizada na intervenção de um poço marítimo está intimamente atrelado à profundidade da cabeça do poço, ou seja, quão distante a sonda está do leito marinho. O termo utilizado para esta distância é lâmina d’água (LDA). Para LDAs menores, é mais comum o uso dos dois primeiros tipos de plataforma (Plataforma Fixa e Plataforma Auto elevatória), porém, para LDAs dos

campos de petróleo mais recentes (como os do pré-sal), as operações só são viáveis com o uso dos dois últimos tipos de unidade de intervenção (Plataforma Semissubmersível e Navio-sonda), devido à grande profundidade do leito marinho (geralmente em torno de 2000 metros).

Estes dois tipos de unidade de intervenção se mantêm em posição utilizando propulsão própria (não há uso de âncoras tão pouco estruturas fixas), sendo denominadas unidades de posicionamento dinâmico (*DP – dynamic positioning*). Elas devem se manter dentro de um raio que possui como centro a projeção da cabeça de poço submarino na superfície do mar. A Figura 31 ilustra isso:

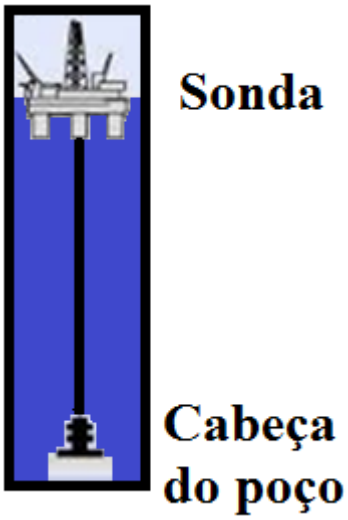

Figura 31: Esquema da plataforma de posicionamento dinâmico.

O foco deste trabalho será nestes tipos de unidade em específico.

As plataformas de posicionamento dinâmico possuem vários sistemas complexos e integrados que podem ser representados de forma resumida na Figura 32 a seguir.

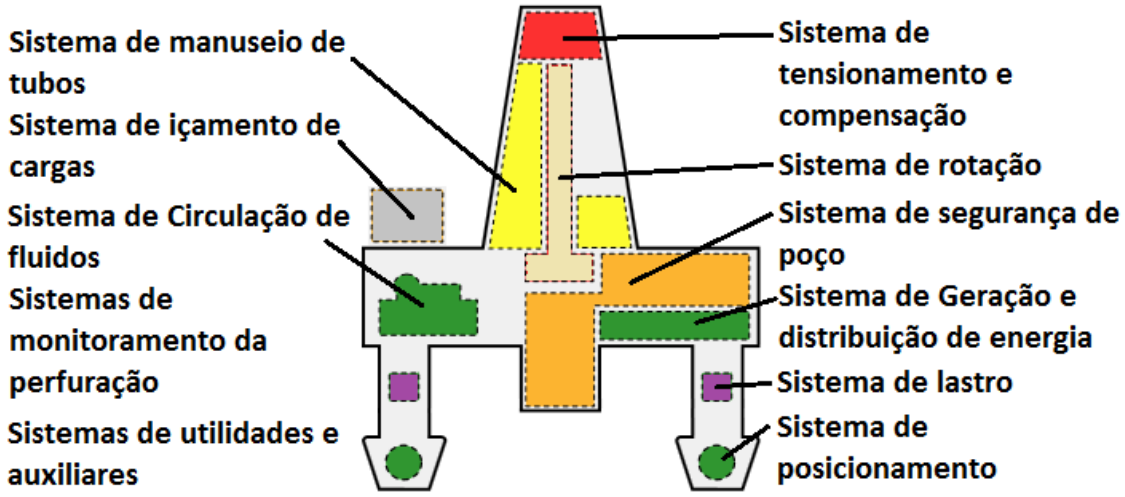

Figura 32: Sistemas presentes nas unidades de intervenção. Figura adaptada de (Petrobras, 2017).

Cada sistema desempenha um papel importante na intervenção do poço. Os sistemas de monitoramento da perfuração e os sistemas de utilidades e auxiliares não estão apontados na Figura 32. Os sistemas de monitoramento da perfuração se relacionam aos equipamentos que fazem medição de parâmetros relevantes às operações e os sistemas de utilidades e auxiliares são os demais equipamentos que dão apoio às intervenções, mas não se encaixam nos sistemas anteriores, como sistemas de ar comprimido, unidade de cimentação bem como o *remoted operated underwater vehicle* (conhecido como ROV) que é um veículo operado remotamente, utilizado para avaliar visualmente a cabeça de poço submarina, bem como executar outras atividades de atuação de válvulas, limpeza e coleta de material.

Até agora foram apontados todos os sistemas de equipamentos residentes na unidade de intervenção. Porém há também as ferramentas destinadas a operações específicas que são utilizadas de forma temporária nas operações e em seguida são desembarcadas para uso em outras unidades. Estas são denominadas equipamentos de trabalho no poço.

Os equipamentos de trabalho no poço envolvem desde brocas de perfuração, equipamentos de “geodirecionamento” da perfuração, passando também por equipamentos específicos de superfície.

Finalmente, para que o poço atue plenamente na drenagem do óleo do campo, é necessário instalar vários equipamentos que ficarão residentes no poço por vários anos. Estes são denominados equipamentos do poço.

3.3. Banco de dados de operação

Todas as operações mencionadas anteriormente na Seção 3.1 são registradas por meio de relatórios digitais. Eles são como diários de bordo, relatando as ocorrências diárias das operações. No caso específico da Petrobras, utiliza-se o software denominado *Open Wells* (Landmark, 2017) (gerido pela Petrobras com apoio da empresa *Landmark*, subsidiária da empresa *Halliburton*). O *Open Wells* é a interface onde o coordenador de operações a bordo efetua o registro de todas as operações ocorridas durante as intervenções nos poços. Estes registros são realizados em boletins diários de perfuração, os BDPs (para intervenções de perfuração de poços), ou em boletins diários de completação e avaliação, os BDCAs. Estes abarcam as demais intervenções: instalação de equipamentos, obtenção de informações, manutenções e abandono de poços. O foco deste trabalho se dará nos boletins diários de perfuração (BDPs).

3.4. Anormalidades na construção de poços marítimos

Conforme apontado no início deste capítulo, uma anormalidade caracteriza-se por um desvio não planejado durante a construção de um poço. Os motivos podem ser diversos: desde falhas relacionadas à ação humana ou de equipamentos, passando por condições meteorológicas adversas, até acidentes geológicos (grandes falhas na rocha perfurada, por exemplo).

No caso particular da Petrobras, todos estes desvios são avaliados e categorizados manualmente por um operador experiente (o coordenador de operações, também conhecido como engenheiro fiscal) de acordo com um sistema de classificação de anormalidades elaborado internamente na Petrobras. A Figura 33 exemplifica uma anormalidade, bem como as informações relacionadas à mesma e que são preenchidas nos BDPs:

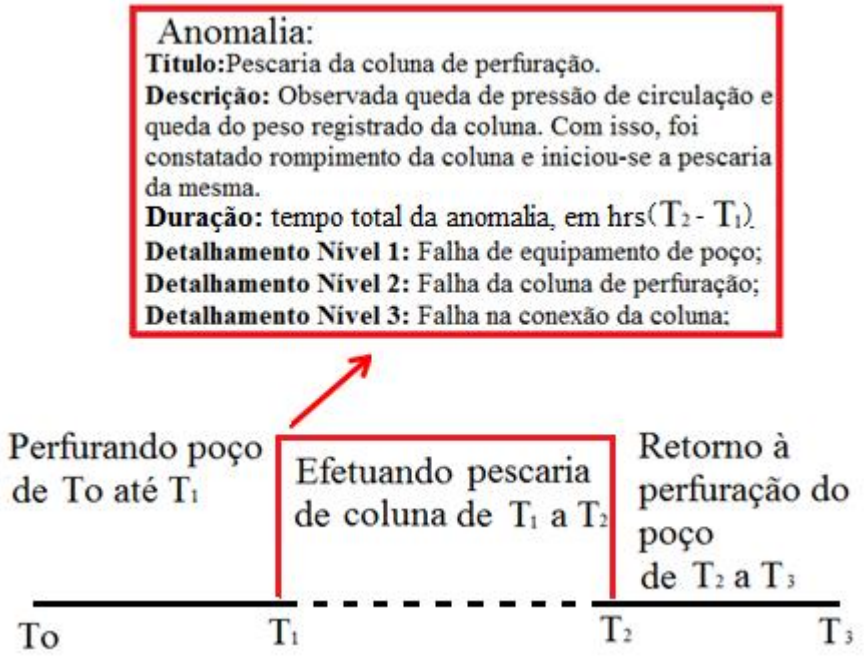

Figura 33 Exemplo de uma anormalidade e seu registro na base de dados.

A Figura 33 ilustra a ocorrência de uma queda de parte da coluna de perfuração no poço, por rompimento de tubulares. Esta ocorrência acarretou a necessidade da retirada (ou pescaria) do trecho que caiu no poço, para em seguida retomar a operação de perfuração. Isso causa uma interrupção no caminho planejado das operações, conforme evidenciado pelo segmento em vermelho, do ponto T_1 até T_2, da Figura 33. Destacam-se na caixa em vermelho as informações registradas no software para a anormalidade ocorrida, com dois campos principais: Título e Descrição da anormalidade, os quais aceitam texto livre.

O campo Título contém um resumo da ocorrência e o campo Descrição apresenta um maior detalhamento do evento, para análises futuras. Há também a informação sobre a duração da anormalidade (tempo decorrido em horas de T_1 até T_2). Além de todas essas informações, há também um apontamento sobre qual gerência interna à Petrobras será responsável por tratar a anormalidade (não mencionado na figura). Finalmente há informações inseridas pelo operador que devem ser selecionadas a partir de uma lista fixa e pré-determinada, para classificar manualmente os três níveis de detalhamento existentes.

Esta lista é disposta de forma hierarquizada, ou seja, quando o operador seleciona o primeiro nível (mais abrangente), automaticamente a lista abaixo desta

é liberada para preenchimento, trazendo um detalhamento maior para a anormalidade (nível 2); e sucessivamente, após preencher o nível 2, o nível 3 (quando existente) também é liberado para escolha. Com isso, parte-se de uma classe mais abrangente, a qual será detalhada nos níveis inferiores, de forma a melhor especificar o incidente.

Nos boletins diários de perfuração (foco desta dissertação), há uma seção exclusiva para o registro destas anomalias, ou anormalidades (desvios operacionais de planejamento), ocorridas durante as operações. Um exemplo desta seção específica pode ser visualizado na Figura 34:

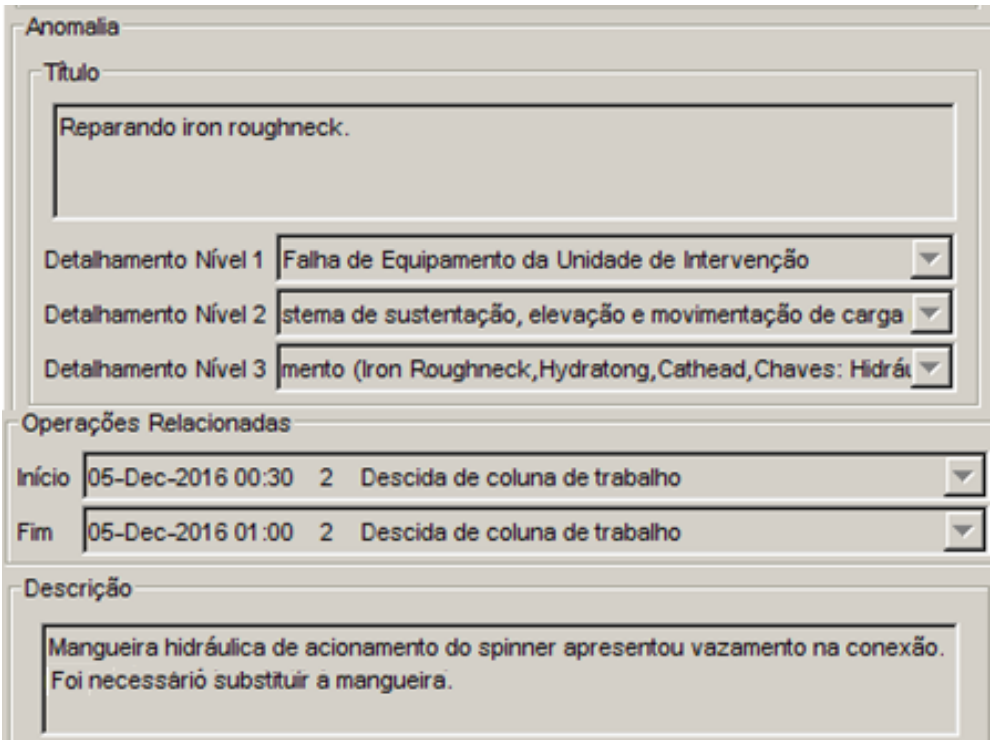

Figura 34: Interface do software Open Wells (Landmark, 2017) para registro de anormalidades.

3.5. Atualização nas classes de anormalidades

Conforme mencionado na seção anterior, as anomalias (ou anormalidades) são classificadas de maneira hierárquica, por meio de três níveis de detalhamento, os quais contemplam uma miríade de opções de classificação. Mesmo assim, como esta lista foi concebida há anos atrás, a mesma não possuía detalhamento

suficiente para algumas anormalidades. De forma a aumentar este detalhamento, vários especialistas da Petrobras se reuniram e foi elaborada uma revisão desta classificação, adequando classes já existentes e inserindo outras novas. Esta nova atualização melhorou a abordagem para tratamento das anormalidades, assim como facilitou o preenchimento por parte do operador, porém gerou um passivo de anormalidades que ainda possuíam a classificação antiga (incompleta).

Conforme evidenciado na Figura 35, na parte superior (em destaque) consta a classe antiga, sem detalhamento adicional (esta classe continha as três novas classes). Na sua parte inferior a figura mostra as novas classes criadas para aumentar o detalhamento na atribuição das anormalidades:

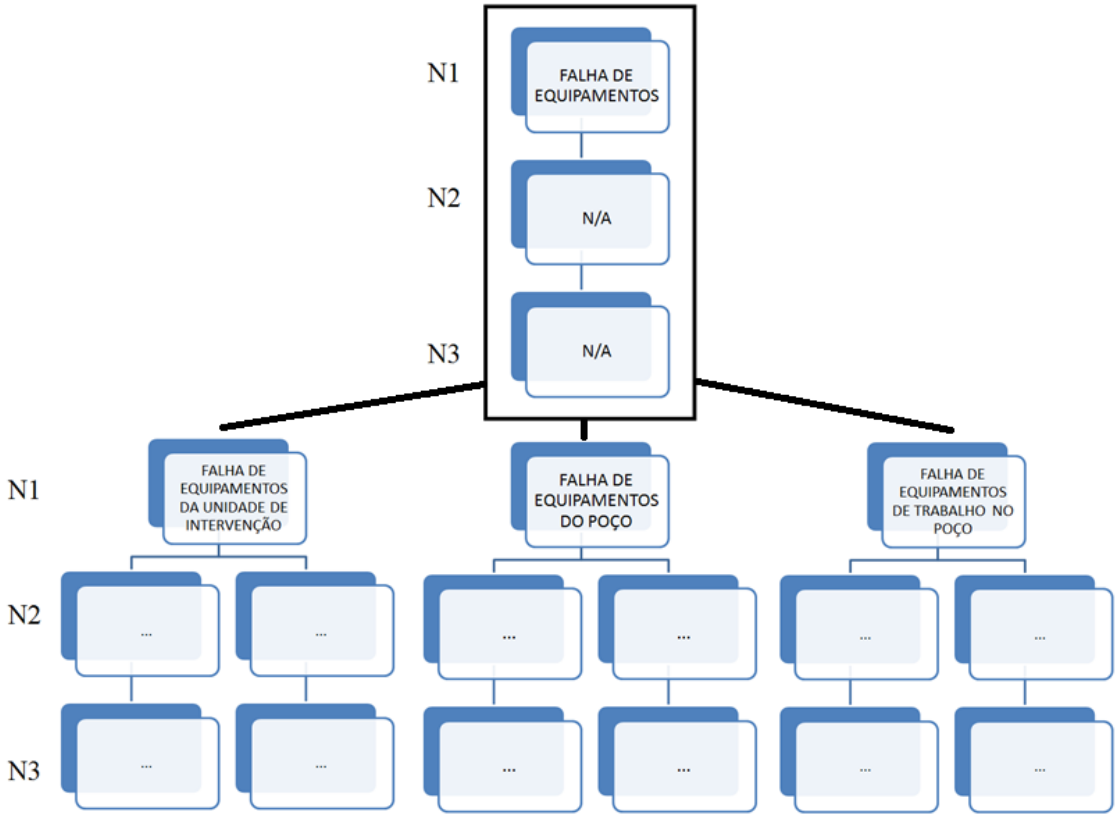

Figura 35: Desmembramento das classes de anormalidades de equipamentos.

A antiga classe “FALHA DE EQUIPAMENTOS” agrupava falhas tanto de equipamentos da unidade de intervenção (sonda) quanto falhas de equipamentos de trabalho no poço, além de falhas de equipamentos residentes no poço. Como fator agravante, esta classe não possuía nenhum item em lista para detalhamento nível 2, tão pouco nível 3.

Devido ao fato da ocorrência de falhas de equipamentos da sonda ser altamente relevante, a classe antiga foi separada em uma classe distinta no nível 1, denominada “FALHA DE EQUIPAMENTOS DA UNIDADE DE INTERVENÇÃO”. As demais falhas foram também desmembradas em “FALHA DE EQUIPAMENTOS DE TRABALHO NO POÇO” e “FALHAS DE EQUIPAMENTOS DE POÇO”.

Para estes três tipos de falha nível 1 também foram criados níveis inferiores (2 e 3) de detalhamento (antes não presentes). No caso específico das anormalidades relacionadas às unidades de intervenção (sondas), foram criadas 11 classes para o detalhamento nível 2, e várias outras classes abaixo destas, no detalhamento nível 3.

É possível fazer uma comparação quanto aos campos presentes nas anormalidades antes e depois da atualização da antiga classe “FALHA DE EQUIPAMENTOS”:

Tabela 3: Comparação dos campos antes e após a atualização da classe “FALHA DE EQUIPAMENTOS”.

Campo de preenchimento	Classe antiga	Novas classes
Título	X	X
Descrição detalhada	X	X
Duração	X	X
Detalhamento nível 1	X	X
Detalhamento nível 2		X
Detalhamento nível 3		X

Após a atualização (ocorrida em junho de 2011), o histórico de falhas de equipamentos começou a ser preenchido sob o molde das novas classes, à medida que as novas operações foram ocorrendo. Porém, a atualização gerou uma descontinuidade na análise das anormalidades para o caso específico de falhas de equipamentos. Este histórico é importante para verificar a evolução das falhas e confirmar se falhas semelhantes (antes debeladas com ações técnicas e gerenciais) não ocorreriam novamente.

As anormalidades relacionadas a equipamentos de trabalho no poço bem como equipamentos de poço foram prontamente tratadas pelas gerências internas

da Petrobras, as quais eram responsáveis pelos contratos dos referidos equipamentos. Porém, para os equipamentos de sonda, não há uma gerência específica para cada sistema da unidade e com isso esta distinção não foi possível de ser realizada. Ou seja, perdeu-se o histórico do banco de dados e gerou-se um passivo de anormalidades de sonda sem classificação detalhada (um total de 3384 anormalidades, somente para as intervenções de perfuração, no período de janeiro de 2008 até junho de 2011). A partir deste problema, surgiu a demanda de classificar estas anormalidades e recompor o histórico tão importante do banco de dados relacionados à falha de equipamentos de sonda.

Conforme se observa na Tabela 3, os únicos campos que trazem informações suficientes para um operador classificar o passivo de anormalidades são campos de informação textual (tanto título quanto a descrição das anormalidades). Com isso, seria necessário alocar operadores para ler cada anormalidade e efetuar sua classificação nos novos moldes. Isto seria inviável, pois tomaria demasiado tempo dos operadores, que são um recurso humano escasso na companhia. Uma ferramenta que pudesse efetuar esta classificação de forma automática seria muito bem-vinda.

Como a partir de junho de 2011 os operadores vêm classificando as novas anormalidades ocorridas já com base na última atualização das classes, gerou-se uma base de dados de anormalidades classificada de acordo com os novos níveis de detalhamentos. Esta base já possui um histórico suficiente para o desenvolvimento de um modelo de classificação, que possa ser ajustado e testado. Após a obtenção da melhor configuração do modelo de classificação, este poderia ser utilizado para analisar e classificar o passivo de anormalidades de forma automatizada.

Desde a atualização dos níveis até o início deste trabalho na elaboração dos modelos de classificação, 9202 anormalidades da classe nível 1 “FALHA DE EQUIPAMENTOS DA UNIDADE DE INTERVENÇÃO” foram preenchidas. A Tabela 4 mostra a distribuição das 11 subclasses (nível 2 de detalhamento) para esta classe nível 1 (por motivos de sigilo serão atribuídos códigos (de 1 a 11) às classes):

Tabela 4: Distribuição das anormalidades de nível 2 de falha de equipamentos de sonda, a serem utilizadas para modelagem do classificador.

Classe	Frequência	Fração da base
1	3077	33,44%
2	1607	17,46%
3	1576	17,13%
4	1522	16,54%
5	432	4,69%
6	266	2,89%
7	209	2,27%
8	187	2,03%
9	172	1,87%
10	142	1,54%
11	12	0,13%

Como pode se observar, a distribuição das classes para o nível de detalhamento 2 é bem desigual. Com isso, a demanda se restringiu em classificar o passivo das anormalidades sem classificação adequada somente até o nível de detalhamento 2 da nova classe nível 1 “FALHA DE EQUIPAMENTOS DA UNIDADE DE INTERVENÇÃO”. O nível 3, portanto, não foi contemplado.

Vale destacar que se espera, além da nova classificação do passivo, uma análise da estrutura atual das classes, de forma a observar se há classes redundantes e se há classes que o operador tenha classificado de forma equivocada e recorrente.

O próximo capítulo apresenta o modelo de classificação automática de anormalidades relacionadas às falhas de equipamentos de sonda, desenvolvido neste trabalho.

