**Instituto de Pesquisas Tecnológicas do Estado de São Paulo**

**Mateus Espadoto**

# Método para Análise de Risco de Modelos de Classificação de Documentos Digitalizados

**São Paulo**

**2016**

Mateus Espadoto

# Método para Análise de Risco de Modelos de Classificação de Documentos Digitalizados

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Computação.

Área de Concentração: Engenharia de Software

Orientador: Fernando A. de Castro Giorno

São Paulo
Novembro/2016

Ficha Catalográfica
Elaborada pelo Departamento de Acervo e Informação Tecnológica – DAIT
do Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT

E77m **Espadoto, Mateus**
Método para análise de risco de modelos de classificação de documentos digitalizados. / Mateus Espadoto. São Paulo, 2016.
55p.

Dissertação (Mestrado em Engenharia de Computação) - Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Área de concentração: Engenharia de Software.

Orientador: Prof. Dr. Fernando A. de Castro Giorno

1. Classificação de documentos digitalizados 2. Análise de risco 3. Gestão de documentos 4. Tese I. Giorno, Fernando A. de Castro, orient. II. IPT. Coordenadoria de Ensino Tecnológico III. Título

16-60 CDU 004.91(043)

Mateus Espadoto

# Método para Análise de Risco de Modelos de Classificação de Documentos Digitalizados

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Computação.

Data da aprovação ____ / ____ / _____

Prof. Dr. Fernando A. de Castro Giorno (Orientador)
Mestrado Engenharia de Computação

Membros da Banca Examinadora:

Prof. Dr. Fernando A. de Castro Giorno (Orientador)
Mestrado Engenharia de Computação

Prof. Dr. Victor Fossaluza (Membro)
USP - Universidade de São Paulo

Prof. Dr. Carlos Eduardo de Barros Paes (Membro)
PUC-SP - Pontifícia Universidade Católica de São Paulo

*Para Adriana, que me ensinou a viver.*

## AGRADECIMENTOS

Em primeiro lugar, agradeço ao meu orientador, Prof. Fernando Giorno, pelo apoio incondicional e pela tranquilidade passada durante a elaboração do trabalho. Sem ele este trabalho não teria sido possível.

Agradeço também ao professor Marcelo Rezende pela conversa que deu origem à ideia central deste trabalho, e aos professores José Eduardo Deboni e Mario Miyake pelas sugestões valiosas.

Por último mas não menos importante, agradeço à equipe do IPT pela disposição e paciência para esclarecer minhas inúmeras dúvidas durante o curso.

*“We are here to laugh at the odds and live our lives so well that Death will tremble to take us.”*

*(Charles Bukowski)*

## RESUMO

As empresas brasileiras realizam muitas transações por meio de documentos impressos, o que ocorre principalmente devido a obrigações legais. O processo de gestão de documentos impressos requer, além de amplo espaço físico, grande esforço manual para classificação e organização, o que o torna sujeito a erros que podem gerar riscos para as empresas. Este trabalho apresenta um método para análise de risco de modelos de classificação de dados aplicados a documentos de negócio. Para tanto, o método proposto considera aspectos como disponibilidade, precisão e agilidade, e permite comparar o desempenho de diferentes modelos de classificação de acordo com os tipos de documentos a serem classificados, técnicas de pré-processamento aplicadas e quantidade de dados de treinamento disponíveis.



**Palavras-chaves**: classificação de documentos. análise de risco. aprendizado de máquina.

## ABSTRACT

### Risk Analysis Method for Classification Models of Scanned Documents

Brazilian companies do many transactions using printed documents, mainly due to legal obligations. Paper document management is a process that demands great effort for classification and organisation in terms of personnel, and requires large physical spaces for storage, all of which makes it vulnerable to errors that may translate into risk for companies. This work presents a method for risk analysis of data classification models applied to business documents. The method considers aspects such as availability, accuracy and agility, and allows the comparison among different classification models, according to the types of documents to be classified, pre-processing techniques being used and the amount of training data available.