4. Desenvolvimento do modelo

4.1. Introdução

Conforme especificado na Figura 36, a modelagem do processo de classificação das anormalidades inicia-se na coleta de dados (textuais e numéricos) relacionados às falhas de equipamentos de sonda na base de dados das operações de construção de poços marítimos. Em seguida há o tratamento destes dados, na etapa de pré-processamento, que é feito utilizando ferramentas de mineração textual para os dados textuais, que por sua vez são transformados em dados numéricos (dispostos na matriz de termos por documentos). Ainda no pré-processamento a normalização é aplicada tanto para os dados originalmente numéricos (duração das anormalidades) bem como para os dados textuais que foram transformados em dados numéricos (representados pela matriz de termos por documentos). Numa próxima etapa os dados pré-processados são submetidos ao classificador, que pode executar a tarefa em etapa única (usual) ou em mais de uma etapa (hierárquico). De forma opcional, é possível alterar os dados de entrada pré-processados por meio do balanceamento da base de dados e em seguida submetê-los ao classificador. Finalmente, o processo termina com o classificador atribuindo as classes para os documentos analisados e pela avaliação final do seu desempenho:

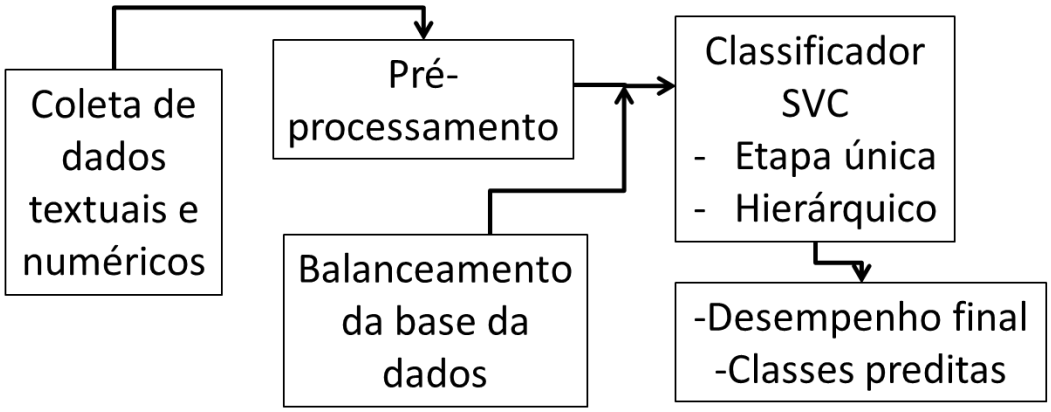

Figura 36: Diagrama básico do modelo de classificação utilizado.

Este diagrama será detalhado nas próximas seções.

4.2. Coleta de dados textuais e numéricos

Conforme apontado em seções anteriores, os dados para este trabalho são em sua maioria textuais, mas também há informações numéricas (duração das anormalidades). Ambos os dados foram utilizados neste trabalho e cada um deles demanda um tratamento específico, o que será abordado na próxima seção.

4.3. Pré-processamento

O pré-processamento foi efetuado tanto para as informações textuais quanto para dados numéricos (duração das anormalidades). Conforme apontado no capítulo 3, em termos de dados textuais há disponível tanto o título quanto à descrição detalhada da anormalidade. O primeiro é basicamente um resumo do segundo. Com isso foram gerados dois corpora diferentes, um para o título e outro para a descrição. Esses corpora foram submetidos ao pré-processamento textual. A Figura 37 traz o diagrama relativo à etapa de pré-processamento:

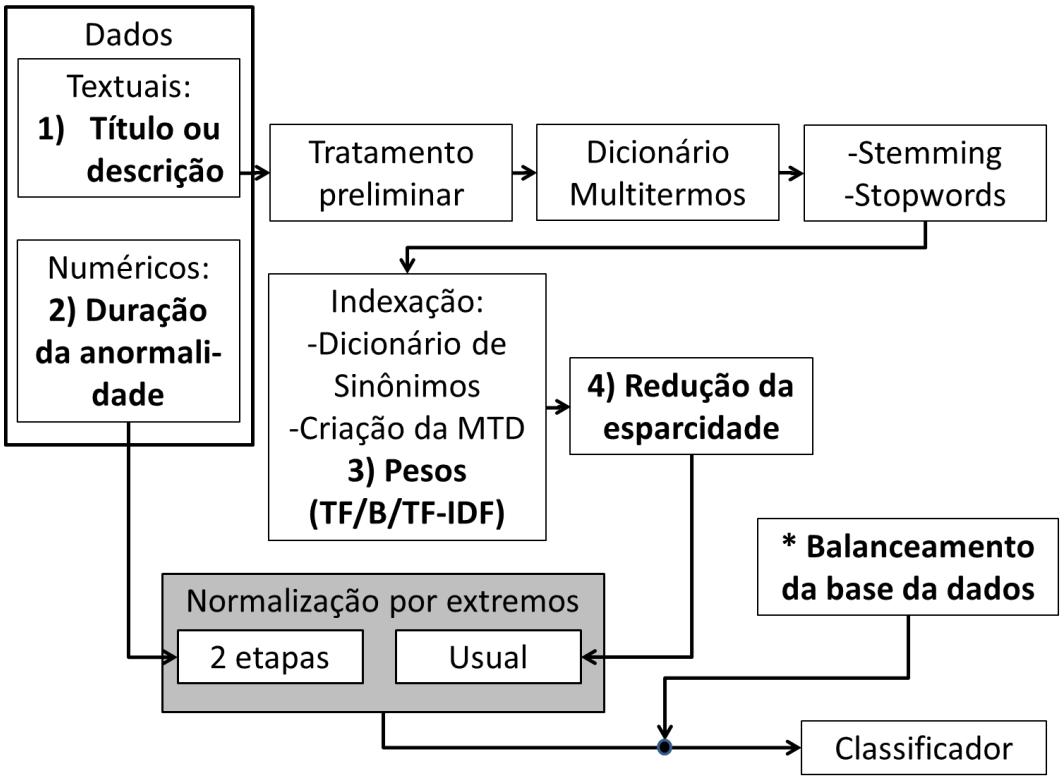

Figura 37: Diagrama da etapa de pré-processamento.

As seções seguintes detalham cada elemento do diagrama.

4.3.1. Dados textuais

Foram utilizadas ferramentas básicas de mineração de texto para, ao final do processo, transformar os dados textuais em dados numéricos. Existem duas opções para o dado textual a ser utilizado: corpus relativo ao campo do título das anormalidades, bem como o corpus relacionado à sua descrição detalhada.

4.3.1.1. Tratamento preliminar

Nesta etapa removeram-se espaços adicionais entre os termos dos documentos, bem como os números dentro dos campos analisados. Além disso, foi aplicado o processo de *Case Folding*, onde todas as letras maiúsculas foram

reduzidas a letras minúsculas (descaracterizando nomes próprios). Também foram normalizados os documentos retirando todo e qualquer sinal de pontuação e acentuação gráfica.

4.3.1.2. Dicionário multi-termos

Esta etapa do processo demandou muito tempo para ser realizada. A área de construção de poços possui uma extensa quantidade de palavras próprias e específicas principalmente no que tange a diversas denominações de ferramentas existentes, além dos jargões específicos da área, o que tornou o processo laborioso. A criação dos dicionários foi feita por especialistas da área.

4.3.1.3. *Stemming & Stopwords*

Nesta etapa efetuou-se o *stemming* (radicalização) e também a remoção de palavras de pouca relevância do corpus (*stopwords*). Para o processo de radicalização aplicou-se o algoritmo de Porter, sendo o *stemmer* utilizado neste trabalho. As *stopwords*, ou lista de palavras a serem removidas, é subdividida em uma lista básica (em sua maioria preposições, pronomes de tratamento, etc) e outra lista complementar elaborada pelos especialistas com termos de pouco relevância no cenário aplicado.

4.3.2. Indexação textual

Primeiramente esta etapa consiste em indexar os termos textuais restantes após o pré-processamento. Como se pode observar, para deixar o banco de dados textuais adequado ao processamento, vários termos não significativos são retirados da análise (números, pontuações, termos não importantes etc.). Em seguida são efetuadas as referências via dicionários de sinônimos, com o mesmo intuito de condensar a informação para um menor número de termos, aprimorando

sua indexação. Finalmente, os documentos são representados via modelo de espaço vetorial. Isto é alcançado pela criação da matriz de termos por documentos, conforme apontado na Figura 38:

COLEÇÃO (CORPUS)	
Documento 1	deposite o dinheiro e o cheque no banco
Documento 2	o barco está preso no banco de areia
Documento 3	sentei no banco da praça
Documento 4	foi para o banco dos réus

MTD: MATRIZ DE TERMOS POR DOCUMENTOS

	depositar	o	dinheiro	e	cheque	no	banco	barco	estar	preso	de	areia	sentar	praça	ir	para	dos	réus
d1	1	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
d2	0	1	0	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0
d3	0	0	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0
d4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1

Figura 38: Exemplo de uma MTD (matriz de termos por documentos).

4.3.2.1. Dicionário de sinônimos

Da mesma forma que a criação do dicionário de termos múltiplos, a criação do dicionário de sinônimos também consumiu razoável tempo dos especialistas devido às particularidades técnicas da área de engenharia de poços.

4.3.2.2. Pesos dos termos contidos na matriz de termos por documentos

Na representação pelo modelo de espaço vetorial, os valores a serem assumidos por cada termo, quando relacionado a um documento (após criação da matriz de termos por documentos), podem assumir diversos valores. Os seguintes foram aplicados para os estudos de caso:

- *Term Frequency* (TF): frequência simples dos termos dentro de cada documento;
- Binário (B): verifica ou não a presença do termo (valor da frequência não é considerado);

- *Term Frequency-Inverse Document Frequency* (TF-IDF): métrica elaborada de forma a ponderar a presença do termo com relação a toda coleção de documentos.

Vale ressaltar que esta indexação foi aplicada tanto no corpus pré-processado do campo do título quanto da descrição das anormalidades. Portanto, há disponível até este ponto seis tipos de matriz de termos por documentos:

Tabela 5: MTDs (Matrizes de termos por documentos) disponíveis após indexação.

MTD	CORPUS	PESO
1	TÍTULO	TF
2		BINÁRIO
3		TF-IDF
4	DESCRIÇÃO	TF
5		BINÁRIO
6		TF-IDF

4.3.2.3. Redução da esparcidade

Conforme apontado na Seção 2.3.3.4, foi efetuada a retirada de termos menos representativos com a redução da esparcidade da matriz de termos por documentos. Com isso, os termos que possuíam mais “zeros” na matriz (ou seja, menos frequentes na coleção) foram retirados. Os redutores de esparcidade atuam de maneira a reduzir drasticamente a quantidade de termos e, portanto, foi feita uma análise de sensibilidade (a ser detalhada no próximo capítulo), tanto para redução de termos do campo do título quanto da descrição das anormalidades.

Após a análise, os seguintes valores foram utilizados para o fator:

- E1=1: sem redução de esparcidade;
- E2=0,998: reduz aproximadamente pela metade o número inicial de termos;
- E3=0,9995: reduz aproximadamente a um quarto do número original de termos.

Como a esparcidade é um dos parâmetros dos estudos de caso, haverá a seguinte disposição de matrizes de termos por documentos:

Tabela 6: MTDs (matrizes de termos por documentos) para modelagem considerando esparcidade.

MTD	CORPUS	PESO	REDUTOR ESPARC.
1	TÍTULO	TF	E1
2		BINÁRIO	E1
3		TF-IDF	E1
4	TÍTULO	TF	E2
5		BINÁRIO	E2
6		TF-IDF	E2
7	TÍTULO	TF	E3
8		BINÁRIO	E3
9		TF-IDF	E3
10	DESCRIÇÃO	TF	E1
11		BINÁRIO	E1
12		TF-IDF	E1
13	DESCRIÇÃO	TF	E2
14		BINÁRIO	E2
15		TF-IDF	E2
16	DESCRIÇÃO	TF	E3
17		BINÁRIO	E3
18		TF-IDF	E3

4.3.3. Dado numérico

Apesar de a maior parte dos dados serem apresentados em campos textuais, avaliou-se a inserção da duração das anormalidades no desempenho do modelo. Este dado, único originalmente numérico, foi avaliado em um histograma:

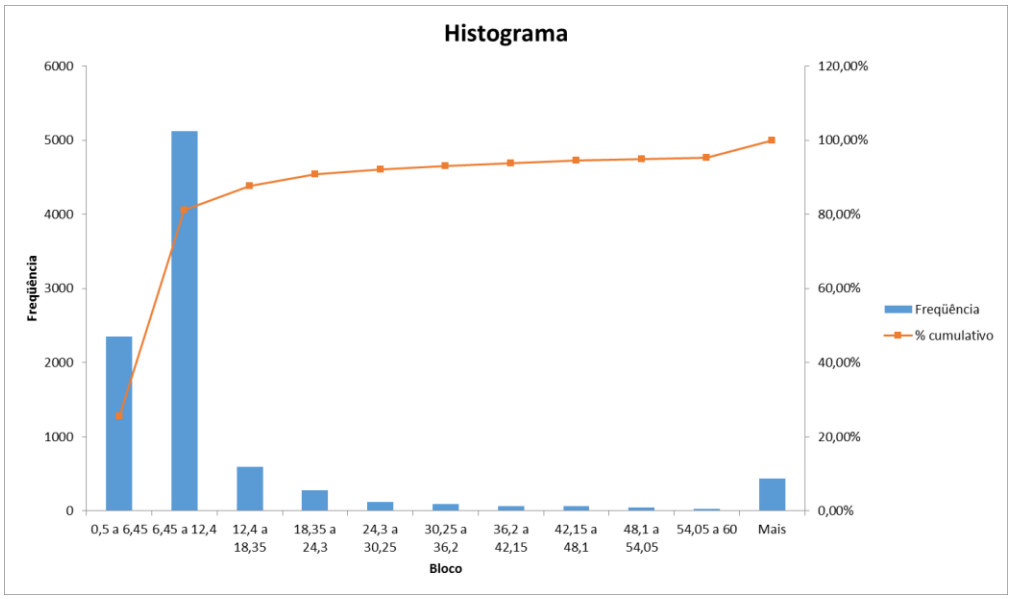

Figura 39: Histograma da duração das anormalidades.

Como pode se observar, a maioria das anormalidades (em torno de 80%) está contida em intervalos de duração de até 12 horas, enquanto as demais variam de 12,5 até um valor máximo que foi truncado para facilitar a visualização. Com isso observa-se a necessidade de normalizar estes dados devido à grande magnitude encontrada para os valores observados.

Como a inserção ou não da duração das anormalidades é um dos parâmetros de configuração de dado de entrada dos estudos de caso e, observando que pela Tabela 6 já há dezoito tipos de matriz de termos por documentos como configurações de entrada a serem avaliadas, haverá 18 x 2 = 36 tipos possíveis de dados de entrada para avaliação.

4.3.4. Normalização dos dados textuais e numéricos

Conforme apontado por (Hsu, Chang, Lin, 2010), se faz necessária a normalização dos dados de entrada antes dos mesmos serem submetidos aos SVCs. Além de evitar distorções nos dados (principalmente quando há dados de magnitude variada) a normalização acaba por tornar a execução do algoritmo SVC mais rápida e sem incidência de erros. O intervalo utilizado foi de [0,1], conforme sugerido na mesma referência.

Importante destacar que a normalização foi executada em duas etapas distintas. A primeira etapa foi executada para as informações textuais que foram transformadas em informações numéricas e, devido à baixa amplitude de valores, estes foram submetidos à normalização usual por extremos (abordada no capítulo 2). Esta modalidade consiste em dividir os valores existentes por faixa de valor mínimo e máximo, de modo a se obter valores variando somente de [0,1].

Por outro lado, a normalização dos dados originalmente numéricos (relacionados à duração das anormalidades) foi efetuada em duas etapas (por partes), de maneira diferente da normalização executada sobre os dados numéricos provenientes de informações textuais. Esta abordagem foi seguida, pois os valores dos dados brutos de duração das anormalidades possuíam uma grande amplitude e com distribuição não uniforme, conforme evidenciado no histograma da Figura 39.

4.4. Balanceamento da base de dados

As classes do banco de dados utilizado estavam altamente desbalanceadas e uma das propostas foi de balancear a base de dados inicial. Com isso altera-se o número de observações submetidas como dado de entrada do classificador, de modo a reduzir o impacto do desbalanceamento. Como já foi abordado na última seção, o método de classificação hierárquico seria outra maneira de tratar esta particularidade do banco de dados.

De acordo com o que foi levantado na seção específica sobre balanceamento da base de dados (Seção 2.3.4.3), foram utilizados métodos para redução de classes numerosas bem como métodos para aumento de classes de menor tamanho.

4.5. Classificador SVC

Nesta seção será abordado como se deu o processo de classificação das anormalidades por meio de SVCs. Será abordada a divisão da base de dados em

treinamento/validação e teste. Em seguida serão tratados os classificadores de etapa única e de mais de uma etapa. Finalmente serão abordadas as métricas para seleção das melhores configurações. A Figura 40 traz o diagrama desta etapa do processo. Vale observar que a escolha do melhor modelo na fase de treinamento e validação, bem como a avaliação do desempenho final, foram balizadas por meio de métricas apontadas ao final do Capítulo 2.

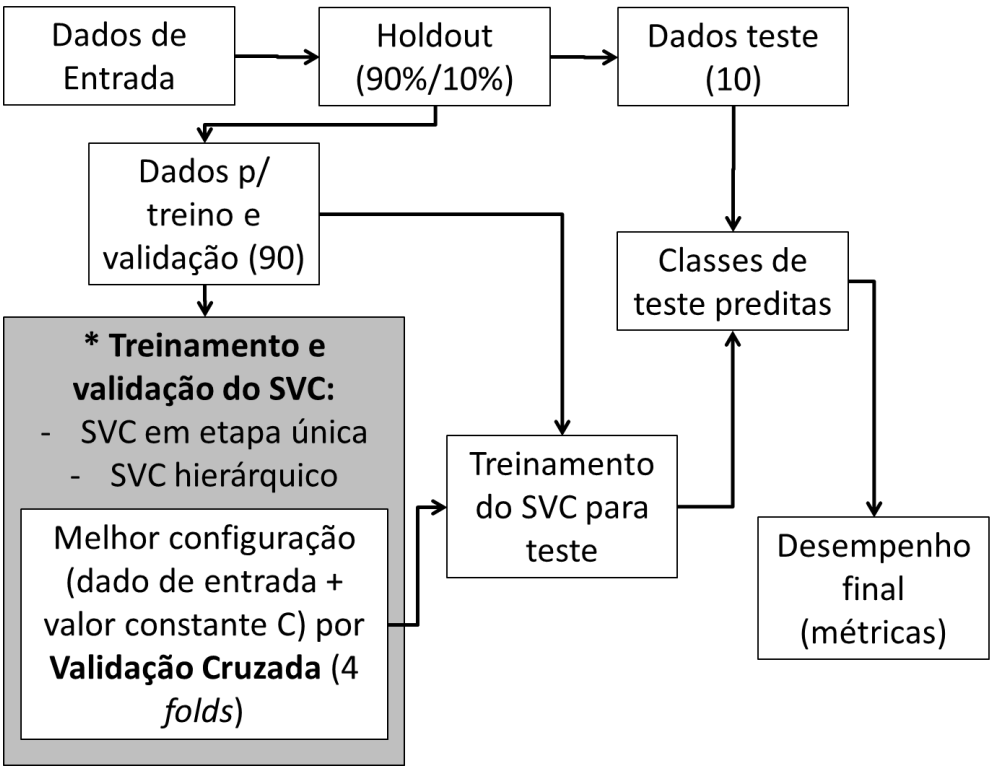

Figura 40: Diagrama da etapa de classificação dos documentos.

As próximas seções detalham o diagrama da etapa de classificação.

4.5.1. Divisão da base de dados em treinamento, validação e teste

Para as etapas de treino, validação e teste do modelo foram utilizadas as técnicas de *Holdout* e Validação cruzada (*Cross Validation*) em conjunto. Inicialmente valeu-se do *Holdout* para separar desde o início as observações utilizadas para treino e validação das observações utilizadas para o teste final, em uma proporção de 90% e 10% respectivamente. Em seguida, utilizou-se a

validação cruzada sobre os dados de treino e validação para selecionar a melhor disposição de dados de entrada (corpus, esparcidade, peso dos termos) em conjunto com os melhores parâmetros de SVC. Vale destacar que o balanceamento da base de dados também altera os dados de entrada. Porém, como sua alteração é mais elaborada, foram separados estudos de caso para cada balanceamento proposto.

Após obter a melhor configuração na etapa preliminar de treino e validação, submeteu-se o classificador ao teste final, obtendo o desempenho de sua configuração para a fase de teste para o estudo de caso.

Após repetir este processo para vários estudos de caso, foi possível comparar as melhores configurações dos modelos utilizando os resultados dos seus respectivos testes finais.

Para um primeiro estudo de caso, na etapa preliminar da validação cruzada (treino e validação), foram efetuadas as K iterações para todas as variações do modelo dentro do estudo de caso. Esta etapa preliminar teve como resultado o desempenho médio E^j (j variando de 1 até o número de configurações de modelos avaliados dentro do estudo de caso) de cada configuração analisada, conforme abordado na Seção 2.3.4.1. Ao final, escolheu-se o melhor desempenho médio da etapa preliminar de treinamento e validação, finalizando a etapa de validação cruzada.

Este modelo foi submetido ao teste final, de forma a se obter seu respectivo desempenho de teste. Vale destacar que o treinamento da melhor configuração do classificador antes do teste final foi feito com toda a base de dados utilizada na etapa de treino e validação, ou seja, todos os *folds* (frações) são consolidados para treinar o melhor classificador, de modo que este possa efetuar a classificação para o teste final.

4.5.2. Classificação em etapa única por SVC

Conforme abordado no capítulo 2, por recomendação das referências bibliográficas, utilizou-se na configuração básica dos SVCs o *Kernel* linear, que alcança robusto desempenho e não possui parâmetros adicionais de ajuste. As

referências apontaram que, para mineração textual pelo método de representação dos dados textuais pela matriz de termos por documentos, o *Kernel* linear desempenha tão bem quanto os *Kernels* mais elaborados, com o ganho da simplificação do algoritmo.

Neste cenário, o parâmetro C (*penalty parameter*) é a única constante a ser alterada para encontrar o SVC de melhor desempenho. O ajuste sugerido deste parâmetro, conforme (Hsu, Chang, Lin, 2010) é de variar o valor da constante de forma exponencial, como por exemplo: 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2} ,...,2^{5},2^{6} e assim sucessivamente.