**Keywords**: document classification. risk analysis. machine learning.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ANN | *Artificial Neural Network*, ou Rede Neural Artificial |
| CSV | *Comma-Separated Values*, ou Valores Separados por Vírgula |
| EQM | Erro Quadrático Médio |
| FAIR | *Factor Analysis for Information Risk* |
| FGTS | Fundo de Garantia por Tempo de Serviço |
| FN | *False Negatives*, ou Falsos Negativos |
| FP | *False Positives*, ou Falsos Positivos |
| GED | Gerenciamento Eletrônico de Documentos |
| GPS | Guia da Previdência Social |
| IBGE | Instituto Brasileiro de Geografia e Estatística |
| NLP | *Natural Language Processing*, ou Processamento de Linguagem Natural |
| PCA | *Principal Component Analysis*, ou Análise de Componentes Principais |
| PCMSO | Programa de Controle Médico de Saúde Ocupacional |
| PMC | *Perceptron* de Múltiplas Camadas |
| RGB | *Red, Green, Blue*, ou Vermelho, Verde e Azul, as cores utilizadas para representação digital de imagens |
| SVM | *Support Vector Machine*, ou Máquina de Vetor Suporte |
| TN | *True Negatives*, ou Negativos Verdadeiros |
| TP | *True Positives*, ou Positivos Verdadeiros |

**SUMÁRIO**

# 1 INTRODUÇÃO

## 1.1 Motivação

Mesmo com o advento da era digital, muitas atividades comerciais ainda são realizadas com o uso de documentos impressos, como faturas, boletos, contratos, entre outros. Estes documentos tipicamente possuem grande variedade de forma e conteúdo, o que dificulta a criação de *software* para o seu processamento automatizado. Como consequência, este trabalho comumente é feito de forma manual e por grande quantidade de pessoas, em um processo suscetível a erros e que representa grande risco para as empresas, que pode se materializar na forma de prejuízo financeiro ou de imagem perante seus clientes.

De acordo com Lyman e Varian (2003), empresas ainda fazem uso intensivo de papel para a condução dos seus negócios, apesar do surgimento do conceito de escritório sem papel (*paperless office*) por volta dos anos 1970 (Business Week, 1975), que prometia redução de custos e aumento de eficiência em relação ao sistema tradicional. A seguir são apresentados dois fatores que contribuem para a perpetuação do uso de papel nas empresas.

a) Exigências legais: o artigo 1.194 do Código Civil Brasileiro (BRASIL, 2002) diz que: “O empresário e a sociedade empresária são obrigados a conservar em boa guarda toda a escrituração, correspondência e mais papéis concernentes à sua atividade, enquanto não ocorrer prescrição ou decadência no tocante aos atos neles consignados.”

   Com base nessa determinação, as diversas áreas do governo determinam prazos legais para a guarda de documentos de acordo com os assuntos de interesse da administração pública. Na tabela 1 é possível observar exemplos de prazos de guarda de documentos:

Tabela 1 – Prazos para guarda de documentos.

| **Tipo de Documento** | **Tempo de Guarda** |
|---|---|
| GPS - Guia da Previdência Social | 5 anos |
| PCMSO - Programa de Controle Médico de Saúde Ocupacional | 20 anos |
| FGTS - Depósitos e documentos relacionados | 30 anos |

Fonte – Elaborado pelo autor com base em (FENACON, 2010)

   A adoção das tecnologias de certificação digital possibilitará que, em alguns anos, grande parte dos documentos sejam gerados e mantidos apenas em

formato digital. Até que isso ocorra, o legado de documentos impressos que deverão ser geridos pelos próximos anos só crescerá.

b) Oferta de mão de obra barata: o trabalho de armazenamento e classificação de documentos é bastante laborioso e, portanto, requer grande quantidade de pessoas para executá-lo. No entanto, por se tratar de trabalho simples e repetitivo e que não requer qualificação específica, é realizado tipicamente por pessoas de baixa escolaridade e que se dispõem a trabalhar em troca de baixos salários. A abundância de mão de obra barata nas grandes cidades pode ser um fator que desestimula a busca por alternativas para a automação desta atividade. Na tabela 2 são apresentados alguns exemplos de salários pagos em funções relacionadas ao serviço de manipulação de documentos.