Após encontrar o desempenho médio (E^{j}) na etapa de treinamento e validação para cada valor de C, constrói-se o gráfico ilustrado na Figura 41 para verificar pontos de inflexão da curva, para cada configuração de cada modelo. Em seguida, um ajuste fino é efetuado nestas regiões, para as configurações mais promissoras. Atentar que cada curva (representada na figura por uma cor diferente) representa uma configuração fixa de dado de entrada, de forma que foi variado somente o valor da constante C:

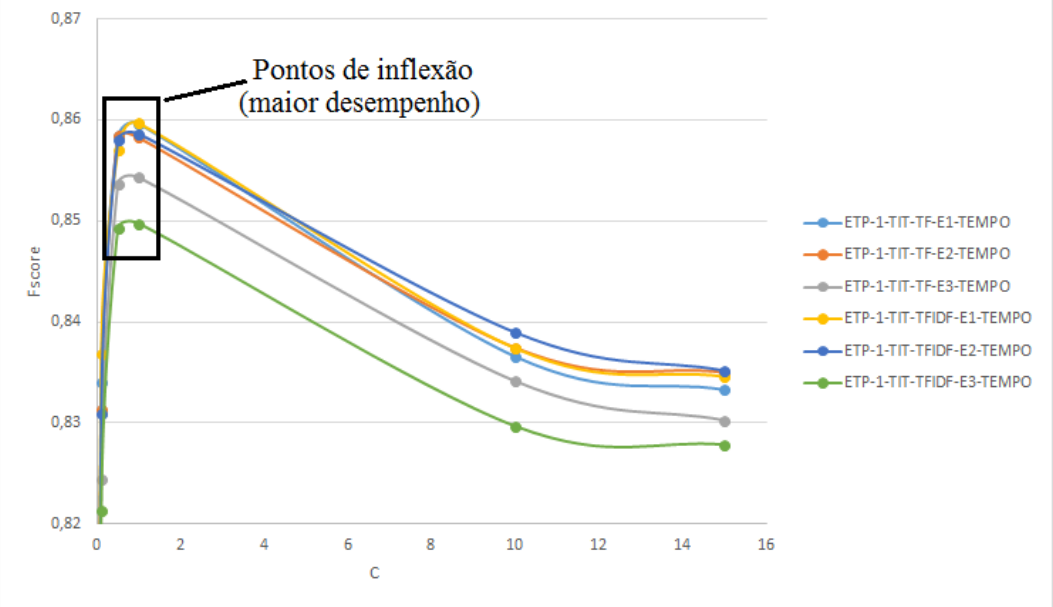

Figura 41: Exemplo de heurística para a constante de regularização (C).

4.5.3.
Classificação em mais de uma etapa com SVC

Durante a elaboração e análise dos resultados verificou-se, por meio da matriz de confusão dos melhores classificadores, uma baixa taxa de acerto para as classes menores (principalmente as classes de 5 até 11). Uma maneira de contornar o acentuado desbalanceamento das classes foi efetuar uma classificação por etapas, ou classificação hierárquica. Vale destacar que também foi previsto efetuar o balanceamento destas classes menores (assunto a ser tratado no próximo capítulo).

O objetivo da classificação hierárquica é utilizar mais de um classificador ao invés de somente um classificador generalista e dividir o problema em várias partes. Para este trabalho utilizaram-se dois classificadores. Em uma primeira etapa, o primeiro classificador apontou somente as quatro primeiras classes (maiores) e as demais classes foram todas reunidas somente em uma classe “R”. Este primeiro classificador foi treinado com base nesta nova divisão de (4+1) classes.

As classes apontadas pelo primeiro classificador como classe “R” foram submetidas a um segundo classificador que foi treinado somente com observações destas classes menores. O esquema da Figura 42 resume a abordagem proposta:

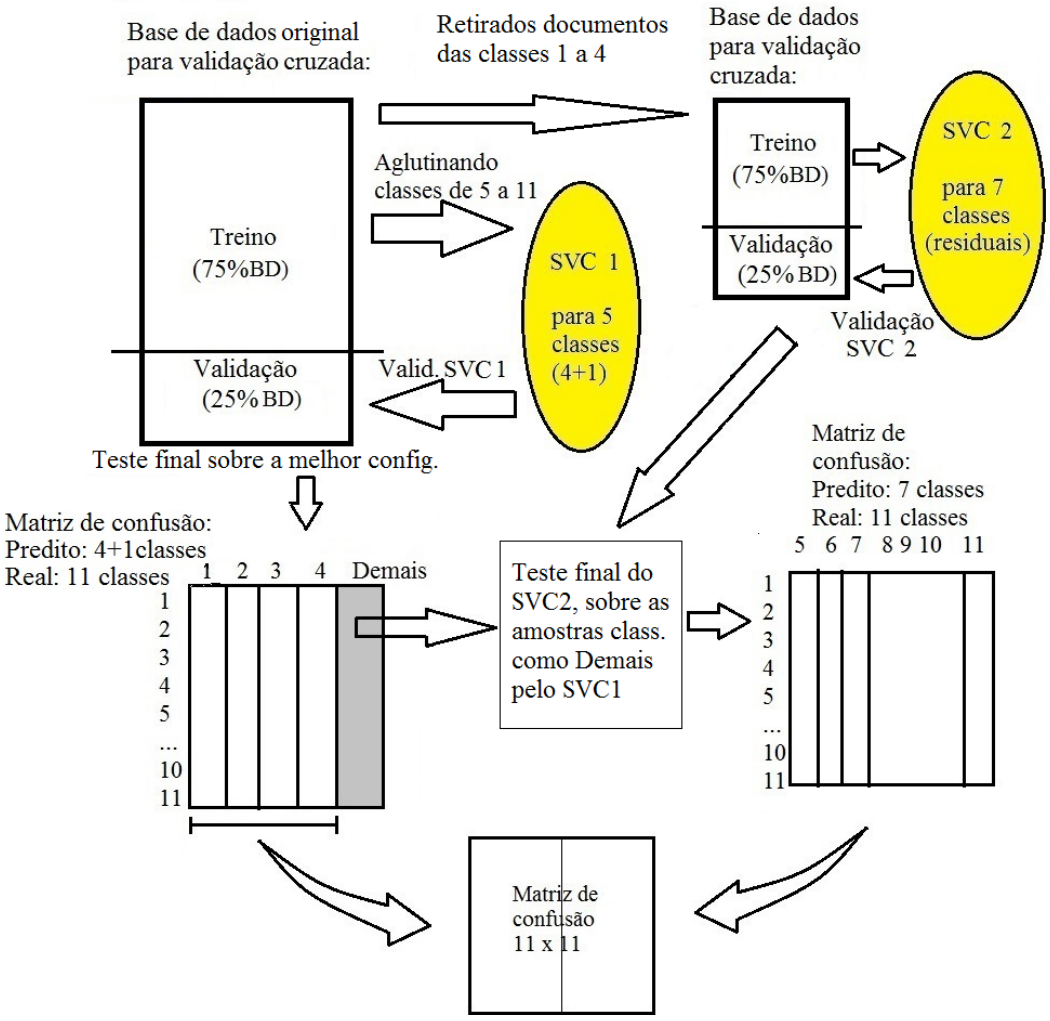

Figura 42: Esquema proposto para classificação hierárquica.

O passo a passo do esquema proposto será detalhado no Capítulo 5, mas a base de treino e validação da Figura 42 já considera os 90% de toda a base de dados, separada por *hold out* (os demais 10% são separados para teste). Ainda com relação à Figura 42, o valor relacionado a 75% da base de dados se refere às observações que foram utilizadas para treino e validação (correspondente a 3 dos 4 folds da técnica de validação cruzada). Os demais 25% correspondem ao *fold* restante.

4.5.4.
Métricas de avaliação

Foram selecionadas três métricas para avaliar o desempenho das configurações do modelo: micro F-score, macro F-score bem como a média de acerto percentual das classes. Conforme apontado, como a base de dados é altamente desbalanceada, o uso do macro F-score permite avaliar melhor as configurações com bom desempenho tanto nas classes majoritárias quanto nas minoritárias, diferentemente do micro F-score, que geralmente privilegia o desempenho das classes de maior número.

5. Estudo de casos para o classificador de anormalidades de falha de equipamento de unidade de intervenção

5.1. Ajustes preliminares

Antes de serem iniciadas as avaliações das diversas configurações do modelo proposto (estudos de caso), alguns pontos de ajuste bem como o detalhamento de aplicação dos algoritmos foram realizados. Em seguida foram abordados os diversos estudos de caso.

Todo o processamento se baseou em um computador Intel Core i7 2,9 GHz com 16 GB de memória RAM.

Como valores aproximados, do tempo dispendido nas etapas de obtenção dos resultados deste trabalho, em torno de 60 % foi utilizado para a etapa de pré-processamento e na correção das bases de dados. Em seguida, em torno de 25 % deste tempo foi utilizado para o processamento dos modelos. Os demais 15 % foram utilizados para analisar os dados de saída.

5.1.1. Software R

A linguagem R (R Core Team, 2013) foi criada por Ross Ihaka e Robert Gentleman em 1993, na Universidade de Auckland, na Nova Zelândia. Este ambiente de programação provê base para o desenvolvimento integrado de cálculos e análises. Possui expansão de suas funcionalidades através do uso de pacotes, que são bibliotecas para funções avançadas e específicas, disponibilizados através de uma rede de distribuição do R (CRAN em inglês).

Como o código de programação é aberto, os pacotes são construídos de modo colaborativo e, com isso, além de não ter custo de aquisição, possui grande abrangência e robustez na execução das tarefas de análise de dados.

Esta foi a linguagem de programação utilizada neste trabalho. A versão aplicada foi a 3.1.2.

5.1.1.1. Pacotes utilizados

O ambiente de programação em R possui funções fundamentais, de modo a prover a base para a elaboração dos trabalhos. Porém, quando tarefas específicas são demandadas, faz-se necessária a instalação de pacotes adicionais. Esta seção irá tratar os pacotes externos utilizados neste trabalho.

5.1.1.1.1. Tm

O pacote Tm (versão 0.7-1) foi desenvolvido por (Feinerer, Hornik, 2015), (Feinerer, Hornik, Meyer, 2008). É um pacote básico para mineração textual, via método *bag of words* (matriz de termos por documentos). Ele desempenha importação de dados textuais, manipulação do corpus, pré-processamento, além de gerenciamento adequado dos dados (nas grandes matrizes esparsas). O pacote possui as seguintes funções internas:

- Retira espaços em excesso entre as palavras do texto;
- Altera as palavras, de modo a remover toda a pontuação (pontos de interrogação, exclamação, vírgulas) bem como acentos gráficos e etc.;
- Remove todos os números da coleção;
- Remove uma lista pré-determinada no pacote de preposições e outras palavras de pouco significado para a língua portuguesa (lista base de *stop words*). Também permite o usuário inserir outras palavras que este julgue como pouco relevantes (lista customizada de *stop words*);
- Faz conexão com outros pacotes para efetuar a radicalização das palavras da coleção utilizando o pacote SnowballC neste trabalho (a ser abordado mais adiante);

- Transforma todo o corpus na matriz de termos por documentos.
    - Preenchimento dos pesos na matriz: após criação da estrutura matricial, os pesos dos termos são automaticamente preenchidos como TF (*term frequency*), ou seja, o pacote faz a contagem das palavras dentro de cada documento. Caso se necessite alterar o peso, há outras opções: aplicação nos termos do peso TF-IDF ou aplicação dos pesos de forma binária;
- Redução de termos esparsos dentro da matriz construída.

5.1.1.1.2. SnowballC

O pacote SnowballC (versão 0.5.1) implementa a radicalização dos termos, segundo o algoritmo de Porter. Este pacote é acionado pelo pacote Tm para efetuar a radicalização para língua portuguesa, pois o pacote Tm em si não possui tal opção. Este pacote foi desenvolvido por (Bouchet-Valat, 2015).

5.1.1.1.3. e1071

O pacote e1071 (versão 1.6-4) desenvolvido por (Meyer et al, 2015) é a conjunção de vários algoritmos de mineração de dados e análise de dados. Este pacote faz interface com o pacote em linguagem C++ denominado LibSVM, que foi desenvolvido por (Chang, Lin, 2011). O LibSVM implementa o algoritmo SVM (*Support Vector Machines*) para uma gama de aplicações. No caso das SVMs utilizadas para classificação (SVCs), o pacote possui duas funções básicas:

- Treinamento do SVC;
- Teste do SVC;

A função de treinamento permite normalizar os dados de entrada, bem como selecionar o *Kernel* e o parâmetro do SVC, a depender da configuração utilizada.

Conforme apontado no capítulo 2, o *Kernel* utilizado será o linear e para a seleção do parâmetro C será utilizada a heurística também apontada no capítulo 2. Foi arbitrado para cada rodada do modelo a avaliação de 6 valores do parâmetro C. O método utilizado neste pacote para o problema de classificação com mais de duas classes por meio de SVCs é o método “um contra um”.

5.1.1.1.4. Unbalanced

Desenvolvido por (Pozzolo, Caelen, Bontempi, 2015), este pacote é utilizado para efetuar o balanceamento de classes minoritárias. A versão utilizada foi a 2.0. As seguintes funções foram utilizadas:

- Aumento de uma classe minoritária através do algoritmo SMOTE;
- Aumento de uma classe minoritária pela geração aleatória de observações;
- Remoção dos dados da classe majoritária pelo algoritmo *Edited Nearest Neighbor* (ENN);
- Redução de uma classe majoritária pela retirada aleatória de observações;
- Remoção dos dados da classe majoritária pelo algoritmo *Neighborhood Cleaning Rule* (NCL);
- Remoção dos dados da classe majoritária pelo algoritmo *Tomek Links*.

5.1.1.1.5. CARET

Pacote desenvolvido por (Kuhn, 2016) com contribuição de vários autores, este pacote é a reunião de vários outros pacotes e ferramentas para análise e tratamento de dados. Uma destas ferramentas auxilia na criação das frações da base de dados (*folds*) para efetuar o treinamento por validação cruzada. Importante destacar que a geração dos *folds* por este pacote é sempre estratificada, ou seja, as classes são distribuídas dentro de cada fração de acordo com a

proporção que estão presentes na base de dados completa. Também foi utilizada esta função para separar os dados da etapa de treino/validação da etapa de teste (*holdout*), de modo a também contemplar uma base de teste bem distribuída. A versão utilizada foi a 6.0-52.

5.1.2. Base de dados

Conforme abordado no Capítulo 3, a base de dados consiste em 9202 anormalidades relativas às falhas de equipamentos de unidades de intervenção (sondas). Estas foram extraídas da base de dados Petrobras, alimentada via software *Open Wells* (Landmark, 2017) e consistem nos registros efetuados com base na nova classificação (estabelecida após junho de 2011). Estes dados serão utilizados para treinamento, validação e teste dos modelos. A Tabela 7 ilustra os tipos de dados disponíveis:

Tabela 7: Tipos de dados disponíveis para a análise.

Campo Disponível	Tipo de dado
Título	Textual
Descrição	Textual
Duração	Numérico
Detalhamento nível 1	Categórico (1 classe)
Detalhamento nível 2	Categórico (11 classes)

O título e descrição da anormalidade serão tratados de forma a gerar a matriz de termos por documentos. O campo duração, relacionado ao intervalo de tempo de duração da anormalidade em horas, possui fração mínima de 0,5 horas (30 minutos). Este campo será tratado, normalizado à parte e posteriormente acoplado à matriz de termos por documentos resultante dos dados textuais.

Conforme apontado no capítulo 3, a variável “detalhamento nível 1” foi útil durante o tratamento inicial dos dados, de modo a possibilitar a separação das falhas de equipamentos de sonda das demais falhas de equipamentos. Porém, nesta etapa de classificação ele não será utilizado. Por outro lado, as 11 subclasses (“detalhamento nível 2”) relacionadas às falhas de equipamentos de sonda foram as classes utilizadas nas rodadas do modelo.

5.1.3.
Normalização dos dados de entrada

O único dado originalmente numérico, a duração das anormalidades, foi submetido ao processo de normalização. Conforme abordado no capítulo 4, foram avaliados os dados disponíveis para a duração das anormalidades por meio de um histograma:

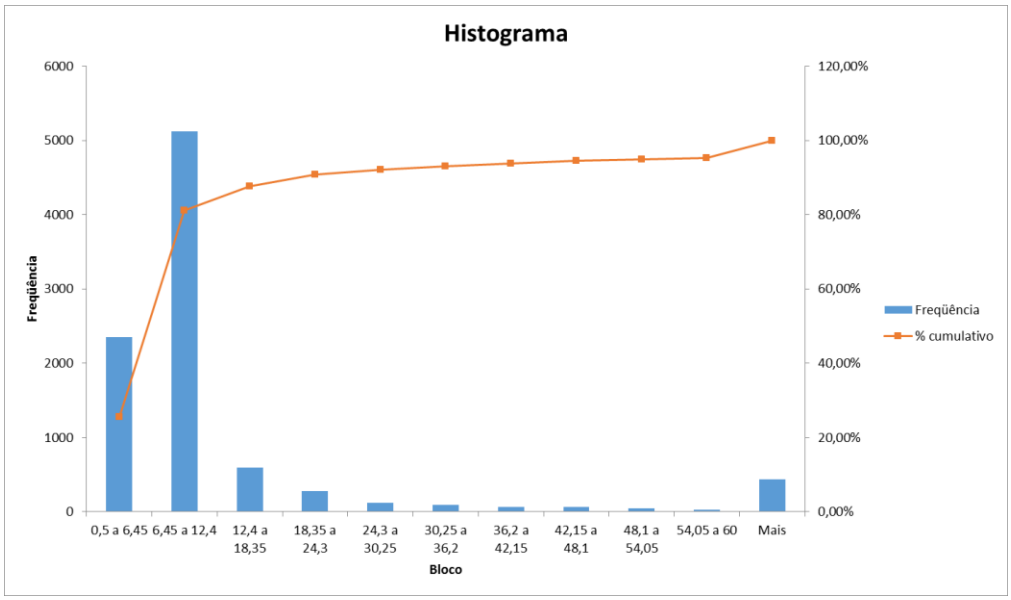

Figura 43: Histograma da duração das anormalidades.

A distribuição da frequência da duração das anormalidades é concentrada basicamente em valores de até 12 horas. Normalizar as variáveis de [0,1] sem nenhum tratamento prévio iria reduzir bastante os valores destas durações de valor mais baixo, pois o valor máximo possui magnitude elevada. Com isso, considerando que aproximadamente 80% dos dados numéricos de duração estão no intervalo de 0,5 a 12 horas, efetuou-se a normalização por partes da seguinte forma:

- 0,5 horas (mínimo) até 12,5 horas (máximo), resultando em valor de 0 a 0,8 (incluindo 0,8), com a seguinte formulação:
-

$$X_{i,\ 0\ a\ 0,8} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

Equação 14

Onde X_{min} = 0,5 (mínimo) e X_{max} = 12 horas (máximo), incluindo este valor, resultando em valor X_i, de 0 a 0,8 (incluindo 0,8);

- 12,5 horas (mínimo) até o valor máximo, resultando em valor de 0,8 a 1 (não incluindo 0,8), com a seguinte formulação:

$$X_{i,\ 0,8\ a\ 1} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

Equação 15

Onde X_{min} = 12,5 horas (mínimo) até o valor máximo (X_{max}), resultando em valor X_i, de 0,8 a 1 (não incluindo 0,8).

A inserção destes dados seria feita por meio de uma coluna extra, posteriormente anexada à matriz de termos por documentos já normalizada, de modo agregar mais uma característica ao modelo.

Importante destacar que os dados da matriz de termos por documentos foram normalizados em etapa distinta desta, pelo método de normalização por extremos em etapa única (no mesmo intervalo de [0,1]), devido à sua distribuição ter amplitude de valores menor.

5.1.4. Redução de esparcidade – testes preliminares

Foi efetuada uma análise de sensibilidade para a escolha do valor a ser atribuído para o redutor de esparcidade (conforme apontado no capítulo 4). A

Tabela 7 ilustra a quantidade de termos quando não há aplicação do redutor (destacado na cor cinza escuro da tabela), comparando-se com outros valores utilizados para o redutor (ver Seção 2.3.3.4 para maiores detalhes sobre a redução de esparcidade).

Tabela 8: Análise de atuação do redutor de esparcidade para os campos de título e de descrição detalhada das anormalidades.

Título	Redutor de esparcidade	1	0,9998	0,9995	0,99888
	Total de termos	2303	1320	780	418
	Células vazias na MTD	46713	45730	44069	42092
Descrição detalhada	Redutor de esparcidade	1	0,9998	0,9995	0,99888
	Total de termos	5040	2918	1746	1107
	Células vazias na MTD	103514	101392	98250	93795

Após alguns testes preliminares, observou-se que a redução além do fator 0,9995 impactou negativamente nos resultados. Por isso o redutor será utilizado até este valor.

5.1.5. Divisão da base de dados para treinamento, validação e teste

A escolha do número de frações K (*folds*) foi altamente impactada pela distribuição das classes na base de dados. Como o desbalanceamento é acentuado, optou-se por separar 90% da base de dados para a etapa inicial de treinamento e validação (total de 8280 observações), e os demais 10% para teste (total de 922 observações). Para a etapa preliminar de treinamento e validação, a quantidade de 10 *folds* tem sido o valor usual sugerido nas referências para valor de K. Porém para este trabalho escolheu-se um valor menor para separar as frações (*folds*) na validação cruzada, sendo utilizado K = 4 (2070 observações por *fold*). A Tabela 9 mostra a distribuição por classe na base de dados para todas as etapas de escolha dos modelos:

Tabela 9: Distribuição dos documentos por classe para cada etapa de avaliação do modelo.

Classe	Base de dados	Base de treino/val.	Base por fold	Base de teste
1	3077	2774	694	303
2	1607	1456	364	151
3	1576	1414	354	162
4	1522	1370	343	152
5	432	389	97	43
6	266	236	59	30
7	209	181	45	28
8	187	167	42	20
9	172	156	39	16
10	142	127	32	15
11	12	10	3	2

O valor de K = 4 foi justificado pela disposição das classes na base de dados, de modo que o uso de valor usual de K = 10 iria dividir as classes de menor tamanho em parcelas cada vez menores, o que prejudicaria a análise na etapa de treinamento e validação.

Como a divisão da base de treinamento e validação para os quatro *folds* não é exata, o algoritmo CARET (pacote R), que faz essa separação, ajusta alguns valores, de forma que ao se multiplicar por quatro os valores da coluna “Base por *fold*” não se encontrará o mesmo valor da coluna relativa à base de treinamento/validação.

5.1.6. Aplicação do balanceamento da base de dados

As classes do banco de dados utilizado estão fortemente desbalanceadas e uma das propostas (além da abordagem de classificação hierárquica abordada no capítulo anterior) foi efetuar testes reduzindo a numerosa classe 1 a um tamanho próximo à classe 2 (ou seja, de 3077 para aproximadamente 1600 documentos), bem como aumentando as classes de 6 a 11, que possuíam poucas observações, de modo a se mitigar o problema de balanceamento da base. Esse aumento das classes menores se deu de forma a torná-las de porte semelhante ao da classe 5 (que possui 432 documentos).

Os estudos de caso a princípio tratam da redução da classe majoritária 1. Em seguida há outro estudo somente para o aumento das 6 classes menores. Após encontrar o método com melhor desempenho para cada estudo de caso, os mesmos foram testados em conjunto.