Tabela 2 – Exemplo de salários pagos para funções relacionadas à manipulação de documentos.

| **Função** | **Carga Horária (horas/mês)** | **Salário Médio (R$)** |
|---|---|---|
| Digitador | 180 | 1.062,20 |
| Auxiliar de Expedição | 220 | 1.432,10 |
| Conferente de Expedição | 220 | 1.505,20 |
| Arquivista | 220 | 1.641,40 |

Fonte – Elaborado pelo autor com base em (DATAFOLHA, 2015)

Há produtos no mercado para processamento de documentos, que comumente são classificados como sistemas de Gerenciamento Eletrônico de Documentos (GED), e que possibilitam a execução de tarefas como digitalização, ingestão, classificação e indexação manuais de documentos. No entanto, por serem operados por humanos, requerem grande esforço para a classificação e indexação correta de documentos, e consequentemente são sujeitos a erros.

O problema de classificação automatizada de documentos é estudado pelo menos desde os anos 1960 (BORKO; BERNICK, 1963), quando já se pensava em soluções; mas a limitação de capacidade dos computadores da época não permitiu implementações viáveis.

Com a capacidade dos computadores atuais, técnicas de aprendizado de máquina, também conhecidas como técnicas de modelagem preditiva, se tornaram viáveis para uso em larga escala, o que possibilitou grande volume de pesquisas sobre o assunto, que exploram o uso de modelos como Redes Neurais Artificiais, Máquinas de Vetores Suporte (*Support Vector Machines*, ou SVM) e técnicas de Processamento de Linguagem Natural (*Natural Language Processing*, ou NLP), entre outras, para classificação automatizada de documentos.

Os trabalhos existentes, de modo geral, podem ser agrupados de acordo com a

sua escolha de abordagem para o problema:

a) Centrados em análise de imagem (MARINAI; GORI; SODA, 2005; GACEB; EGLIN; LEBOURGEOIS, 2011), que buscam classificar documentos com base em similaridade de *forma*;

b) Centrados em análise de texto (MANEVITZ; YOUSEF, 2002; BLEI; NG; JORDAN, 2003), que buscam classificar documentos com base em similaridade de *conteúdo*.

Estes trabalhos fornecem fundamentação teórica para os tipos de análise propostos. Adicionalmente, trabalhos como o de Chen e Blostein (2007) buscam definir parâmetros de avaliação de desempenho de diferentes técnicas.

Apesar de haver muita pesquisa sobre modelos de classificação, não foram encontrados trabalhos que estudem os riscos associados à classificação automatizada de documentos, bem como formas de mitigá-los. Alguns possíveis motivos para a inexistência de pesquisa sobre o assunto podem ser mencionados, como a já citada abundância de mão de obra barata que desestimula a pesquisa, o desconhecimento por parte das empresas de que tal automação é possível, dado que boa parte da pesquisa sobre o assunto fica restrita à academia, e a aversão a riscos de modo geral pelas empresas, em particular aos que são difíceis de serem estimados.

## 1.2 Objetivos

Tem-se como objetivos deste trabalho:

a) Propor um método para avaliação de risco de modelos de classificação de documentos digitalizados encontrados na literatura;

b) Comparar, segundo o método de avaliação de risco a ser proposto, o desempenho de bibliotecas de *software* existentes de Redes Neurais e Máquinas de Vetor Suporte para classificação de documentos, por meio da criação de um protótipo para cálculo de risco e testes com dados reais, ou seja, não fabricados especificamente para os testes;

c) Identificar, dentre as técnicas de classificação de documentos escolhidas, as que possuem o menor risco associado em termos de disponibilidade, precisão e agilidade, critérios propostos de acordo com o *framework* de gestão de risco escolhido.

O método a ser proposto deve estar de acordo com princípios de gestão de risco estabelecidos na literatura, como os descritos nos *frameworks* 4A (WESTERMAN; HUNTER, 2009) e FAIR (JONES, 2006), que serão discutidos posteriormente.