5.1.7. Adequação das métricas de avaliação

Conforme apontado no capítulo 4, três métricas foram selecionadas para avaliar o desempenho das configurações do modelo: micro F-score, macro F-score e a média de acerto percentual das classes.

Porém, à medida que os modelos foram testados, observou-se erro nos cálculos da métrica macro F-score, pois um de seus denominadores resultava em valor zero. A Figura 44 apresenta um exemplo deste problema. Suponha que no caso evidenciado na figura, o modelo não classifique nenhuma observação como pertencente à classe 3. Neste caso, os valores de **fp3** e de **tp3** alcançariam valor zero.

		Predito pelo modelo			
		1	2	3	4
Classe real	1	tn 3	fn 3	fp3	fn 3
	2	fn 3	tn 3	fp3	fn 3
	3	fn 3	fn 3	tp 3	fn 3
	4	fn 3	fn 3	fp3	tn 3

Figura 44: Exemplo de matriz de confusão para cálculo de F-score.

Conforme a formulação abordada no capítulo 2, o valor calculado de precisão macro (M) para a classe 3 teria como denominador o valor zero, implicando em erro no cálculo do macro F-score (que depende do cálculo da precisão macro – (M)). Neste trabalho, em particular, algumas vezes o classificador não aponta nenhum padrão de entrada para as classes com poucas observações (a classe 11 possui 12 observações, por exemplo), causando o erro mencionado.

Portanto, apesar de a literatura indicar o uso do macro F-score, o cálculo do mesmo não é viável em um cenário de base altamente desbalanceada (vários cenários neste trabalho se deparam com este problema).

Como forma alternativa, foram efetuados testes de forma que quando o denominador resultasse em valor nulo, zerava-se o valor da fração para a referida classe. Porém esta medida penalizou severamente os modelos onde isso ocorria.

Por este motivo, escolheu-se trabalhar com as demais métricas: o micro F-score como métrica principal de desempenho e como métrica auxiliar foi utilizada a média de acerto percentual das classes (ao invés do macro F-score).

5.1.8. Legendas

Foram estipuladas legendas para melhor visualização das tabelas de resultados. Essas tabelas foram ordenadas de acordo com o desempenho obtido pela métrica escolhida (no caso, a métrica micro F-Score). Os modelos mais bem avaliados dentro de cada estudo de caso tiveram suas matrizes de confusão analisadas.

A Tabela 10 ilustra como foram dispostos os parâmetros e características das configurações dos modelos, bem como seu desempenho:

Tabela 10: Exemplo de apresentação dos resultados.

Modelo	Legenda	C	Desempenho (métrica)
1	TIT-TF-E1-TEMPO	0,5	0,9
2	OCOR-BIN-E2-TEMPO	0,1	0,85
3	OCOR-TFIDF-E3-S/ TEMPO	10	0,8
4	TIT-TF-E3-TEMPO	0,01	0,75
5	...	...	...

Cada linha representa uma configuração do modelo submetido à etapa de treino e validação. O valor médio de desempenho (E^j) encontrado será disponibilizado na última coluna. Abaixo a explicação das legendas:

Para o corpus utilizado:

- TIT: utilizado campo do título das anormalidades

- OCOR: utilizado o campo de descrição detalhada das anormalidades.

Para o peso utilizado na transformação dos dados:

- TF: frequência dos termos no documento.
- BIN: presença (1) ou não (2) do termo no documento.
- TFIDF ou somente IDF: peso do termo considerando sua frequência local e global (TF-IDF).

Para o fator de redução de esparcidade:

- E1: sem redução de esparcidade.
- E2: redutor utilizado de 0,998.
- E3: redutor utilizado de 0,995.

Para a duração da anormalidade:

- TEMPO: utilizado campo de duração;
- S/ TEMPO: não utilizado campo de duração.

Isso totaliza 36 tipos diferentes de dados de entrada a serem avaliados nos estudos de caso.

Para a escolha do parâmetro de penalidade do SVC:

- C: variando conforme heurística.

Ao final de cada estudo de caso foram elencadas as configurações com maior desempenho, em seguida a melhor configuração foi submetida ao teste final. Finalmente foi avaliada a matriz de confusão desta configuração. Este processo se repetiu para cada estudo de caso.

5.1.9. Taxa de acerto do operador

Como forma de estabelecer o mínimo desempenho esperado para o classificador (*base line)*, selecionaram-se 461 observações da base de treinamento/validação para que um especialista avaliasse a taxa de acerto do operador. Este valor também permite ter uma prévia dos possíveis erros a serem observados nas classificações utilizando inteligência artificial. Vale destacar que

todas as observações classificadas de forma incorreta pelo operador (após serem analisadas pelo especialista) tiveram suas classes alteradas na base de dados de treinamento/validação.

Todo processo está sujeito a falhas e esta análise colabora para um melhor entendimento de quais classes o operador possui mais dificuldade na sua atribuição. As Figuras 45 e 46 ilustram as matrizes de confusão elaboradas para o teste efetuado do operador:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	127	0	0	22	5	2	15	2	0	1	0
	2	3	49	7	0	1	0	0	0	0	0	0
	3	1	1	40	2	9	1	0	0	0	0	0
	4	8	5	3	91	0	0	0	2	1	0	0
	5	2	0	1	2	12	0	2	1	0	0	0
	6	0	0	0	0	0	12	0	0	0	0	0
	7	7	0	0	0	0	0	3	0	0	0	0
	8	3	0	0	0	2	0	0	4	0	0	0
	9	1	0	0	0	1	0	0	0	8	0	0
	10	0	0	0	0	0	0	0	0	0	1	1
	11	0	0	0	0	0	0	0	0	0	0	0

Figura 45: Matriz de confusão do operador (valores absolutos).

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	72,99%	0,00%	0,00%	12,64%	2,87%	1,15%	8,62%	1,15%	0,00%	0,57%	0,00%
	2	5,00%	81,67%	11,67%	0,00%	1,67%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	1,85%	1,85%	74,07%	3,70%	16,67%	1,85%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	7,27%	4,55%	2,73%	82,73%	0,00%	0,00%	0,00%	1,82%	0,91%	0,00%	0,00%
	5	10,00%	0,00%	5,00%	10,00%	60,00%	0,00%	10,00%	5,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	70,00%	0,00%	0,00%	0,00%	0,00%	0,00%	30,00%	0,00%	0,00%	0,00%	0,00%
	8	33,33%	0,00%	0,00%	0,00%	22,22%	0,00%	0,00%	44,44%	0,00%	0,00%	0,00%
	9	10,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	80,00%	0,00%	0,00%
	10	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
	11											
			Média das classes			67,59%	Micro-Fscore		35,63%			

Figura 46: Matriz de confusão do operador (valores percentuais por classe).

O micro F-score do operador foi de 35,63% e a média de acerto entre classes foi de 67,59%. Este será o desempenho base do trabalho (*base line)*. Importante observar que a classe 1 (mais numerosa) é confundida com várias classes, principalmente com a classe 4 e com a classe 7. Outro ponto importante é que o

operador somente acertou em sua totalidade a classe 6 e teve baixa taxa de acerto para as classes 7, 8 e 10. Não havia elementos da classe 11 neste teste.

5.1.10. Correção da base

A aplicação dos classificadores a uma base de dados já rotulada assume o fato de que esta classificação foi apontada de maneira correta em todas suas observações. Porém, conforme evidenciado no teste realizado com o operador, observa-se dificuldade do mesmo em apontar algumas classes. Isso traz indícios de que a base de dados utilizada neste trabalho também possua erros semelhantes de classificação.

De forma a contornar este problema, seria ideal a correção de toda a base de dados, mas tal tarefa seria considerada inviável pelo tempo a ser dispendido. Deste modo, adotou-se a seguinte medida: após obter a melhor configuração de modelo, após todos os estudos de caso, fez-se uma análise dos erros deste classificador de melhor desempenho, via matriz de confusão. Suspeitou-se que classes muito confundidas pelo classificador poderiam ser, na verdade, erros de apontamento do operador. Isto se baseia no fato de que o classificador recebe como entrada basicamente informações textuais condensadas e transformadas em uma matriz de termos por documentos. Supondo que o erro do operador seja baixo, infere-se que a maioria dos documentos possui classificação adequada, o que irá influenciar positivamente no desempenho do classificador. Se, por exemplo, a classe W estava apontada de forma incorreta no documento como classe Z, o modelo tenderá a classificar este mesmo documento como W (apesar de possuir um rótulo incorreto Z).

Todo este processo é próximo de uma classificação manual, porém os dados apontados como erro pelo SVC (classificador) trouxeram direcionamento e mais objetividade nestas correções. O detalhamento das correções será feito nas próximas seções.

Após esta etapa de correção das classes todo o processo foi repetido, a fim de calibrar novamente o classificador com as classes apontadas, agora, corretamente. Este longo trabalho de conferência dos dados, apesar de demorado, trouxe

informações importantes para melhoria no processo de classificação de anormalidades, de forma a subsidiar o treinamento dos operadores, além de colaborar para o desempenho superior do classificador.

5.2. Estudos de caso sem correção da base de dados

Os primeiros estudos de caso consideraram a base de dados como estava disposta, assumindo que as classes apontadas eram todas verdadeiras. Posteriormente correções foram efetuadas. Para esta primeira fase foi avaliado se a duração agregava ao processo. Em seguida, aplicou-se a abordagem de classificação hierárquica. Posteriormente, foram efetuados os balanceamentos da base de dados (redução da classe 1 e aumento das classes de 6 a 11, tanto em separado como concomitantemente). Finalmente, utilizaram-se todos os métodos em conjunto para obter o melhor modelo.

5.2.1. Estudo de caso 1: caso base

Para o primeiro estudo de caso, foram ajustadas as configurações básicas do modelo, sem alterar a base de dados tão pouco o método de classificação. Com isso, avaliou-se o resultado de um caso base para comparação com os demais estudos de caso. Seguem os parâmetros variados dentro deste estudo de caso:

36 combinações de configuração para os dados de entrada:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):
    - Termos para título (E1, E2 e E3): 2303, 1320 e 708;
    - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;
- (2) - Duração: presente ou não na matriz final;
- (2 x 6 = 12) - Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada. Ou seja, efetua-se um ajuste inicial com seis valores da

constante C. Em seguida, após analisar o gráfico de desempenho das configurações, efetua-se um ajuste fino nos entornos do ponto de inflexão de maior desempenho.

A Figura 47 mostra os primeiros resultados obtidos após o primeiro ajuste para obtenção do valor da constante C que traz o melhor desempenho. Cada linha do gráfico é uma configuração de entrada do modelo, alterando o valor da constante C. São avaliados os valores em torno das inflexões destas curvas. Das 36 diferentes configurações avaliadas a Figura 47 apresenta 14 delas apenas para ilustrar o processo de busca pelo melhor valor do parâmetro C.

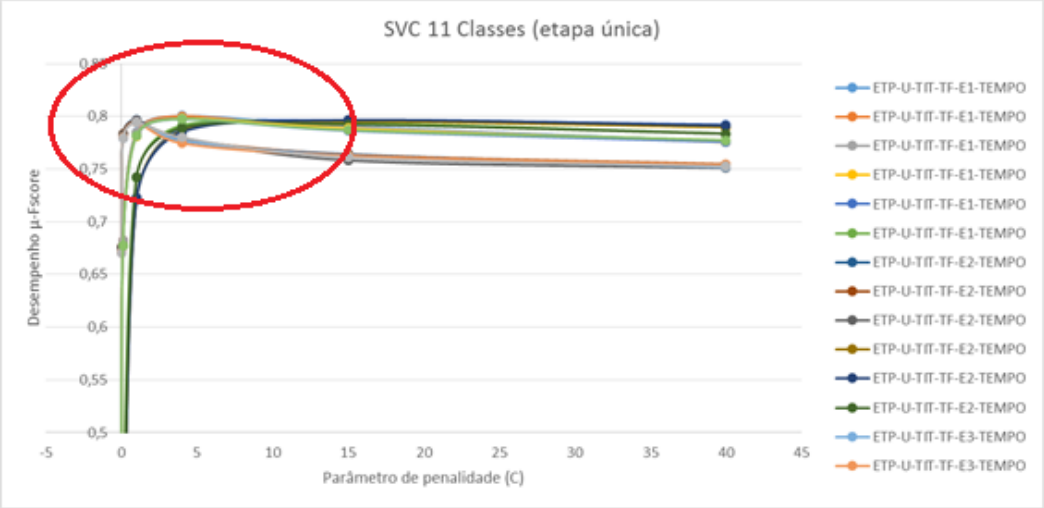

Figura 47: Curvas de desempenho para o primeiro estudo de caso. Círculo em vermelho destaca as inflexões das curvas.

Em seguida, restringe-se o intervalo em torno dos melhores resultados (próximos aos pontos de inflexão, conforme destacado em vermelho na Figura 47). Utilizando 6 novos valores para a constante C efetua-se o ajuste fino nos entornos dos pontos de inflexão das configurações de melhor desempenho, (conforme evidenciado na Figura 48) para as 14 configurações (de um total de 36) de forma a ilustrar o processo de obtenção do melhor desempenho (para o caso abaixo, os melhores desempenhos estão próximos dos valores 1 e 4, para a constante C):

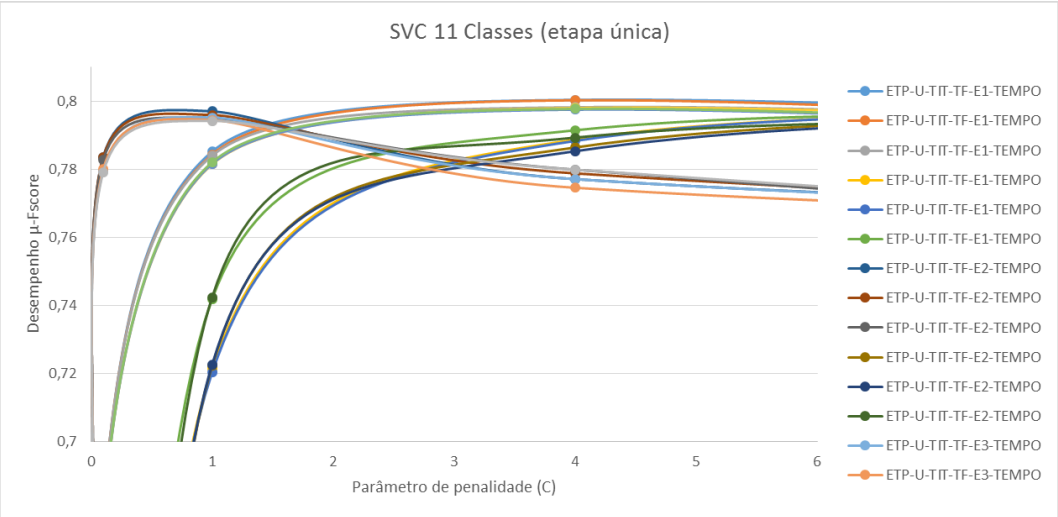

Figura 48: Visão ampliada das curvas de desempenho para primeiro estudo de caso.

Após esta etapa, ordenam-se os resultados do melhor ao pior desempenho de acordo com a métrica micro F-score (legendas: ETP-U = Etapa única, ou seja, 1 SVC classifica 11 classes), conforme a Tabela 11:

Tabela 11: Resultados da etapa de treinamento e validação para o primeiro estudo de caso, após ajuste fino do parâmetro de penalidade do SVC.

Dado de entrada	C	Micro F-score
ETP-U-TIT-TF-E1-TEMPO	3	53,91%
ETP-U-TIT-TF-E2-TEMPO	3	53,85%
ETP-U-TIT-TF-E1-TEMPO	4	53,82%
ETP-U-TIT-TF-E2-TEMPO	5	53,81%
ETP-U-TIT-TF-E1-SEM TEMPO	3	53,78%

Não serão apresentados todos os resultados, pois o total de variações é de 432 modelos avaliados (36 variações de entrada testados em 6 valores da constante C na primeira avaliação e mais 6 valores da constante para efetuar o ajuste fino, de modo que há {(36 x 6) + (36 x 6)} avaliações). O tempo médio dispendido para processamento destes modelos foi de aproximadamente 5 minutos.

Conforme observado acima, o melhor modelo para este estudo de caso foi ETP-U-TIT-TF-E1-Tempo. Pode-se observar que o uso da duração das anormalidades (legenda TEMPO) se sobressai à sua não utilização (SEM

TEMPO). Também se observa que o corpus relacionado à descrição detalhada das anormalidades não apresentou resultados expressivos como o título das anormalidades: o melhor modelo utilizando este corpus ficou na 52º posição em desempenho geral (dos 432 modelos avaliados).

A melhor configuração foi submetida à etapa de teste. As Figuras 49 e 50 mostram suas matrizes de confusão:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	250	7	0	41	0	0	3	1	0	1	0
	2	10	133	6	2	0	0	0	0	0	0	0
	3	8	10	131	10	2	0	0	1	0	0	0
	4	9	3	3	137	0	0	0	0	0	0	0
	5	9	0	3	0	29	0	2	0	0	0	0
	6	1	2	0	0	0	27	0	0	0	0	0
	7	14	0	0	1	0	0	13	0	0	0	0
	8	3	3	1	4	0	0	0	9	0	0	0
	9	6	0	0	1	0	0	0	0	9	0	0
	10	3	2	1	0	2	0	0	0	2	5	0
	11	0	1	0	0	0	0	0	0	0	0	1

Figura 49: Matriz de confusão (valores absolutos) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 50: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento.

Após esta primeira etapa de avaliação, observa-se desempenho razoável do modelo (considerando tanto micro F-score quando a média das classes). Porém, há de se destacar que o modelo confunde a classe 1, principalmente com as classes 4,

5, 7 e 9. Há de se destacar que a classe 6 possui bom desempenho, sendo este melhor que das classes 3 e 5, mais numerosas.

Observa-se também, pela matriz de confusão, que via de regra o modelo possui bom desempenho para as classes maiores (de 1 a 4) e para a classe 6, porém as demais classes não atingiram resultados expressivos. Coincidentemente são as classes de menor número de documentos na base de dados (conforme evidenciado na Tabela 9).

5.2.2. Estudo de caso 2: classificação hierárquica (4+7 classes)

Uma abordagem para melhorar o desempenho do modelo foi apontada no capítulo 4 e será avaliada neste estudo de caso: a classificação hierárquica. O objetivo é separar o problema da classificação em duas etapas, utilizando 2 SVCs diferentes, ao invés de classificar todos os documentos em uma etapa única. A primeira SVC se especializa em separar as classes maiores das menores. A segunda se especializa somente nas classes minoritárias. A proposta para a divisão das classes será a seguinte:

- Etapa 1: 1 x 2 x 3 x 4 x DEMAIS;
- Etapa 2: DEMAIS classes: 5 x 6 x 7 x 8 x 9 x 10 x 11.

Apesar do bom desempenho, a classe 6 foi deixada para a segunda etapa pois há a expectativa de que a classificação hierárquica auxilie no problema de balanceamento das classes menores. Como a classe 6 também possui baixo número de documentos, a mesma foi inserida na segunda etapa com intuito de aprimorar ainda mais seu resultado.

Importante destacar que, para a primeira etapa, as demais classes (que serão avaliadas pelo segundo classificador) serão consolidadas em uma só classe (aqui referida como “DEMAIS”), para que sejam separadas na primeira etapa.

.

Novamente serão utilizados os dados disponíveis com as seguintes configurações de dados de entrada:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):

    - Termos para título (E1, E2 e E3): 2303, 1320 e 708;
    - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;
- (2) - Duração: presente ou não na matriz final;
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada.

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 9 minutos.

Após efetuar uma primeira rodada, os gráficos dos resultados foram avaliados. Em seguida, fez-se o ajuste fino do parâmetro C dos SVCs. Finalmente os resultados obtidos no treinamento e validação foram ordenados nas seguintes tabelas, sendo que cada tabela se refere a um classificador (legenda ETP-1 = etapa 1):

Tabela 12: Resultados para a primeira etapa da classificação hierárquica.

Dado de entrada	C	Micro F-score
ETP-1-TIT-TFIDF-E1-SEM TEMPO	10	54,73%
ETP-1-TIT-TFIDF-E2-SEM TEMPO	10	54,71%
ETP-1-TIT-TFIDF-E2-TEMPO	10	54,60%
ETP-1-TIT-TFIDF-E1-TEMPO	10	54,60%
ETP-1-TIT-TFIDF-E2-TEMPO	5	54,57%

Ainda na etapa de treinamento, foram separadas somente as classes de 5 a 11 e efetuado o treinamento e validação do segundo SVC. As configurações dos dados de entrada utilizados foram os seguintes:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):
    - Termos para título (E1, E2 e E3): 2303, 1320 e 708;
    - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;
- (2) - Duração: presente ou não na matriz final;
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada.

Os gráficos foram gerados, o ajuste fino do parâmetro C foi efetuado e em seguida obteve-se os melhores resultados de treinamento e validação para a etapa 2 (legenda ETP-2 = etapa 2):

Tabela 13: Resultados para a segunda etapa da classificação hierárquica.

Dado de entrada	C	Micro F-score
ETP-2-TIT-TFIDF-E1-SEM TEMPO	10	54,05%
ETP-2-TIT-B-E1-SEM TEMPO	1	53,80%
ETP-2-TIT-TF-E2-SEM TEMPO	1	53,75%
ETP-2-TIT-B-E2-SEM TEMPO	1	53,75%
ETP-2-TIT-TF-E1-SEM TEMPO	1	53,75%

- Melhores SVCs de treinamento e validação para as etapas 1 e 2, com seus respectivos micro F-score, foram:
    - Etapa 1: Título/TF-IDF/E1/Sem tempo/C=10 – 54,73%;
    - Etapa 2: Título/TF-IDF/E1/Sem tempo/C=10– 54,05%.

Conforme esclarecido no capítulo 4, para cada etapa foram geradas duas matrizes de confusão (uma para cada etapa) e estas foram consolidadas em uma matriz final:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	247	7	1	40	3	0	3	0	0	2	0
	2	8	131	6	4	1	0	0	0	0	1	0
	3	4	10	135	11	1	0	0	1	0	0	0
	4	8	2	3	138	1	0	0	0	0	0	0
	5	8	0	3	1	28	0	2	0	0	1	0
	6	1	3	0	1	9	14	0	0	0	2	0
	7	14	1	0	1	0	0	11	0	0	1	0
	8	1	2	3	3	1	0	0	9	0	1	0
	9	3	0	1	0	3	0	0	0	9	0	0
	10	2	2	0	0	4	0	0	0	1	6	0
	11	0	0	1	0	0	0	0	0	0	1	0

Figura 51: Matriz de confusão (valores absolutos) para a classificação hierárquica.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	81,52%	2,31%	0,33%	13,20%	0,99%	0,00%	0,99%	0,00%	0,00%	0,66%	0,00%
	2	5,30%	86,75%	3,97%	2,65%	0,66%	0,00%	0,00%	0,00%	0,00%	0,66%	0,00%
	3	2,47%	6,17%	83,33%	6,79%	0,62%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,26%	1,32%	1,97%	90,79%	0,66%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	18,60%	0,00%	6,98%	2,33%	65,12%	0,00%	4,65%	0,00%	0,00%	2,33%	0,00%
	6	3,33%	10,00%	0,00%	3,33%	30,00%	46,67%	0,00%	0,00%	0,00%	6,67%	0,00%
	7	50,00%	3,57%	0,00%	3,57%	0,00%	0,00%	39,29%	0,00%	0,00%	3,57%	0,00%
	8	5,00%	10,00%	15,00%	15,00%	5,00%	0,00%	0,00%	45,00%	0,00%	5,00%	0,00%
	9	18,75%	0,00%	6,25%	0,00%	18,75%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	13,33%	13,33%	0,00%	0,00%	26,67%	0,00%	0,00%	0,00%	6,67%	40,00%	0,00%
	11	0,00%	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	0,00%
			Média das classes			63,47%	Micro-Fscore		40,56%			

Figura 52: Matriz de confusão (valores percentuais) para classificação hierárquica.

O retângulo azul ilustra as classes menores (5 a 11) que foram preditas como classes maiores (1 a 4) logo na primeira etapa. Com isso não há chance de o segundo classificador acertar estes documentos, pois eles já foram classificados e separados como pertencentes às classes maiores na primeira etapa. O retângulo verde evidencia as classes preditas erroneamente na primeira etapa como as demais classes (5 a 11), mas que pertenciam às classes maiores (1 a 4). Estes valores são enviados ao segundo classificador, também sem chance de acerto.

Observa-se que o revés de utilizar este método é que o conjunto das demais classes deve ser bem classificado na etapa 1, caso contrário o erro propaga-se para a etapa 2.

Analisando o desempenho geral via classificação hierárquica observa-se que a classe 1 ainda é bastante confundida pelo classificador. Para a classificação da segunda etapa do método hierárquico (classes de 5 a 11), as classes 5 e 9

tiveram um desempenho razoável, porém as classes 6 a 11 (com exceção da 9) tiveram desempenho baixo. Este fato não está totalmente atrelado à proporção das classes menores (quando analisadas de forma separada), pois há classes com poucos exemplos, como a classe 9, que possuem desempenho melhor que classes mais numerosas, como a classe 6. Porém, há uma tendência de que a quantidade de documentos esteja diretamente relacionada com seu desempenho de classificação. A Tabela 15 traz a frequência das classes menores bem como o percentual destas desconsiderando as quatro classes maiores. Ou seja, segregando as classes menores, ainda há desbalanceamento acentuado:

Tabela 14: Proporção das classes dentro do segundo classificador (não considerando as classes maiores, de 1 a 4).

Classe	Freq.	
5	432	30,42%
6	266	18,73%
7	209	14,72%
8	187	13,17%
9	172	12,11%
10	142	10,00%
11	12	0,85%

De qualquer forma, se faz desejável o aumento no número de documentos das classes menores, de modo a avaliar se há aumento no desempenho do classificador como um todo.

Para ter uma ideia do desempenho final da classificação hierárquica, na Figura 53 compara-se a matriz de confusão de teste com valores percentuais pelo estudo de caso base (Figura 54):

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	81,52%	2,31%	0,33%	13,20%	0,99%	0,00%	0,99%	0,00%	0,00%	0,66%	0,00%
	2	5,30%	86,75%	3,97%	2,65%	0,66%	0,00%	0,00%	0,00%	0,00%	0,66%	0,00%
	3	2,47%	6,17%	83,33%	6,79%	0,62%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,26%	1,32%	1,97%	90,79%	0,66%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	18,60%	0,00%	6,98%	2,33%	65,12%	0,00%	4,65%	0,00%	0,00%	2,33%	0,00%
	6	3,33%	10,00%	0,00%	3,33%	30,00%	46,67%	0,00%	0,00%	0,00%	6,67%	0,00%
	7	50,00%	3,57%	0,00%	3,57%	0,00%	0,00%	39,29%	0,00%	0,00%	3,57%	0,00%
	8	5,00%	10,00%	15,00%	15,00%	5,00%	0,00%	0,00%	45,00%	0,00%	5,00%	0,00%
	9	18,75%	0,00%	6,25%	0,00%	18,75%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	13,33%	13,33%	0,00%	0,00%	26,67%	0,00%	0,00%	0,00%	6,67%	40,00%	0,00%
	11	0,00%	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	0,00%
			Média das classes			63,47%	Micro-Fscore		40,56%			

Figura 53: Matriz de confusão (valores percentuais) para a classificação hierárquica.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 54: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento, caso base.

Com exceção das classes 3, 4 e 10, o classificador hierárquico possui pior desempenho que o caso base. Pretende-se mais adiante analisar novamente o método hierárquico, inclusive efetuando-o em conjunto com o balanceamento das classes menores.

5.2.3. Estudo de caso 3: redução da classe 1

Conforme já apontado ao final do estudo de caso 1, observa-se um desbalanceamento acentuado da base de dados. Para mitigar este efeito foram

propostas duas abordagens: o método de classificação hierárquico, apresentado na seção anterior e o balanceamento da base de dados. Nesta seção de estudos de caso, será abordado o balanceamento da base de dados, no intuito de reduzir a numerosa classe 1.

No capítulo 4 foram apresentados métodos de balanceamento para redução de classes numerosas. A Tabela 16 mostra a distribuição das quatro maiores classes, bem como seu percentual em relação a toda base de dados:

Tabela 15: Divisão percentual da base de dados disponível com foco nas classes majoritárias.

Classe	Frequência	
1	3077	33,44%
2	1607	17,46%
3	1576	17,13%
4	1522	16,54%

A ideia é reduzir a classe 1 até um tamanho próximo às classes 2, 3 e 4. Para esta redução foram utilizados os seguintes métodos de balanceamento:

- ENN: *Edited Nearest Neighbors*;
- NCL: *Neighborhood Cleaning Rule*;
- *Tomek Links*;
- *Undersampling*: redução aleatória da classe.

Foram avaliadas as seguintes combinações de configurações de dados de entrada:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):
    - Termos para título (E1, E2 e E3): 2303, 1320 e 708;
    - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;
- (2) - Duração: presente ou não na matriz final;
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada;
    - SVC utilizado: etapa única (11 classes).

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 6 minutos.

Este processo ocorreu de forma semelhante aos estudos de caso anteriores: para cada método de redução da classe 1 foram efetuadas duas etapas de busca do melhor parâmetro de penalidade (C), onde a primeira consistia em intervalos maiores para o parâmetro. Em seguida, fez-se um ajuste fino e obteve-se para cada método o melhor valor do parâmetro C. A Tabela 17 resume os melhores resultados da etapa preliminar de treinamento e validação para cada método de balanceamento:

Tabela 16: Resumo dos resultados de treinamento e validação para redução da classe 1 (SVC de etapa única: 11 classes).

Método	Dado de entrada	C	Micro F-score
ENN	ETP-2-TIT-TFIDF-E1-SEM TEMPO	10	56,19%
NCL	ETP-2-TIT-B-E1-SEM TEMPO	1	56,25%
Tomek	ETP-2-TIT-TF-E2-SEM TEMPO	1	54,75%
Unders.	ETP-2-TIT-B-E2-SEM TEMPO	1	46,18%

Conforme apontado na tabela, o NCL obteve o melhor desempenho (ligeiramente maior que do ENN). Vale destacar que foram removidos 1319 documentos da classe 1, reduzindo de 3077 para 1758 documentos dentro da classe 1. Seguem as matrizes de confusão obtidas para a etapa de teste:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	237	10	10	43	0	0	3	0	0	0	0
	2	8	131	10	2	0	0	0	0	0	0	0
	3	3	11	137	8	1	0	0	1	1	0	0
	4	3	3	7	139	0	0	0	0	0	0	0
	5	9	0	4	0	28	0	2	0	0	0	0
	6	1	4	0	0	0	25	0	0	0	0	0
	7	12	1	2	1	0	0	12	0	0	0	0
	8	7	3	4	3	0	0	0	3	0	0	0
	9	5	1	0	1	0	0	0	0	9	0	0
	10	1	2	7	0	2	0	0	0	3	0	0
	11	0	0	2	0	0	0	0	0	0	0	0

Figura 55: Matriz de confusão (valores absolutos) para a classificação em etapa única com redução da classe 1 por NCL.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	78,22%	3,30%	3,30%	14,19%	0,00%	0,00%	0,99%	0,00%	0,00%	0,00%	0,00%
	2	5,30%	86,75%	6,62%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	1,85%	6,79%	84,57%	4,94%	0,62%	0,00%	0,00%	0,62%	0,62%	0,00%	0,00%
	4	1,97%	1,97%	4,61%	91,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	9,30%	0,00%	65,12%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	13,33%	0,00%	0,00%	0,00%	83,33%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	42,86%	3,57%	7,14%	3,57%	0,00%	0,00%	42,86%	0,00%	0,00%	0,00%	0,00%
	8	35,00%	15,00%	20,00%	15,00%	0,00%	0,00%	0,00%	15,00%	0,00%	0,00%	0,00%
	9	31,25%	6,25%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	6,67%	13,33%	46,67%	0,00%	13,33%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%
	11	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
			Média das classes			60,35%	Micro-Fscore		39,47%			

Figura 56: Matriz de confusão (valores percentuais) para a classificação em etapa única com redução da classe 1 por NCL.

A Figura 57 mostra o melhor resultado obtido até então para comparação:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 57: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento.

Observa-se que a redução da classe 1, mesmo que pelo método de melhor desempenho, acaba por prejudicar o classificador. Como o método tende a remover exemplos próximos às outras classes (*borderline*), é provável que tenham sido removidos alguns vetores de suporte que residem nestas fronteiras, prejudicando o SVC.

5.2.4.
Estudo de caso 4: redução da classe 1 em conjunto com a classificação hierárquica

Para este estudo de caso será efetuada a classificação hierárquica, porém com prévia redução da classe 1. Como a classificação hierárquica demanda 2 SVCs (um para classes maiores e outro para menores), só será treinado novamente o SVC para classes maiores, devido à redução da classe 1. Como não há alteração nas classes menores, será utilizado o mesmo SVC do estudo de caso feito para classificação hierárquica. O SVC já obtido para a segunda etapa possui a seguinte configuração:

- Corpus: título das anormalidades;
- Peso: TF-IDF;
- Esparcidade: E1;
- Sem utilizar a duração da anormalidade como dado de entrada;
- Valor do parâmetro de penalidade do SVC: 10.

Para a redução da classe majoritária associada ao treinamento e à validação do SVC da primeira etapa serão avaliados todos os métodos apontados para este tipo de balanceamento. Segue as combinações de dados de entrada avaliados:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):
    - Termos para título (E1, E2 e E3): 2303, 1320 e 708;
    - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;
- (2) - Duração: presente ou não na matriz final;
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada.

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 11 minutos.

Após avaliar todos as opções e se efetuar o ajuste fino para cada método de balanceamento via redução da classe 1, segue uma tabela resumo dos melhores resultados por método:

Tabela 17: Resumo dos resultados de treinamento e validação para redução da classe 1 (SVC para as 4 classes maiores, por meio de classificação hierárquica).

Método	Dado de entrada	C	Micro F-score
ENN	ETP-2-TIT-TFIDF-E1-SEM TEMPO	0,5	57,08%
NCL	ETP-2-TIT-B-E1-SEM TEMPO	0,5	57,34%
Tomek	ETP-2-TIT-TF-E2-SEM TEMPO	10	55,82%
Unders.	ETP-2-TIT-B-E2-SEM TEMPO	10	47,54%

Como pode ser observado, mais uma vez o método NCL se sobressaiu aos demais (houve 1309 remoções de documentos da classe 1). Seguem as matrizes de confusão do teste final:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	185	6	1	43	44	4	19	0	1	0	0
	2	3	126	6	2	8	0	5	0	1	0	0
	3	0	9	129	8	12	1	3	0	0	0	0
	4	1	3	5	139	4	0	0	0	0	0	0
	5	7	0	3	0	31	0	2	0	0	0	0
	6	0	3	0	0	2	25	0	0	0	0	0
	7	2	1	0	1	5	1	18	0	0	0	0
	8	0	1	1	3	11	0	1	3	0	0	0
	9	1	0	0	1	2	0	3	0	9	0	0
	10	1	1	0	0	10	0	0	0	2	1	0
	11	0	0	0	0	2	0	0	0	0	0	0

Figura 58: Matriz de confusão (valores absolutos) para a classificação em duas etapas (hierárquico) com redução da classe 1 por NCL.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	61,06%	1,98%	0,33%	14,19%	14,52%	1,32%	6,27%	0,00%	0,33%	0,00%	0,00%
	2	1,99%	83,44%	3,97%	1,32%	5,30%	0,00%	3,31%	0,00%	0,66%	0,00%	0,00%
	3	0,00%	5,56%	79,63%	4,94%	7,41%	0,62%	1,85%	0,00%	0,00%	0,00%	0,00%
	4	0,66%	1,97%	3,29%	91,45%	2,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	16,28%	0,00%	6,98%	0,00%	72,09%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	10,00%	0,00%	0,00%	6,67%	83,33%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	7,14%	3,57%	0,00%	3,57%	17,86%	3,57%	64,29%	0,00%	0,00%	0,00%	0,00%
	8	0,00%	5,00%	5,00%	15,00%	55,00%	0,00%	5,00%	15,00%	0,00%	0,00%	0,00%
	9	6,25%	0,00%	0,00%	6,25%	12,50%	0,00%	18,75%	0,00%	56,25%	0,00%	0,00%
	10	6,67%	6,67%	0,00%	0,00%	66,67%	0,00%	0,00%	0,00%	13,33%	6,67%	0,00%
	11	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
			Média das classes			61,32%	Micro-Fscore		32,11%			

Figura 59: Matriz de confusão (valores percentuais) para a classificação em duas etapas (hierárquico) com redução da classe 1 por NCL.

Pode-se comparar o desempenho deste modelo com o modelo onde também foi reduzida a classe 1 por NCL, porém com classificação em etapa única:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	78,22%	3,30%	3,30%	14,19%	0,00%	0,00%	0,99%	0,00%	0,00%	0,00%	0,00%
	2	5,30%	86,75%	6,62%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	1,85%	6,79%	84,57%	4,94%	0,62%	0,00%	0,00%	0,62%	0,62%	0,00%	0,00%
	4	1,97%	1,97%	4,61%	91,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	9,30%	0,00%	65,12%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	13,33%	0,00%	0,00%	0,00%	83,33%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	42,86%	3,57%	7,14%	3,57%	0,00%	0,00%	42,86%	0,00%	0,00%	0,00%	0,00%
	8	35,00%	15,00%	20,00%	15,00%	0,00%	0,00%	0,00%	15,00%	0,00%	0,00%	0,00%
	9	31,25%	6,25%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	6,67%	13,33%	46,67%	0,00%	13,33%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%
	11	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
			Média das classes			60,35%	Micro-Fscore		39,47%			

Figura 60: Matriz de confusão (valores percentuais) para a classificação em etapa única com redução da classe 1 por NCL.

Em termos da métrica micro F-score a classificação hierárquica foi inferior à classificação direta, para o mesmo método de redução da classe 1 (NCL), porém o acerto médio das classes no hierárquico foi ligeiramente superior na classificação hierárquica.

Finalmente, observa-se que o modelo do caso base ainda é superior a todos os modelos propostos nos demais estudos de caso:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 61: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento, caso base.

5.2.5.
Estudo de caso 5: balanceamento das classes 6 a 11

As classes menores possuem poucos documentos quando comparadas às classes maiores. A classificação hierárquica proposta prejudicou o desempenho do classificador para estas classes de menor número (com algumas exceções), conforme observado nos resultados apresentados. Neste estudo de caso será avaliado o aumento destas classes menores, somente.

Para o aumento das classes minoritárias foram utilizados os seguintes métodos:

- SMOTE;
- *Oversampling*: aumento aleatório das classes.

Todas as classes foram equiparadas à classe 5, que possui 432 documentos. Portanto o método foi aplicado a cada classe em separado até se obter o tamanho desejado.

Novamente, foram empregadas as seguintes combinações de dados de entrada:

- (2) - Corpus: título e descrição da anormalidade (ocorrência)
- (3) - Peso: TF, TF-IDF e binário
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3)
    - Termos para título (E1, E2 e E3): 2303, 1320 e 708
    - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746
- (2) - Duração: presente ou não na matriz final.
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada;
    - SVC utilizado: etapa única (11 classes).

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 30 minutos.

Este processo ocorreu de forma semelhante aos estudos de caso anteriores: para cada método de aumento das classes menores foram efetuadas duas etapas de busca do melhor parâmetro de penalidade (C), onde a primeira consistia em intervalos maiores para o parâmetro. Em seguida, fez-se um ajuste fino e obteve-se para cada método o melhor valor do parâmetro C. A Tabela 19 resume os

melhores resultados da fase de treinamento e validação para cada método de balanceamento:

Tabela 18: Resumo dos resultados de treinamento e validação para aumento das classes de 6 a 11 (SVC de etapa única: 11 classes).

Método	Dado de entrada	C	Micro F-score
Oversamp.	ETP-U-TIT-TF-E1-TEMPO	0,5	54,78%
SMOTE	ETP-U-TIT-TFIDF-E2-TEMPO	1	58,04%

O método SMOTE foi selecionado para etapa de teste, resultando nas matrizes de confusão:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	253	4	0	40	1	0	3	1	0	1	0
	2	15	125	8	2	0	0	0	1	0	0	0
	3	11	8	132	8	2	0	0	1	0	0	0
	4	9	3	3	137	0	0	0	0	0	0	0
	5	9	0	2	0	30	0	2	0	0	0	0
	6	1	2	0	0	0	27	0	0	0	0	0
	7	14	1	0	1	0	0	12	0	0	0	0
	8	4	1	1	3	0	0	0	11	0	0	0
	9	5	0	0	1	0	0	0	0	10	0	0
	10	2	2	1	0	2	0	0	1	2	5	0
	11	1	0	0	0	0	0	0	0	0	0	1

Figura 62: Matriz de confusão (valores absolutos) para a classificação em etapa única com aumento das classes de 6 a 11 por SMOTE.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	83,50%	1,32%	0,00%	13,20%	0,33%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	9,93%	82,78%	5,30%	1,32%	0,00%	0,00%	0,00%	0,66%	0,00%	0,00%	0,00%
	3	6,79%	4,94%	81,48%	4,94%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	4,65%	0,00%	69,77%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	3,57%	0,00%	3,57%	0,00%	0,00%	42,86%	0,00%	0,00%	0,00%	0,00%
	8	20,00%	5,00%	5,00%	15,00%	0,00%	0,00%	0,00%	55,00%	0,00%	0,00%	0,00%
	9	31,25%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	62,50%	0,00%	0,00%
	10	13,33%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	6,67%	13,33%	33,33%	0,00%
	11	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			69,14%	Micro-Fscore		43,01%			

Figura 63: Matriz de confusão (valores percentuais) para a classificação em etapa única com aumento das classes de 6 a 11 por SMOTE.

Comparando com o melhor modelo até então obtido:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 64: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento, caso base.

Observa-se que o modelo do caso base tem desempenho muito próximo a este último modelo, mas ainda é superior na métrica micro F-score. Porém, para a métrica de média de acerto entre classes, o balanceamento das classes menores efetuado traz um resultado ligeiramente superior.

5.2.6. Estudo de caso 6: aumento das classes 6 a 11 em conjunto com a classificação hierárquica

Neste estudo de caso será avaliado somente o impacto do aumento das classes minoritárias sobre a classificação hierárquica (sem efetuar a redução da classe majoritária).

A classificação hierárquica possui dois SVCs que necessitam ser treinados novamente, pois este balanceamento interfere tanto no primeiro SVC quanto no segundo.

Para o aumento da classe majoritária serão avaliados os dois métodos apontados para este tipo de balanceamento. Seguem as combinações de dados de entrada avaliados:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):

- Termos para título (E1, E2 e E3): 2303, 1320 e 708;
- Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;

- (2) - Duração: presente ou não na matriz final;
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada.

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 35 minutos.

Após efetuar o ajuste fino para cada método de balanceamento, segue tabela resumo com os melhores modelos obtidos na etapa de treinamento e validação:

Tabela 19: Resumo dos resultados de treinamento e validação para aumento das classes de 6 a 11, para cada etapa de classificação hierárquica.

Método	Dado de entrada	C	Micro F-score
Oversamp.	ETP-1-TIT-TF-E2-TEMPO	0,5	55,15%
SMOTE	ETP-1-TIT-TF-E1-TEMPO	0,5	55,83%
Oversamp.	ETP-2-TIT-TFIDF-E3-TEMPO	40	61,09%
SMOTE	ETP-2-TIT-TFIDF-E1-TEMPO	30	61,51%

Também neste cenário o método SMOTE se sobressaiu ao *Oversampling*. Para estas configurações das etapas 1 e 2 foi efetuado o teste final:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	251	6	0	41	0	1	2	1	0	1	0
	2	17	124	6	2	0	0	0	1	0	1	0
	3	12	9	128	8	1	1	1	1	1	0	0
	4	11	3	4	134	0	0	0	0	0	0	0
	5	9	0	3	0	24	0	2	2	3	0	0
	6	1	3	0	0	0	24	1	0	0	1	0
	7	14	1	0	1	0	0	11	1	0	0	0
	8	10	1	1	3	0	0	1	3	0	0	1
	9	4	0	0	1	0	0	1	0	9	0	1
	10	7	0	0	0	2	0	0	1	1	4	0
	11	1	0	0	0	0	0	0	0	0	1	0

Figura 65: Matriz de confusão (valores absolutos) para a classificação em duas etapas (hierárquico) com aumento das classes de 6 a 11 por SMOTE.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,84%	1,98%	0,00%	13,53%	0,00%	0,33%	0,66%	0,33%	0,00%	0,33%	0,00%
	2	11,26%	82,12%	3,97%	1,32%	0,00%	0,00%	0,00%	0,66%	0,00%	0,66%	0,00%
	3	7,41%	5,56%	79,01%	4,94%	0,62%	0,62%	0,62%	0,62%	0,62%	0,00%	0,00%
	4	7,24%	1,97%	2,63%	88,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	55,81%	0,00%	4,65%	4,65%	6,98%	0,00%	0,00%
	6	3,33%	10,00%	0,00%	0,00%	0,00%	80,00%	3,33%	0,00%	0,00%	3,33%	0,00%
	7	50,00%	3,57%	0,00%	3,57%	0,00%	0,00%	39,29%	3,57%	0,00%	0,00%	0,00%
	8	50,00%	5,00%	5,00%	15,00%	0,00%	0,00%	5,00%	15,00%	0,00%	0,00%	5,00%
	9	25,00%	0,00%	0,00%	6,25%	0,00%	0,00%	6,25%	0,00%	56,25%	0,00%	6,25%
	10	46,67%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,67%	6,67%	26,67%	0,00%
	11	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	0,00%
			Média das classes			60,51%	Micro-Fscore		38,14%			

Figura 66: Matriz de confusão (valores percentuais) para a classificação em duas etapas (hierárquico) com aumento das classes de 6 a 11 por SMOTE.

Para efeitos de comparação, segue a matriz de confusão onde também foram aumentadas as classes 6 a 11 por SMOTE, porém com classificação em etapa única:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	83,50%	1,32%	0,00%	13,20%	0,33%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	9,93%	82,78%	5,30%	1,32%	0,00%	0,00%	0,00%	0,66%	0,00%	0,00%	0,00%
	3	6,79%	4,94%	81,48%	4,94%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	4,65%	0,00%	69,77%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	3,57%	0,00%	3,57%	0,00%	0,00%	42,86%	0,00%	0,00%	0,00%	0,00%
	8	20,00%	5,00%	5,00%	15,00%	0,00%	0,00%	0,00%	55,00%	0,00%	0,00%	0,00%
	9	31,25%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	62,50%	0,00%	0,00%
	10	13,33%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	6,67%	13,33%	33,33%	0,00%
	11	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			69,14%	Micro-Fscore		43,01%			

Figura 67: Matriz de confusão (valores percentuais) para a classificação em etapa única com aumento das classes de 6 a 11 por SMOTE.

O balanceamento das classes menores para a classificação hierárquica não superou o mesmo balanceamento em etapa única.

Finalmente, observa-se novamente que o modelo do caso base ainda é superior a todos os modelos propostos nos demais estudos de caso na métrica micro F-score:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 68: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento, caso base.

5.2.7.
Estudo de caso 7: abordagens simultâneas

Nesta seção serão implementadas simultaneamente as abordagens de balanceamento da base de dados (redução da classe 1 e aumento das classes de 6 a 11). Isto será feito tanto para classificadores em etapa única (11 classes) quanto para classificação hierárquica (etapa dupla).

Para o balanceamento, serão utilizados os métodos que tiveram melhor desempenho até então:

- Redução classe 1: NCL;
- Aumento classes de 6 a 11: SMOTE;

Em ambos os casos de classificação, os seguintes dados de entrada foram avaliados:

- (2) - Corpus: título e descrição da anormalidade (ocorrência);
- (3) - Peso: TF, TF-IDF e binário;
- (3) - Fator redução esparcidade: 1,0 (E1), 0,9998(E2) e 0,995(E3):
  - Termos para título (E1, E2 e E3): 2303, 1320 e 708;
  - Termos para ocorrência (E1, E2 e E3): 5040, 2918, 1746;
- (2) - Duração: presente ou não na matriz final;
- Variação da constante de custo conforme os pontos dos gráficos de obtenção do melhor valor de C para cada SVC treinada.

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 32 minutos.

Para a etapa de treinamento e validação para classificação em etapa única (1 SVC para 11 classes), os seguintes resultados foram encontrados:

Tabela 20: Melhores resultados de treinamento e validação para balanceamento simultâneo, utilizando SVC em etapa única.

Dado de entrada	C	Micro F-score
ETP-U-TIT-TFIDF-E2-TEMPO	0,5	58,04%
ETP-U-TIT-TFIDF-E1-TEMPO	1	57,81%
ETP-U-TIT-TF-E1-TEMPO	1	57,81%
ETP-U-TIT-TFIDF-E2-TEMPO	1	57,74%
ETP-U-TIT-TF-E2-TEMPO	0,5	57,73%

Utilizando o melhor modelo desta etapa, submeteu-se o mesmo ao teste final, que resultou nas seguintes matrizes de confusão:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	256	4	0	41	1	0	0	1	0	0	0
	2	18	124	7	2	0	0	0	0	0	0	0
	3	12	9	129	8	1	0	0	2	1	0	0
	4	11	3	4	134	0	0	0	0	0	0	0
	5	11	0	3	0	28	0	1	0	0	0	0
	6	1	2	0	0	0	27	0	0	0	0	0
	7	17	1	0	1	0	0	9	0	0	0	0
	8	7	0	1	3	0	0	0	9	0	0	0
	9	5	1	0	1	0	0	0	0	9	0	0
	10	7	0	0	0	2	0	0	1	3	2	0
	11	1	0	0	0	0	0	0	0	0	0	1

Figura 69: Matriz de confusão (valores absolutos) para a classificação em etapa única com redução da classe 1 por NCL e aumento das classes de 6 a 11 por SMOTE.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	84,49%	1,32%	0,00%	13,53%	0,33%	0,00%	0,00%	0,33%	0,00%	0,00%	0,00%
	2	11,92%	82,12%	4,64%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	7,41%	5,56%	79,63%	4,94%	0,62%	0,00%	0,00%	1,23%	0,62%	0,00%	0,00%
	4	7,24%	1,97%	2,63%	88,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	25,58%	0,00%	6,98%	0,00%	65,12%	0,00%	2,33%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	60,71%	3,57%	0,00%	3,57%	0,00%	0,00%	32,14%	0,00%	0,00%	0,00%	0,00%
	8	35,00%	0,00%	5,00%	15,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	31,25%	6,25%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	46,67%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,67%	20,00%	13,33%	0,00%
	11	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			63,62%	Micro-Fscore		40,56%			

Figura 70: Matriz de confusão (valores percentuais) para a classificação em etapa única com redução da classe 1 por NCL e aumento das classes de 6 a 11 por SMOTE.

Já para a classificação hierárquica, foram obtidos na etapa de treinamento e de validação os modelos de melhor desempenho considerando também o balanceamento das classes (redução da classe 1 e aumento das minoritárias). Vale destacar que, como são dois SVCs a ser treinados, a base de dados a qual o primeiro SVC foi submetido teve a classe 1 reduzida bem como as minoritárias aumentadas. Em seguida, o segundo SVC (que necessita ser treinado somente com base nas classes menores) utilizou-se da base de dados balanceada, porém retirando as classes de 1 a 4.

O tempo aproximado para processamento de todas as configurações para este estudo de caso foi de 35 minutos.

Os melhores resultados obtidos para cada etapa da classificação hierárquica foram:

Tabela 21: Melhores modelos obtidos para balanceamento das classes para classificação hierárquica.

Método	Dado de entrada	C	Micro F-score
NCL+SMOTE	ETP-1-TIT-TF-E1-TEMPO	0,5	57,34%
SMOTE	ETP-2-TIT-TF-E1-TEMPO	20	61,19%

Este modelo foi testado e os resultados podem ser observados nas matrizes de confusão abaixo:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	184	3	0	24	68	2	19	1	2	0	0
	2	3	129	6	1	10	0	1	1	0	0	0
	3	0	10	131	8	12	0	1	0	0	0	0
	4	9	3	5	102	33	0	0	0	0	0	0
	5	7	0	0	0	34	0	2	0	0	0	0
	6	0	3	0	0	6	21	0	0	0	0	0
	7	2	0	0	0	8	0	18	0	0	0	0
	8	0	2	1	1	11	0	1	4	0	0	0
	9	1	0	0	0	7	0	1	0	7	0	0
	10	0	1	0	0	11	0	0	0	3	0	0
	11	0	0	0	0	2	0	0	0	0	0	1

Figura 71: Matriz de confusão (valores absolutos) para a classificação em duas etapas (hierárquico) com redução da classe 1 por NCL e aumento das classes de 6 a 11 por SMOTE.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	60,73%	0,99%	0,00%	7,92%	22,44%	0,66%	6,27%	0,33%	0,66%	0,00%	0,00%
	2	1,99%	85,43%	3,97%	0,66%	6,62%	0,00%	0,66%	0,66%	0,00%	0,00%	0,00%
	3	0,00%	6,17%	80,86%	4,94%	7,41%	0,00%	0,62%	0,00%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	3,29%	67,11%	21,71%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	16,28%	0,00%	0,00%	0,00%	79,07%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	10,00%	0,00%	0,00%	20,00%	70,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	7,14%	0,00%	0,00%	0,00%	28,57%	0,00%	64,29%	0,00%	0,00%	0,00%	0,00%
	8	0,00%	10,00%	5,00%	5,00%	55,00%	0,00%	5,00%	20,00%	0,00%	0,00%	0,00%
	9	6,25%	0,00%	0,00%	0,00%	43,75%	0,00%	6,25%	0,00%	43,75%	0,00%	0,00%
	10	0,00%	6,67%	0,00%	0,00%	73,33%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%
	11	0,00%	0,00%	0,00%	0,00%	66,67%	0,00%	0,00%	0,00%	0,00%	0,00%	33,33%
			Média das classes			57,12%	Micro-Fscore		28,21%			

Figura 72: Matriz de confusão (valores percentuais) para a classificação em duas etapas (hierárquico) com redução da classe 1 por NCL e aumento das classes de 6 a 11 por SMOTE.

Finalmente, observa-se mais uma vez que o modelo do caso base ainda é superior a todos os modelos propostos nos demais estudos de caso avaliados pela métricas micro F-score.

Tendo em vista todos os estudos de caso avaliados até então, foi elaborada uma tabela resumo, consolidando os melhores modelos obtidos na fase de teste. A nomenclatura “11C” se refere à classificação em etapa única (1 SVC) e a nomenclatura “4+7C” se refere à classificação hierárquica (2 SVCs):

Tabela 22: Consolidação dos resultados obtidos até esta etapa.

		Balanceamento			Sem balanceamento
		Redução classe 1	Aumento das classes menores	Ambos balanc.	
		NCL	SMOTE	NCL+SMOTE	
Micro F-score	SVC 11C	39,47%	**43,01%**	**40,56%**	**43,18%**
	SVC 4+7C	32,11%	38,14%	28,21%	40,56%
Média das classes	SVC 11C	60,35%	**69,14%**	**63,62%**	**68,00%**
	SVC 4+7C	61,32%	60,51%	57,12%	63,47%

Como pode se observar, o caso base apresentou o maior desempenho na métrica micro F-score. Por outro lado, com resultado muito próximo, o aumento das classes minoritárias para o SVC de classificação única obteve desempenho ligeiramente superior na métrica de média das classes. Em terceiro lugar, o balanceamento, tanto com redução da classe 1 e aumento das minoritárias.

5.3. Estudos de caso após primeira correção da base de dados

Foi escolhido o melhor modelo obtido até então para efetuar a primeira correção da base de dados. O melhor classificador obtido foi o SVC de classificação das onze classes diretamente (etapa única), valor da constante C=3, sem balanceamento. A matriz de dados utilizada foi:

- Corpus: Título;
- Peso: TF;
- Redutor de esparcidade: E1;
- Duração das anormalidades contabilizada;

Como o processo de correção das classes consome muito tempo dos operadores (recurso crítico), escolheram-se classes as quais o modelo se confundia mais, após análise da matriz de confusão. Como a base de treinamento é maior, a correção foi de menor monta. Já a base para a base de teste, de tamanho menor, todos os erros do SVC foram avaliados. O intuito desta correção foi de avaliar se

o classificador realmente falhou em classificar os documentos ou se as classes foram apontadas erroneamente na base de dados.

Resumo das correções:

- Base de dados de treino: 321 (3,8% de 8290 documentos).
- Base de dados de teste: 106 (11,5% de 922 documentos)

A Tabela 23 traz uma síntese das correções, mostrando quando houve falha do classificador, do operador, ou de ambos:

Tabela 23: Comparação dos resultados da correção.

	Anormalidades	% do total corrigido
Acerto do classificador e erro do operador	341	79,86%
Erro do classificador e acerto do operador	63	14,75%
Ambos erraram	23	5,39%
Total	**427**	

Em torno de 21% das classes corrigidas da base de treino eram classes apontadas como 4, 5 ou 8 e que, na verdade, pertenciam a classe 1. Vale ressaltar que no primeiro teste efetuado com o operador (disposto no capítulo 4) estas classes também tiveram maior taxa de erro.

Após esta correção, os modelos foram avaliados novamente (já considerando a base de dados corrigida) e comparou-se o melhor modelo antes e após a correção efetuada:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 73: Matriz de confusão antes da primeira correção (melhor modelo: SVC de etapa única, sem balanceamento).

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	92,47%	1,71%	0,34%	5,14%	0,34%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	2	5,81%	90,32%	1,29%	2,58%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	1,31%	3,92%	91,50%	1,31%	1,31%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%
	4	1,12%	0,00%	0,00%	97,75%	0,00%	0,56%	0,00%	0,56%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	9,09%	3,03%	87,88%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	3,57%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	23,08%	3,85%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	16,00%	8,00%	4,00%	24,00%	0,00%	0,00%	0,00%	48,00%	0,00%	0,00%	0,00%
	9	15,38%	7,69%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	5,88%	0,00%	0,00%	0,00%	5,88%	52,94%	0,00%
	11	0,00%	50,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
			Média das classes			80,34%	Micro-Fscore		61,84%			

Figura 74: Matriz de confusão após a primeira correção (melhor modelo: SVC de etapa única, sem balanceamento).

Como se pode observar, o desempenho do modelo aumentou 2,67 pontos percentuais na métrica micro F-score. Notória melhoria no acerto de várias classes. Também se reduziu substancialmente os erros do modelo entre classes, especialmente entre as classes 4 e 5 com relação à classe 1. A classe 4 possui a melhor taxa de acerto, bem próxima à classe 6, que já possuía alto desempenho.

Como este melhor modelo foi selecionado com base em um banco de dados enviesado, efetuaram-se novamente as rodadas para todos os modelos avaliados até então.

Um ponto importante para essa nova etapa de estudos de caso reside na alteração do classificador hierárquico, pois houve sensível melhora no desempenho das classes 5 e 6. Com isso se propôs a seguinte divisão para nova classificação hierárquica:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	92,47%	1,71%	0,34%	5,14%	0,34%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	2	5,81%	90,32%	1,29%	2,58%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	1,31%	3,92%	91,50%	1,31%	1,31%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%
	4	1,12%	0,00%	0,00%	97,75%	0,00%	0,56%	0,00%	0,56%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	9,09%	3,03%	87,88%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	3,57%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	23,08%	3,85%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	16,00%	8,00%	4,00%	24,00%	0,00%	0,00%	0,00%	48,00%	0,00%	0,00%	0,00%
	9	15,38%	7,69%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	5,88%	0,00%	0,00%	0,00%	5,88%	52,94%	0,00%
	11	0,00%	50,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
			Média das classes			80,34%	Micro-Fscore		61,84%			

Figura 75: Sugestão de separação de classes para a classificação hierárquica.

Ou seja, na primeira etapa da classificação hierárquica foram classificadas as classes de 1 a 6, e as demais foram conjugadas em outra classe, para avaliação posterior por outro SVC específico para as classes menores.

O tempo aproximado para processamento de todas as configurações para este estudo de caso (nova configuração de classificação hierárquica) foi de 9 minutos.

Após se avaliar a eficácia no balanceamento da base de dados, se propôs o aumento somente para as classes 10 e 11, que tiveram baixo desempenho do classificador bem como possuem poucas observações na fase de treinamento e validação (1,54% da base para a classe 10 e 0,13% dos elementos da base para a classe 11). A redução da classe 1 foi suprimida nesta nova etapa, pois prejudicou o desempenho em todos os casos. O aumento das classes 6, 7, 8 e 9 também foi suprimido, pois se espera obter melhor resultado com a classificação hierárquica, somente. Além disso, o aumento dessas classes não trouxe ganho relevante.

O tempo aproximado para processamento de todas as configurações para este estudo de caso (balanceamento das classes 10 e 11) foi de 8 minutos.

Finalmente, foram testados em conjunto a nova configuração hierárquica proposta e o balanceamento das classes 10 e 11. O tempo aproximado de processamento para este caso foi de 15 minutos.

5.3.1. Estudo de caso 8: classificação hierárquica (6+5 classes) e Balanceamento das classes 10 e 11

Sob estas condições, foram efetuados novamente o treinamento/validação bem como o teste de todos os classificadores, com base nesta correção efetuada na base de dados. A Tabela 25 apresenta um resumo dos resultados consolidados obtidos após efetuar a correção da base de dados. Na primeira coluna pode-se observar qual o corpus avaliado. Na segunda coluna nota-se qual abordagem de classificação foi utilizada: etapa única ou em duas etapas, podendo ser a classificação hierárquica de 4 classes na primeira etapa, seguido das demais 7 classes (4-7) ou ainda 6 classes atribuídas na primeira etapa pelo classificador, seguido do apontamento das demais 5 classes (6-5). Segue resumo final dos resultados:

Tabela 24: Resumo dos testes finais após primeira correção. Obs.: “4-7” refere-se ao hierárquico com classificação de 4 classes inicialmente, seguido das demais 7 classes. A mesma lógica se aplica ao termo “6-5”.

Melhores resultados ***(μ F-Score)***		Balanceamento		Sem balanceamento
		Aumento 10 E 11	Aumento 6 a 11	
Corpus	Modelo SVM	Smote	Smote	
Título	Etapa única	**64,16%**	N/A	61,84%
	2 etapas (4-7)	**62,70%**	60,18%	61,28%
	2 etapas (6-5)	53,00%	N/A	**65,07%**
Descrição	Etapa única	N/A	N/A	48,74%
	2 etapas (4-7)	N/A	N/A	48,94%

Melhores resultados ***(Média acerto classes)***		Balanceamento		Sem balanceamento
		Aumento 10 E 11	Aumento 6 a 11	
Corpus	Modelo SVM	Smote	Smote	
Título	Etapa única	**80,26%**	N/A	80,34%
	2 etapas (4-7)	**80,48%**	77,53%	79,45%
	2 etapas (6-5)	75,25%	N/A	**79,27%**
Descrição	Etapa única	N/A	N/A	70,97%
	2 etapas (4-7)	N/A	N/A	70,55%

Para o corpus relativo à descrição detalhada das anormalidades, dado que o uso deste tipo de dado não obteve bom desempenho, foi avaliado somente a configuração em etapa única somente (sem balanceamento ou classificação hierárquica). Observa-se que o mesmo continuou a apresentar resultados inferiores ao corpus relativo ao título das anormalidades. O aumento das classes de 6 a 11 foi simulado para um caso somente para comparação.

São apresentadas nas figuras abaixo as matrizes de confusão dos três melhores resultados para micro F-score:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	95,89%	0,68%	0,68%	1,71%	0,00%	0,00%	0,00%	0,68%	0,34%	0,00%	0,00%
	2	5,16%	93,55%	0,65%	0,65%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	3,27%	3,27%	91,50%	0,65%	1,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	2,81%	0,56%	0,00%	96,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	6,06%	3,03%	90,91%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	3,57%	3,57%	0,00%	0,00%	0,00%	92,86%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	23,08%	3,85%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	16,00%	4,00%	20,00%	0,00%	0,00%	0,00%	44,00%	8,00%	0,00%	0,00%
	9	23,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	0,00%	0,00%	11,76%	0,00%	0,00%	5,88%	5,88%	41,18%	11,76%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			79,27%	Micro-Fscore		65,07%			

Figura 76: 1º colocado para micro F-score: SVC hierárquico de (6 classes na etapa 1 e 5 classes na etapa 2) sem balanceamento.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	94,52%	1,71%	0,34%	2,40%	0,34%	0,00%	0,00%	0,00%	0,00%	0,68%	0,00%
	2	5,16%	92,26%	1,29%	0,65%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%
	3	3,27%	2,61%	91,50%	0,65%	1,96%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	2,81%	0,56%	0,00%	96,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	6,06%	3,03%	87,88%	0,00%	0,00%	0,00%	0,00%	3,03%	0,00%
	6	3,57%	0,00%	0,00%	0,00%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	23,08%	3,85%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	12,00%	0,00%	24,00%	0,00%	0,00%	0,00%	52,00%	0,00%	4,00%	0,00%
	9	23,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	69,23%	7,69%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	5,88%	0,00%	0,00%	0,00%	5,88%	52,94%	0,00%
	11	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			80,26%	Micro-Fscore		64,16%			

Figura 77: 2º colocado para micro F-score: SVC etapa única (11 classes) com balanceamento das classes 10 e 11 por SMOTE.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	91,44%	2,05%	0,68%	4,45%	0,00%	0,00%	0,00%	0,34%	0,34%	0,68%	0,00%
	2	4,52%	92,26%	1,29%	1,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	0,65%	3,92%	92,16%	1,31%	1,96%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	1,69%	0,00%	0,00%	98,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	6,06%	3,03%	90,91%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	3,57%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	23,08%	3,85%	3,85%	3,85%	0,00%	0,00%	65,38%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	8,00%	8,00%	24,00%	0,00%	0,00%	0,00%	48,00%	0,00%	4,00%	0,00%
	9	23,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	5,88%	0,00%	11,76%	0,00%	0,00%	0,00%	5,88%	52,94%	0,00%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			80,48%	Micro-Fscore		62,70%			

Figura 78: 3º colocado para micro F-score: SVC hierárquico de (4 classes na etapa 1 e 7 classes na etapa 2) com balanceamento das classes 10 e 11 por SMOTE.

Primeiramente observa-se que as novas propostas sugeridas para este estudo de caso, sejam elas: a classificação hierárquica (6 classes na primeira etapa e 5 classes na segunda etapa) e o balanceamento somente das classes 10 e 11, melhoraram o desempenho do modelo em geral, quando se compara os melhores desempenhos antes e após as mudanças (ou seja, comparando as matrizes de confusão da Figura 74 com a Figura 76). Para a métrica micro F-score, o desempenho foi de 61,84 para 65,07% (aumento de 3,23 pontos percentuais) e para a métrica da média de acerto das classes de 80,34 para 79,27% (ligeira queda).

Após analisar as 3 matrizes de confusão para o estudo de caso atual (estudo 8) com os melhores resultados para a métrica micro F-score, observa-se que há uma relação de compromisso entre a métrica micro F-score e a métrica da média de acerto entre as classes: a medida que há melhoria no desempenho do micro F-score, há uma queda no desempenho pela média de acerto das classes (mesmo que pequeno). Ainda, observa-se que esta métrica tem seu desempenho aumentado quando da ocorrência de balanceamento da base de dados.

5.4.
Estudos de caso após segunda correção da base de dados

Devido à acentuada melhoria no desempenho do modelo, decidiu-se efetuar uma última correção, mais abrangente que a primeira. Em seguida, seriam treinados mais uma vez os classificadores.

Para esta correção foi utilizado o melhor modelo até então encontrado, de classificação hierárquica, sendo a primeira etapa para 6 classes e a segunda etapa para 5 classes, sem balanceamento da base de dados.

Todos os erros do classificador nesta configuração (tanto na fase de treinamento/validação bem como na fase de teste) foram analisados. No total, 2047 documentos foram avaliados. Em seguida, construiu-se uma matriz de confusão somente para esta revisão, de modo a avaliar detalhadamente as falhas de preenchimento:

		APONTADO PELO OPERADOR											Total avaliado por classe	%	Total de erros por classe	%
		1	2	3	4	5	6	7	8	9	10	11				
APONTADO PELA CORREÇÃO	1	78	34	6	68	45	10	15	14	11	0	0	281	13,73%	203	72,24%
	2	15	173	85	29	4	11	2	10	2	2	0	333	16,27%	160	48,05%
	3	8	24	211	3	20	0	0	7	3	0	0	276	13,48%	65	23,55%
	4	200	8	33	167	8	8	4	6	1	2	0	437	21,35%	382	87,41%
	5	3	1	7	0	55	0	0	1	3	0	0	70	3,42%	15	21,43%
	6	2	1	0	2	0	54	0	0	0	1	0	60	2,93%	6	10,00%
	7	34	7	2	1	17	0	63	0	0	0	0	124	6,06%	61	49,19%
	8	11	6	12	8	7	2	2	135	6	3	0	192	9,38%	57	29,69%
	9	2	0	1	1	5	0	0	2	106	23	0	140	6,84%	34	24,29%
	10	1	1	3	0	2	0	0	2	20	95	2	126	6,16%	31	24,60%
	11	0	0	0	0	0	0	0	0	0	0	8	8	0,39%	0	0,00%

Figura 79: Matriz de confusão do operador (segunda correção): valores absolutos.

		APONTADO PELO OPERADOR											
		1	2	3	4	5	6	7	8	9	10	11	
APONTADO PELA CORREÇÃO	1	27,76%	12,10%	2,14%	24,20%	16,01%	3,56%	5,34%	4,98%	3,91%	0,00%	0,00%	
	2	4,50%	51,95%	25,53%	8,71%	1,20%	3,30%	0,60%	3,00%	0,60%	0,60%	0,00%	Percentual de acerto
	3	2,90%	8,70%	76,45%	1,09%	7,25%	0,00%	0,00%	2,54%	1,09%	0,00%	0,00%	
	4	45,77%	1,83%	7,55%	38,22%	1,83%	1,83%	0,92%	1,37%	0,23%	0,46%	0,00%	55,94%
	5	4,29%	1,43%	10,00%	0,00%	78,57%	0,00%	0,00%	1,43%	4,29%	0,00%	0,00%	Total acertos
	6	3,33%	1,67%	0,00%	3,33%	0,00%	90,00%	0,00%	0,00%	0,00%	1,67%	0,00%	operador
	7	27,42%	5,65%	1,61%	0,81%	13,71%	0,00%	50,81%	0,00%	0,00%	0,00%	0,00%	1.145
	8	5,73%	3,13%	6,25%	4,17%	3,65%	1,04%	1,04%	70,31%	3,13%	1,56%	0,00%	Total avaliado
	9	1,43%	0,00%	0,71%	0,71%	3,57%	0,00%	0,00%	1,43%	75,71%	16,43%	0,00%	2.047
	10	0,79%	0,79%	2,38%	0,00%	1,59%	0,00%	0,00%	1,59%	15,87%	75,40%	1,59%	
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	

Figura 80: Matriz de confusão do operador (segunda correção): valores percentuais.

Pode-se observar que a classe 4 é altamente confundida com a classe 1 pelo operador. Avaliando com especialistas este resultado, os mesmos observaram que estes equipamentos possuem uma interface física muito próxima e que muitas vezes os equipamentos próximos a essa interface foram confundidos no seu apontamento.

Outro ponto observado durante as correções foi de que as falhas eram apontadas como sua causa raiz e não como sua causa aparente. Seguem dois exemplos para ilustrar esta constatação:

- Apontamento para a falha de rompimento da coluna por compressão (classe X) devido à falha no compensador de movimentos (classe Y);
- Apontamento para falha de propulsão (classe M), mas a causa raiz ocorreu devido à falha no fornecimento de energia ao mesmo (classe N).

Como várias falhas de sonda só são esclarecidas em sua totalidade após investigação do evento, a recomendação é sempre apontar a causa aparente no título das anormalidades e, na descrição da mesma, mencionar possíveis causas raiz. Em seguida, após análise do setor de contratos, deve-se aprofundar a investigação de forma a atribuir a causa raiz. Esta correção levou esta premissa em consideração.

Estas classificações não adequadas do operador (tanto relacionada à classe 4 bem como à definição de causa raiz e causa aparente) trouxeram mais subsídio ao processo de preenchimento de anormalidades de sonda, no cenário específico da Petrobras.

Todos os erros de classificação observados nesta análise tiveram sua classe alterada com base nesta última correção.

Para visualizar de maneira mais prática as alterações na base de dados após esta extensa correção, segue uma tabela comparando os percentuais das classes na base de dados antes e após as correções:

Tabela 25: Comparação da distribuição das classes antes e após as correções, tanto para a base de treinamento/validação quanto para a base de teste.

	BASE DE DADOS - TREINAMENTO/VALIDAÇÃO				BASE DE DADOS - TESTE			
	INICIAL		APÓS CORREÇÃO		INICIAL		APÓS CORREÇÃO	
Classe	Frequência	% em rel. à base de dados de treino/valid.	Frequência	% em rel. à base de dados de treino/valid.	Frequência	% em rel. à base de dados de teste	Frequência	% em rel. à base de dados de teste
1	2774	33,50%	2644	31,93%	303	32,86%	292	31,67%
2	1456	17,58%	1515	18,30%	151	16,38%	155	16,81%
3	1414	17,08%	1323	15,98%	162	17,57%	153	16,59%
4	1370	16,55%	1579	19,07%	152	16,49%	178	19,31%
5	389	4,70%	269	3,25%	43	4,66%	33	3,58%
6	236	2,85%	272	3,29%	30	3,25%	28	3,04%
7	181	2,19%	216	2,61%	28	3,04%	26	2,82%
8	167	2,02%	194	2,34%	20	2,17%	25	2,71%
9	156	1,88%	141	1,70%	16	1,74%	13	1,41%
10	127	1,53%	128	1,55%	15	1,63%	17	1,84%
11	10	0,12%	9	0,11%	2	0,22%	2	0,22%

5.4.1. Estudo de caso 9: rodada final após segunda correção

Pela última vez foi treinada e testada toda a base de dados com base na segunda correção. As Tabelas 27 e 28 trazem um resumo do desempenho das configurações adotadas, comparando os valores antes e após esta última correção:

Tabela 26: Resumo dos melhores modelos antes da segunda correção. Obs.: “4-7” refere-se ao hierárquico com classificação de 4 classes inicialmente, seguido das demais 7 classes. A mesma lógica se aplica ao termo “6-5”.

Melhores resultados ***(μ F-Score)***		Balanceamento		Sem balanceamento
		Aumento 10 E 11	Aumento 6 a 11	
Corpus	Modelo SVM	Smote	Smote	
Título	Etapa única	**64,16%**	N/A	61,84%
	2 etapas (4-7)	**62,70%**	60,18%	61,28%
	2 etapas (6-5)	53,00%	N/A	**65,07%**
Descrição	Etapa única	N/A	N/A	48,74%
	2 etapas (4-7)	N/A	N/A	48,94%

Melhores resultados ***(Média acerto classes)***		Balanceamento		Sem balanceamento
		Aumento 10 E 11	Aumento 6 a 11	
Corpus	Modelo SVM	Smote	Smote	
Título	Etapa única	**80,26%**	N/A	80,34%
	2 etapas (4-7)	**80,48%**	77,53%	79,45%
	2 etapas (6-5)	75,25%	N/A	**79,27%**
Descrição	Etapa única	N/A	N/A	70,97%
	2 etapas (4-7)	N/A	N/A	70,55%

Tabela 27: Resumo dos melhores modelos após segunda correção.

Melhores resultados ***(μ F-Score)***		Balanceamento		Sem balanceamento
		Aumento 10 E 11	Aumento 6 a 11	
Corpus	Modelo SVM	Smote	Smote	
Título	Etapa única	**62,13%**	N/A	**61,84%**
	2 etapas (4-7)	53,23%	60,18%	61,28%
	2 etapas (6-5)	57,05%	N/A	**62,99%**
Descrição	Etapa única	N/A	N/A	51,24%
	2 etapas (4-7)	N/A	N/A	51,67%
Melhores resultados ***(Média acerto classes)***		Balanceamento		Sem balanceamento
		Aumento 10 E 11	Aumento 6 a 11	
Corpus	Modelo SVM	Smote	Smote	
Título	Etapa única	**79,89%**	N/A	**80,66%**
	2 etapas (4-7)	74,49%	78,11%	78,10%
	2 etapas (6-5)	77,14%	N/A	**79,89%**
Descrição	Etapa única	N/A	N/A	75,60%
	2 etapas (4-7)	N/A	N/A	74,48%

Observa-se que o desempenho dos classificadores apresentou ligeira queda. Este fato pode ser explicado pelo método utilizado para efetuar as correções. Para uma correção mais eficiente, foi utilizada a heurística de avaliar as classes atribuídas de forma errada pelo classificador. Porém em momento algum foram conferidas as observações onde o classificador acertou. Dada a taxa de erro do operador, o que garante que o mesmo pode ter atribuído classes erradas e o classificador ter aprendido esta atribuição errônea? O seguinte exemplo pode explicitar esta ideia: supondo que o operador atribuiu à classe X à observação N1, mas esta observação pertencesse realmente à classe W. E que o classificador fosse induzido ao erro em sua aprendizagem (devido à repetição do erro em apontar como W a classe X). Dado que o princípio da correção da base de dados focou nos erros do classificador, quando comparados às classificações do operador (que por vezes não era a real, ou seja, a verdadeira classe), muitas anormalidades foram aprendidas de forma errônea e também avaliadas de forma errônea na etapa de teste.

Quando se efetuou a primeira correção, onde se atuou somente na base de teste, a avaliação começou a ficar mais confiável, mas a base de treinamento ainda

carecia de correções. Após a segunda correção, onde o foco foi maior na base de aprendizado e treinamento, apesar da queda no desempenho, pode-se inferir que a confiabilidade dos resultados aumentou.

Para os três melhores modelos, seguem as matrizes de confusão (tanto absolutas quanto percentuais):

- 1º lugar:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	266	7	1	16	1	0	1	0	0	0	0
	2	7	144	3	0	0	0	0	1	0	0	0
	3	1	6	142	2	2	0	0	0	0	0	0
	4	1	2	0	175	0	0	0	0	0	0	0
	5	0	0	1	1	31	0	0	0	0	0	0
	6	0	0	0	1	0	27	0	0	0	0	0
	7	5	2	0	1	0	0	18	0	0	0	0
	8	2	1	1	7	0	0	0	13	1	0	0
	9	2	0	0	1	0	0	0	0	10	0	0
	10	2	2	2	0	2	0	0	0	2	6	1
	11	0	0	0	0	0	0	0	0	0	1	1

Figura 81: Matriz de confusão com valores absolutos do melhor modelo (micro F-score) após a segunda correção.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	91,10%	2,40%	0,34%	5,48%	0,34%	0,00%	0,34%	0,00%	0,00%	0,00%	0,00%
	2	4,52%	92,90%	1,94%	0,00%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%
	3	0,65%	3,92%	92,81%	1,31%	1,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	0,56%	1,12%	0,00%	98,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	3,03%	3,03%	93,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	3,57%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	19,23%	7,69%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	4,00%	4,00%	28,00%	0,00%	0,00%	0,00%	52,00%	4,00%	0,00%	0,00%
	9	15,38%	0,00%	0,00%	7,69%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	11,76%	0,00%	0,00%	0,00%	11,76%	35,29%	5,88%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			79,89%	Micro-Fscore		62,99%			

Figura 82: Matriz de confusão com valores percentuais do melhor modelo (micro F-score) após a segunda correção.

- 2º lugar:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	265	9	1	14	1	0	1	0	0	1	0
	2	10	143	2	0	0	0	0	0	0	0	0
	3	6	5	140	1	1	0	0	0	0	0	0
	4	2	2	0	174	0	0	0	0	0	0	0
	5	0	0	3	1	28	0	0	0	0	1	0
	6	1	0	0	0	0	27	0	0	0	0	0
	7	5	1	0	1	0	0	19	0	0	0	0
	8	2	1	1	7	0	0	0	14	0	0	0
	9	3	0	0	0	0	0	0	0	10	0	0
	10	2	2	2	0	2	0	0	0	1	8	0
	11	0	0	0	0	0	0	0	0	0	1	1

Figura 83: Matriz de confusão com valores absolutos do modelo com o segundo melhor desempenho (micro F-score) após a segunda correção.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	90,75%	3,08%	0,34%	4,79%	0,34%	0,00%	0,34%	0,00%	0,00%	0,34%	0,00%
	2	6,45%	92,26%	1,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	3,92%	3,27%	91,50%	0,65%	0,65%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	1,12%	1,12%	0,00%	97,75%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	9,09%	3,03%	84,85%	0,00%	0,00%	0,00%	0,00%	3,03%	0,00%
	6	3,57%	0,00%	0,00%	0,00%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	19,23%	3,85%	0,00%	3,85%	0,00%	0,00%	73,08%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	4,00%	4,00%	28,00%	0,00%	0,00%	0,00%	56,00%	0,00%	0,00%	0,00%
	9	23,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	11,76%	0,00%	0,00%	0,00%	5,88%	47,06%	0,00%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			80,66%	Micro-Fscore		61,84%			

Figura 84: Matriz de confusão com valores percentuais do melhor modelo (micro F-score) após a segunda correção.

- 3º lugar:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	266	6	1	12	1	1	2	1	1	1	0
	2	6	142	3	0	0	0	2	1	1	0	0
	3	5	5	139	1	0	0	1	0	1	1	0
	4	2	2	0	174	0	0	0	0	0	0	0
	5	0	0	1	1	31	0	0	0	0	0	0
	6	2	2	0	0	0	23	1	0	0	0	0
	7	5	1	0	1	0	2	17	0	0	0	0
	8	1	2	1	7	0	0	1	13	0	0	0
	9	2	0	1	0	0	0	0	0	9	1	0
	10	2	2	1	0	2	0	1	1	0	8	0
	11	0	0	0	0	0	0	0	0	0	1	1

Figura 85: Matriz de confusão com valores absolutos do modelo com o terceiro melhor desempenho (micro F-score) após a segunda correção.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	92,81%	2,40%	0,34%	3,77%	0,34%	0,00%	0,34%	0,00%	0,00%	0,00%	0,00%
	2	6,45%	91,61%	1,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,58%	3,92%	90,20%	0,65%	0,65%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	1,12%	1,12%	0,00%	97,75%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	9,09%	3,03%	84,85%	0,00%	0,00%	0,00%	0,00%	3,03%	0,00%
	6	3,57%	0,00%	0,00%	0,00%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	23,08%	3,85%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	4,00%	8,00%	28,00%	0,00%	0,00%	0,00%	52,00%	0,00%	0,00%	0,00%
	9	23,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	11,76%	0,00%	0,00%	0,00%	5,88%	47,06%	0,00%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			79,89%	Micro-Fscore		62,13%			

Figura 86: Matriz de confusão com valores percentuais do terceiro melhor modelo (micro F-score) após a segunda correção.

As configurações dos três melhores modelos podem ser visualizadas na Tabela 28:

Tabela 28: Resumo dos parâmetros utilizados nos melhores modelos após segunda correção.

	1º lugar		**2º lugar**	**3º lugar**
	Etapa 1	Etapa 2	Etapa única	Etapa única
Corpus	Título		Título	Título
Peso aplicado à MTD	TF-IDF		Binário	TF
Fator de redução de esparcidade	E1=1	E2=0,9998	E1=1	E1=1
Balanceamento	Não utilizado		Não utilizado	Aumento das classes 10 e 11 por SMOTE
Valores de constante do SVC	15	20	0,7	7
Tempo de processamento (minutos)	9		5	8
Métrica micro F-score do mesmo modelo (sem segunda correção)	65,07%		64,16%	62,70%

Comparação do primeiro estudo de caso com o último estudo de caso (melhores modelos):

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	82,51%	2,31%	0,00%	13,53%	0,00%	0,00%	0,99%	0,33%	0,00%	0,33%	0,00%
	2	6,62%	88,08%	3,97%	1,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	4,94%	6,17%	80,86%	6,17%	1,23%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%
	4	5,92%	1,97%	1,97%	90,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	20,93%	0,00%	6,98%	0,00%	67,44%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%
	6	3,33%	6,67%	0,00%	0,00%	0,00%	90,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	50,00%	0,00%	0,00%	3,57%	0,00%	0,00%	46,43%	0,00%	0,00%	0,00%	0,00%
	8	15,00%	15,00%	5,00%	20,00%	0,00%	0,00%	0,00%	45,00%	0,00%	0,00%	0,00%
	9	37,50%	0,00%	0,00%	6,25%	0,00%	0,00%	0,00%	0,00%	56,25%	0,00%	0,00%
	10	20,00%	13,33%	6,67%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	33,33%	0,00%
	11	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%
			Média das classes			68,00%	Micro-Fscore		43,18%			

Figura 87: Matriz de confusão (valores percentuais) do melhor modelo pela métrica micro F-score: SVC de 11 classes (etapa única) sem balanceamento, caso base.

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	91,10%	2,40%	0,34%	5,48%	0,34%	0,00%	0,34%	0,00%	0,00%	0,00%	0,00%
	2	4,52%	92,90%	1,94%	0,00%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%
	3	0,65%	3,92%	92,81%	1,31%	1,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	0,56%	1,12%	0,00%	98,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	3,03%	3,03%	93,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	3,57%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	19,23%	7,69%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	4,00%	4,00%	28,00%	0,00%	0,00%	0,00%	52,00%	4,00%	0,00%	0,00%
	9	15,38%	0,00%	0,00%	7,69%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	11,76%	0,00%	0,00%	0,00%	11,76%	35,29%	5,88%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			79,89%	Micro-Fscore		62,99%			

Figura 88: Matriz de confusão com valores percentuais do melhor modelo (micro F-score) após a segunda correção.

Comparando o melhor modelo antes e após a segunda correção, observa-se que mesmo que o desempenho geral tenha apresentando leve queda, as classes 4, 5, 6, 8 apresentaram aumento na taxa de acerto. Estas classes foram as mais corrigidas nesta última varredura. Já as demais classes ou apresentaram pequena queda de desempenho (como as classes 1, 2, 3) ou se mantiveram em patamares semelhantes.

5.5. Resultado de teste de significância estatística

O intuito desta seção é avaliar de forma estatística os resultados obtidos, para trazer mais robustez às analises efetuadas. Para isso, foi utilizado o teste não paramétrico de Wilcoxon (*signed-rank*). Um maior detalhamento relativo ao teste pode ser encontrado em (Conover, 1980). O valor do nível de significância (p-value) utilizado foi de 5%. O *base line* (avaliação inicial do operador), bem como o melhor modelo obtido sem efetuar correção na base de dados, foram comparados com o melhor modelo obtido após todas as correções. O teste foi realizado comparando-se o percentual de acerto classe a classe (em pares), para as 11 classes (*paired rank test*).

A Tabela 29 apresenta os resultados para o teste de Wilcoxon, de forma a constatar que o melhor modelo obtido após as correções conseguiu ter diferenças estatisticamente significativas quando comparado aos demais casos.

Tabela 29: Resultados para o Teste de Wilcoxon comparando os casos de Base Line e melhor modelo sem correção com o melhor modelo obtido após todas as correções da base de dados.

Caso	p-valor	Lim. Inferior	Lim. Superior	Conclusão
Base line	0,02441	0,01995	0,28985	Modelo com correção é melhor
Melhor modelo sem correção	0,005922	0,05629	0,17533	Modelo com correção é melhor

Finalmente, foi efetuado o teste de Wilcoxon comparando o *base line* com o melhor modelo obtido sem correção da base de dados. Para este caso também se observou diferença significativa:

Tabela 30: Resultados para o Teste de Wilcoxon comparando o caso de Base Line com o melhor modelo obtido antes das correções da base de dados.

Caso	p-valor	Lim. Inferior	Lim. Superior	Conclusão
Base line	0,5771	-0,08154	0,16429	Modelo sem correção é melhor

5.6. Correção do passivo de anormalidades

Para efetuar a correção do passivo de anormalidades foi escolhida a configuração do modelo que obteve o melhor desempenho na métrica *micro-Fscore*, pois este possui o maior percentual de acerto total, apesar de não possuir o maior percentual de acerto entre classes. Esta escolha se baseia na opção pelo maior acerto absoluto esperado para a classificação do passivo.

Em seguida, a base de dados utilizada para treinamento, validação e teste foi consolidada em uma só base de dados. Esta base consolidada foi utilizada para treinar o modelo de melhor configuração. Finalmente, com este modelo treinado, submeteu-se a base de dados não classificada (ou seja, o passivo das anormalidades sem classificação) a este modelo treinado, para obter as classes destas anormalidades. A Figura 89 esquematiza este processo:

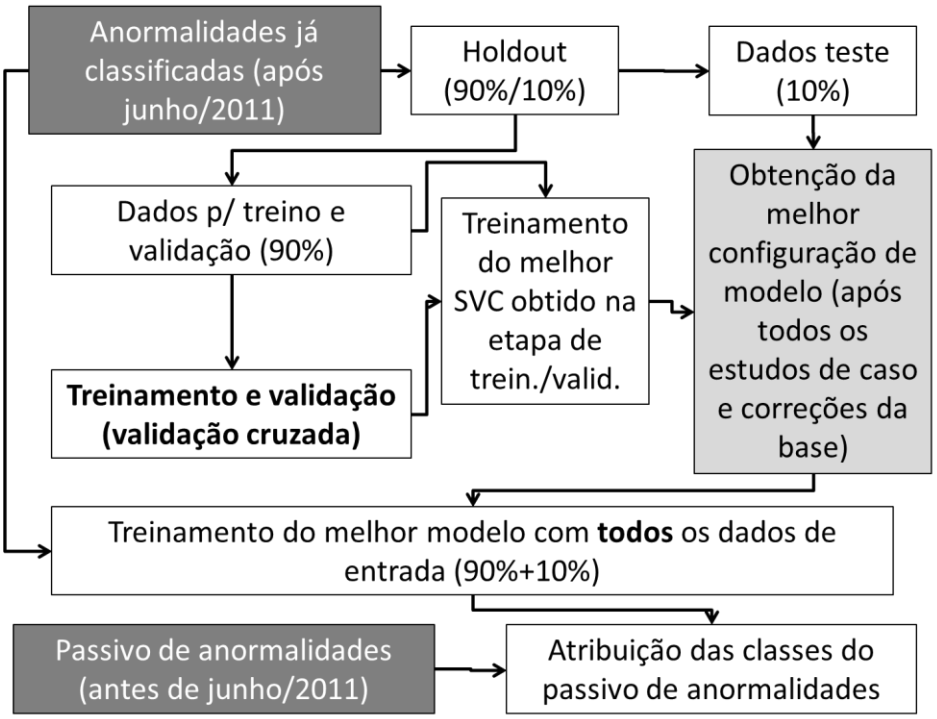

Figura 89: Fluxograma para classificação do passivo de anormalidades após todo o processo de obtenção do melhor modelo.

Importante destacar que os dados de entrada relativos às anormalidades do passivo foram adequados antes de serem submetidos à melhor configuração do modelo, de modo que o classificador recebesse estes dados de entrada de forma coerente com seu treinamento, validação e teste. Abaixo, a configuração do melhor classificador obtido:

- Corpus de ambas as etapas: título da anormalidade;
- Peso aplicado à MTD em ambas as etapas: TF-IDF;
- Sem aplicação do fator de redução de esparcidade na etapa 1 e aplicação do fator E2 na etapa 2;
- Não utilizando a duração da anormalidade na etapa 1 e utilizando-a na etapa 2;
- Sem balanceamento da base de dados de entrada em nenhuma das etapas.

Para o classificador de melhor desempenho, os seguintes parâmetros foram utilizados:

- Classificador SVC Hierárquico (2 etapas: 6 + 5 classes), utilizando os valores de constante de penalização de C1=15 e C2=20 para a primeira e segunda etapas, respectivamente.

A Tabela 31 ilustra a distribuição das classes obtidas após a classificação do passivo de 3384 anormalidades pelo classificador:

Tabela 31: Distribuição das classes obtidas após a classificação do passivo de anormalidades.

Classe	Frequência	Percentual
1	1174	34,69%
2	346	10,22%
3	644	19,03%
4	720	21,28%
5	128	3,78%
6	128	3,78%
7	54	1,60%
8	71	2,10%
9	75	2,22%
10	43	1,27%
11	1	0,03%
	3384	

Finalmente, foi efetuada uma pequena avaliação da aplicação do modelo, sorteando 208 observações do passivo de 3384 anormalidades classificadas pelo modelo e elaborada sua matriz de confusão com valores absolutos:

		PREDITO											
		1	2	3	4	5	6	7	8	9	10	11	Freq.
REAL	1	23	0	0	0	0	0	0	0	4	2	0	29
	2	0	25	1	1	0	0	0	0	0	0	0	27
	3	0	0	23	0	0	0	0	0	0	0	0	23
	4	1	0	0	21	0	0	0	0	0	0	0	22
	5	0	0	1	0	19	0	0	0	0	0	0	20
	6	1	1	0	0	0	21	0	0	0	0	0	23
	7	1	0	0	0	0	0	17	0	0	0	0	18
	8	2	0	0	0	0	0	0	16	0	0	0	18
	9	0	0	0	0	0	0	0	0	13	0	0	13
	10	0	0	0	0	0	0	0	0	0	14	0	14
	11	0	0	0	0	0	0	0	0	0	0	1	1
													208

Figura 90: Matriz de confusão com valores absolutos para classificação do passivo de acordo com o melhor modelo obtido.

Abaixo, segue a matriz de confusão para valores percentuais, bem como os valores das métricas Micro F-score e de média de acerto entre as classes. Com o intuito de comparar a matriz de confusão com valores percentuais do melhor modelo em fase de teste (Figura 82) com o desempenho deste mesmo modelo para classificar o passivo de anormalidades, utilizou-se de um código de cores: na diagonal principal estão destacadas (em cor mais escura e numeração em branco) as classes onde o acerto foi maior que o modelo em fase de teste (após a segunda correção). Os valores na diagonal principal em cor mais clara e numeração em preto correspondem às classes onde o modelo foi superior em fase de teste, quando comparado ao mesmo modelo aplicado sobre o passivo de anormalidades:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	79,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	13,79%	6,90%	0,00%
	2	0,00%	92,59%	3,70%	3,70%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	3	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	4,55%	0,00%	0,00%	95,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	5,00%	0,00%	95,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	4,35%	4,35%	0,00%	0,00%	0,00%	91,30%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	5,56%	0,00%	0,00%	0,00%	0,00%	0,00%	94,44%	0,00%	0,00%	0,00%	0,00%
	8	11,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	88,89%	0,00%	0,00%	0,00%
	9	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
	10	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
			Média das classes			94,27%	Micro-Fscore		70,05%			

Figura 91: Matriz de confusão com valores percentuais do passivo classificado de acordo com o melhor modelo obtido.

Para fins de comparação com a matriz acima, segue matriz de confusão com valores percentuais para o teste do melhor modelo após a segunda correção:

		PREDITO										
		1	2	3	4	5	6	7	8	9	10	11
REAL	1	91,10%	2,40%	0,34%	5,48%	0,34%	0,00%	0,34%	0,00%	0,00%	0,00%	0,00%
	2	4,52%	92,90%	1,94%	0,00%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%
	3	0,65%	3,92%	92,81%	1,31%	1,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	4	0,56%	1,12%	0,00%	98,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	5	0,00%	0,00%	3,03%	3,03%	93,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	6	0,00%	0,00%	0,00%	3,57%	0,00%	96,43%	0,00%	0,00%	0,00%	0,00%	0,00%
	7	19,23%	7,69%	0,00%	3,85%	0,00%	0,00%	69,23%	0,00%	0,00%	0,00%	0,00%
	8	8,00%	4,00%	4,00%	28,00%	0,00%	0,00%	0,00%	52,00%	4,00%	0,00%	0,00%
	9	15,38%	0,00%	0,00%	7,69%	0,00%	0,00%	0,00%	0,00%	76,92%	0,00%	0,00%
	10	11,76%	11,76%	11,76%	0,00%	11,76%	0,00%	0,00%	0,00%	11,76%	35,29%	5,88%
	11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	50,00%
			Média das classes			79,89%	Micro-Fscore		62,99%			

Figura 92: Matriz de confusão com valores percentuais do melhor modelo (micro F-score) após a segunda correção.

Observa-se que o modelo classificou o passivo muito bem, principalmente para as classes menores. O valor de Micro F-score é 7,06 pontos percentuais maior, e a média de acerto entre classes é 14,38 pontos percentuais maior, mostrando bom desempenho do modelo para a amostra de 208 anormalidades (de um total de 3384). Para as classes 2, 4 e 6, apesar do desempenho ter sido menor na comparação, observa-se que mesmo assim o passivo foi corrigido com um alto patamar de acerto. Somente a classe 1 possui um pequena queda no desempenho. Isto pode ser explicado pelo seu tamanho, de forma que o modelo acaba por confundir algumas anormalidades dentro desta numerosa classe.

Além da correção do passivo de anormalidades, há também a expectativa de uso deste modelo para avaliar inconsistências de preenchimento do banco de dados de falha de equipamentos das unidades de intervenção de forma periódica e com isso trazer mais confiabilidade à base de dados de construção de poços de petróleo submarinos da Petrobras.

6.
Conclusões e trabalhos futuros

6.1. Conclusões

Esta dissertação teve como objetivo principal desenvolver um classificador de anormalidades de equipamentos de sonda para a Petrobras. A ferramenta atualizou o passivo de anormalidades relacionadas às falhas de equipamentos de unidades de intervenção ocorridas em intervenções de perfuração em sondas de posicionamento dinâmico preenchidas do início de 2008 até junho de 2011, com base em uma nova classificação que foi aplicada a partir desta data.

Para atender a esta demanda, foi efetuado o tratamento de uma base de dados já classificada de acordo com a nova estrutura de classificação, sendo esta composta majoritariamente por variáveis textuais. A manipulação dos dados se deu em ambiente de linguagem de programação R, utilizando ferramentas de manipulação de dados, mineração textual bem como de aplicação do algoritmo de aprendizado de máquina *Support Vector Machines*, na modalidade mais usual: para classificação (SVC – *Support Vector Classification*).

A ferramenta foi aplicada para avaliação de 9 estudos de caso, por meio da métrica principal micro F-score e da métrica auxiliar da média de acerto percentual das classes. A métrica macro F-score, indicada como ideal pela literatura para o problema de várias classes, não teve aplicação favorável neste trabalho.

Foram avaliados balanceamento das classes, bem como a utilização do método de classificação hierárquico. Estas abordagens foram testadas de forma separada e em conjunto. O classificador final obteve micro F-score de 62,99% e média de acerto entre classes de 79,89%. Comparado a testes efetuados com os operadores, que obtiveram micro F-score de 35,63% e média de acerto entre classes de 67,59%, pode-se concluir que, mesmo considerando as correções efetuadas na base de dados, o classificador obteve desempenho de destaque.

Todo o trabalho demandou o mínimo de participação dos operadores, já sinalizados como recursos humanos de disponibilidade crítica na empresa. A criação da lista de sinônimos, multi-termos bem como *stop lists* foram os pontos de maior necessidade de conhecimento técnico para sua elaboração, além das correções de classificação da base de dados.

A configuração do modelo de melhor desempenho utilizou a seguinte configuração:

- Dados de entrada:
    - Campo textual: título das anormalidades;
    - Peso dos termos da matriz de termos por documentos: TF-IDF;
    - Sem redução de esparcidade na primeira etapa, e com redução parcial de esparcidade na segunda etapa;
    - Não utilizando a duração na primeira etapa e utilizando a duração das anormalidades na segunda etapa;
    - Sem balanceamento da base de dados (apesar desta ser altamente desbalanceada).
- Abordagem de classificação:
    - Classificação em duas etapas, pelo método hierárquico.
    - Tempo de processamento em torno de 9 minutos.

Para alcançar o resultado final foram necessárias duas correções na base de dados. Comparando os resultados do modelo atuando sobre a base de dados sem correção e após as correções, houve aumento do desempenho de 43,18% para o valor já apontado de 62,99% (métrica micro F-score) e para média de acerto entre classes o desempenho foi de 68,00% para 79,89%.

Ao efetuar a avaliação final do desempenho do modelo ao classificar o passivo de anormalidades observou-se que o melhor modelo atendeu as expectativas, com alto índice de acerto tanto para a taxa média de acerto entre classes (94,27%) quanto para a métrica micro F-score (70,05%).

Todo o processo também agregou valor com relação à necessidade de orientação dos operadores quanto ao preenchimento adequado das anormalidades no banco de dados. Destaque especial para classificação da classe 4, muito confundida com a classe 1. Tal fato ocorreu porque os equipamentos destas classes possuem uma interface que muitas vezes é confundida pelo operador,

acarretando em preenchimento não adequado. Esta lição aprendida será incorporada ao processo de preenchimento.

Este trabalho gerou discussões relevantes sobre a forma adequada de preencher o banco de dados quando se trata da classe de falha de equipamentos. Geralmente a falha de equipamento possui causa aparente e algumas vezes, após investigação, é apontada uma causa raiz, diferente da causa aparente. Também esta lição aprendida será incorporada ao processo de preenchimento da base de dados, pois antes disso sua necessidade não havia sido aplicada em plenitude.

Finalmente, destaca-se a necessidade de suprimir algumas classes de tamanhos muito pequenos (como as classes 10 e 11), de modo a inseri-las em outras classes. O mesmo se aplica à numerosa classe 1, que poderia ser dividida em classes menores. Isto traria mais uniformidade, não só em análises de mineração de dados, mas na análise das falhas de equipamentos de sonda como um todo.

6.2. Trabalhos futuros

Podem ser destacadas as seguintes oportunidades para trabalhos futuros:

- Conforme apontado em (Prati, Batista, Silva, 2015), como o balanceamento da base de dados é extenso em problemas que envolvem a mineração textual, sugere-se avaliar o uso de métricas para avaliar a efetividade do processo de balanceamento, de forma a verificar se os tratamentos efetuados realmente agregam valor ou se não interferem no objetivo final. Nesta referência são mencionadas outras métricas para efetuar a avaliação específica do balanceamento de bases onde as classes estão desigualmente distribuídas.
- Avaliar e propor uma nova forma de utilizar a métrica Macro F-score para cenários semelhantes ao deste trabalho: várias classes, sendo que algumas delas possuíam poucas observações quando comparadas às demais classes. A ideia seria adequar a métrica para pleitear casos em que ocorra divisão por zero em alguns dos numeradores de sua fórmula, porém de forma a

não penalizar drasticamente as configurações de modelo onde este evento ocorra.

- Especializar a ferramenta elaborada neste trabalho de forma a aplicá-la de maneira contínua (*online*) sobre os bancos de dados da empresa, agregando valor ao processo de tratamento e adequação de anormalidades preenchidas pelos operadores.
- Expansão do método hierárquico, de modo que sejam inseridos mais SVCs para se especializar e classificar classes menores.

7. Referências bibliográficas

AIZERMAN E., BRAVERMAN E., ROZONOER L. **Theoretical foundations of the potential function method in pattern recognition learning**. Automation and Remote Control, 25 : 821-837, 1964.

ALMEIDA, M. B., BRAGA, A.P. **Training SVMs with EDR Algorithm**. In International Journal of Neural Systems, vol. 11, n. 3, 2001, pp. 257- 263.

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional**, Tese de Doutorado, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, 2007.

BAEZA-YATES, R. BERTIER, R. N. **Modern Information Retrieval**. Harlow: Addison-Wesley. 1999.

BATISTA, G.E.A.P.A. PRATI, R.C. MONARD M.C. **A Study of the Behavior of Several Methods for Balancing Machine Learning Training data**. SIGKDD Explorations, v.6. 2004. p. 20-29.

BOUCHET-VALAT M. **SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library**. R package version 0.5.1. 2015.

CHANG C., LIN C. **LIBSVM : a library for support vector machines.** ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

CHAVES, A. C. F. **Extração de regras fuzzy para máquinas de vetor suporte (SVM) para classificação em múltiplas classes**. 2006. 54 f. Tese (Doutorado em Ciências) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Rio de Janeiro.

CHAWLA N. V., BOWYER K.W., HALL L.O., KEGELMEYER, W.P. **SMOTE: Synthetic Minority Over-Sampling Techinique**. Artificial Intelligence Research, vol. 16, 2002. pp. 321-357.

CONOVER, W. J., Practical Nonparametric Statistics, 2 Ed, New York, John Wiley, 1980.

CORTES C., VAPNIK V. **Support Vector Networks**. Machine Learning 20:273-297. 1995.

POZZOLO, D., CAELEN, A., WATERSCHOOT, O., BONTEMPI, S., **Racing for unbalanced methods selection**. Disponível em: <http://www.ulb.ac.be/di/map/adalpozz/pdf/presentation.pdf 2013>. Acesso em 28/08/2016.

DIAS, M. A. L., MALHEIROS M. G. **Estudo de Técnicas de Radicalização para a Língua Portuguesa**. Centro Universitário UNIVATES, Lajeado, Rio Grande do Sul, Brasil, 2004.

FEINERER I., HORNIK K. **Tm: Text Mining Package**. R package version 0.6-2. 2015.

FEINERER I., HORNIK K., MEYER D. **Text Mining Infrastructure in R**. Journal of Statistical Software 25(5): 1-54. 2008.

FLORES, F. N. **Avaliando o Impacto da Qualidade de um Algoritmo de Stemming na Recuperação de Informações**. Trabalho de graduação. Universidade Federal do Rio Grande do Sul, 2009.

GOMES, N. O. **Categorização de textos - estudo de caso: documentos de pedidos de patente no idioma português**. Tese de Doutorado, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, 2013.

GUILHERME I.R., DE SOUZA SERAPIÃO A.B., RABELO C., MENDES J.R.P. (2006) **An Ontology Based for Drilling Report Classification.** MICAI 2006: Advances in Artificial Intelligence. MICAI 2006. Lecture Notes in Computer Science, vol 4293. Springer, Berlin, Heidelberg. 2006.

GUNN, S. **Support Vector Machines for Classification and Regression**. ISIS Technical Report, 1998.

HE H., GARCIA E. A. **Learning from Imbalanced Data**. IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, 2009. pp. 1263-1284.

HSIEH L. **Rig NPT: the ugly truth.** Disponível em: http://www.drillingcontractor.org/rig-npt-the-ugly-truth-6795. 2010. Acessado em: 13/01/2017.

HSU C., CHANG C., LIN C. **A Practical Guide to Support Vector Classification**. Department of Computer Science, National Taiwan University, Taipei, Taiwan. 2010.

HSU, C., LIN, C. **A Comparison on Methods for Multi-class Support Vector Machines**. In IEEE Transactions on Neural Networks, vol. 13 (2), 2012, pp. 415-425.

IHS, (Information Handling Services Petrodata) **Offshore Rig Day Rate Trends**. Disponível em: https://www.ihs.com/products/oil-gas-drilling-rigs-offshore-day-rates.html. 2016. Acessado em: 01/02/2017.

IMTECH **ASVM based server for rice website**. Disponível em: http://www.imtech.res.in/raghava/rbpred/algorithm.html. 2012. Acessado em: 20/01/2017.

JOACUIMS, T. **Text Categorization with Support Vector Machines: Learning with Many Relevant Features.** Machine Learning: ECML-98. Volume 1398 of the series Lecture Notes in Computer Science. 2005. pp 137-142.

KUHN M. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. **Caret: Classification and Regression Training**. R package version 6.0-71. 2016.

KUMAR A. **Business Intelligence = Structured Data Vs Unstructured**. Disponível em:< http://amitbizintel.blogspot.com.br/2015/02/structured-data-vs-unstructured-data.html>. 2017. Acesso em: 12 de jan. 2017.

LANDMARK. **Open Wells® Operations Reporting – Software para registro de operações em construções de poços.** Disponível em:< https://www.landmark.solutions/OpenWells>. 2017. Acesso em: 12 de jan. 2017.

LAURIKKALA, J. **Improving Identification of Difficult Small Classes by Balancing Class Distribuition.** Proc. Conf. AI in Medicine in Europe: Artificial Intelligence Medicine, pp. 63-66, 2001.

MAKREHCHI M., KAMEL M. S. **Automatic Extraction of Domain-Specific Stopwords from Labeled Documents.** C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 222–233, 2008.

MEYER, D., DIMITRIADOU E., HORNIK K., WEINGESSEL A., LEISCH F. **e1071: Misc Functions of the Department of Statistics, Probability Theory Group** (Formerly: E1071), TU Wien. R package version 1.6-7. 2015.

MINER, G., Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 1. Ed. Oxford: Elsevier, 2012. 1000p.

MIURA K. **Um método para aquisição e representação de conhecimento sobre procedimentos operacionais em serviço de completação de poços marítimos.** Dissertação de mestrado, Departamento de Engenharia de Petróleo, Universidade Estadual de Campinas, Campinas, São Paulo. 1992.

MIURA K., GUILHERME I. R., MOROOKA C. K., MENDES, J. R. P. **Processing Technical Daily Reports in Offshore Petroleum Engineering - An Experience**. JACIII Vol.7 No.2 pp. 223-228. 2003.

ORENGO, V. M., HUYCK, C. A **Stemming Algorithm for the Portuguese Language**. In Proceedings of the SPIRE Conference, Laguna de San Raphael, Chile, November 13-15, 2001.

PAICE, C. D. **Another Stemmer**. Department of Computing. Lancaster University, Reino Unido, 1983.

PETROBRAS S.A. **Sondópolis.** Imagem adaptada de aplicativo em intranet. DP&T-POÇOS/SM. Macaé, Rio de Janeiro, 2017.

POZZOLO A. D., CAELEN O., BONTEMPI G. **Unbalanced: Racing for Unbalanced Methods Selection**. R package version 2.0. 2015.

PRATI, R.C., BATISTA, G.E.A.P.A. & SILVA, D.F. **Class imbalance revisited: a new experimental setup to assess the performance of treatment methods.** Knowledge and Information Systems 2015, Volume 45, Issue 1, pp 247–270. 2015.

R CORE TEAM. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2013

RASCHKA S. **Machine Learning FAQ,** 2012. Disponível em <https://sebastianraschka.com/faq/docs/evaluate-a-model.html>. Acesso em 15/01/2017.

SEWELL, M. **Structural Risk Minimization**. 2008. Disponível em: <http://www.svms.org/srm/srm.pdf>. Acessado em 15/01/2017.

SOARES, F. A. Categorização Automática de Textos Baseada em Mineração de Textos, Tese de Doutorado, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, 2013.

SOARES H., MOURA, R. **A methodology to guide writing Software Requirements Specification document**. Conference: XLI Conferencia Latino Americana de Informática - CLEI 2015, At Arequipa, Peru., Volume: 1. 2015.

SOKOLOVA M., LAPALME, G. **A systematic analysis of performance measures for classification tasks**. Information Processing and Management 45, pp. 427-437. Elsevier. 2009.

STEHMAN, S. V. **Selecting and interpreting measures of thematic classification accuracy**. Remote Sensing of Environment. 62 (1): 77–89. 1997.

TOMEK, I. **Two Modifications of CNN**. IEEE Trans. System, Man, Cybernetics, vol. 6, no. 11, pp. 769-772. 1976.

VAN RIJSBERGEN, C. J. **Information Retrieval**. 2nd edition edn. Butterworth. 1979.

VAPNIK V. N. **An Overview of Statistical Learning Theory**. In IEEE Trans. On Neural Networks, vol.10(5), pp. 988-999, 1999.

VAPNIK V., GOLOWICH S., SMOLA A. **Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing**. In: M. Mozer, M. Jordan, and T. Petsche (eds.): Neural Information Processing Systems, Vol. 9. MIT Press, Cambridge, MA, 1997.

WEISS, G. **Mining with Rarity: A Unifying Framework**. ACM SIGKDD Explorations v.6, 2004. p.7-19

XAVIER B. M., SILVA A. D., GOMES G. R. R. **Analise Comparativa de Algoritmos de Redução de Radicais e sua Importância para a Mineração de Texto**. Revista Eletrônica Pesquisa Operacional para o Desenvolvimento. Rio de Janeiro, v.5, n.1, p. 84-99, 2013.

ZHANG W., TANG X., YOSHIDA T. **Text Classification with Support Vector Machine and Back Propagation Neural Network.** In: Shi Y., van Albada G.D., Dongarra J., Sloot P.M.A. (eds) Computational Science – ICCS 2007. ICCS 2007. Lecture Notes in Computer Science, vol 4490. Springer, Berlin, Heidelberg

ZHANG, J. MANI, I. KNN **Approach to Unbalanced Data Distribuitions: A Case Study Involving Information Extraction**. Proc. Int’l Conf. Machine Learning (ICML 2003), Workshop Learning from Imbalanced Data Sets, 2003.