## 1.3 Justificativa

O método de avaliação de risco pode ser justificado por possibilitar que modelos de classificação sejam avaliados em um contexto que contempla, além da precisão de classificação, variáveis como tempo necessário para treinamento e classificação e demanda por recursos computacionais, e por permitir que pesos sejam atribuídos a essas variáveis de acordo com a necessidade ou preferência do seu usuário; como resultado, espera-se que possa ser escolhido o modelo de classificação mais apropriado para cada situação.

## 1.4 Contribuição

Tem-se como contribuição pretendida por este trabalho a popularização do uso de técnicas de classificação automatizadas nas empresas, ao tornar seu risco quantificável a partir do método proposto, que consiste em avaliar as diferentes técnicas sob os pontos de vista de agilidade, precisão e disponibilidade.

## 1.5 Método de Trabalho

Nas seções a seguir são descritas as atividades a serem desenvolvidas ao longo deste trabalho, visando atingir os objetivos propostos.

- Pesquisa bibliográfica: identificação do estado da arte em termos de técnicas de tratamento de imagem para processamento de documentos, redução de informação e de classificação de documentos, identificação de trabalhos relacionados, revisão crítica de artigos selecionados e posicionamento quanto à relevância dos mesmos para este trabalho.
- Método para avaliação de risco: elaboração do método para avaliação de risco baseado no *framework* 4A, que possibilite a comparação de diferentes modelos de classificação, avaliados quanto aos critérios de precisão, disponibilidade e agilidade, descritos a seguir:
    a) Precisão: risco de documentos serem classificados incorretamente;
    b) Disponibilidade: risco do treinamento de um modelo de classificação requerer mais memória do que a média observada ou do que um valor pré-definido;
    c) Agilidade: risco do treinamento de um modelo de classificação requerer mais tempo do que a média observada ou do que um valor pré-definido.

  O detalhamento do método para avaliação de risco pode ser encontrado na seção 3.4.

- Especificação funcional: levantamento de requisitos e criação de especificação funcional para desenvolvimento do protótipo para cálculo de risco. O protótipo tem como objetivo executar as tarefas a seguir:

    a) Pré-processamento de imagens;
    b) Redução de informação;
    c) Treinamento dos modelos de classificação;
    d) Classificação de documentos;
    e) Armazenamento de dados que possibilitem a avaliação dos resultados da execução;
    f) Cálculo de risco de modelos de classificação, com base no método a ser proposto neste trabalho.

- Obtenção e preparação dos dados de teste: definição de categorias de documentos a serem utilizadas durante os testes, obtenção de conjunto de dados com documentos representativos das categorias selecionadas, como por exemplo, boletos bancários, faturas e notas fiscais, e classificação manual dos documentos obtidos para que possam ser utilizados como dados de treinamento e teste pelos modelos de classificação.

- Desenvolvimento do protótipo para cálculo de risco de acordo com a especificação funcional: foi escolhida a linguagem de programação R (R Core Team, 2015) para o desenvolvimento, com exceção do programa de pré-processamento de imagens, a ser desenvolvido em C++ por questões de desempenho e devido à disponibilidade da biblioteca *Magick++*, que possui as funcionalidades de correção de distorção e limiarização (ImageMagick Studio LLC, 2016). A linguagem R foi escolhida devido à disponibilidade de bibliotecas de aprendizado de máquina como *caret* (WING et al., 2016), *nnet* (VENABLES; RIPLEY, 2002), *e1071* (MEYER et al., 2015) e *kernlab* (KARATZOGLOU et al., 2004), bem como de pacotes para visualização e análise estatística de dados, como *ggplot2* (WICKHAM, 2009).

- Plano de testes: definição e execução de plano de testes que permita observar a influência do tamanho do conjunto de dados de treinamento sobre os aspectos de disponibilidade, precisão e agilidade para cada modelo.

- Análise dos resultados: consolidação de dados obtidos durante a execução dos testes, análise dos resultados alcançados e conclusões.

## 1.6 Organização do Texto

Este texto está organizado nas seguintes seções, em adição à esta seção introdutória:

a) Revisão Bibliográfica: apresenta artigos recentes sobre os temas classificação e pré-processamento de documentos, bem como apresenta outras referências bibliográficas selecionadas como parte da fundamentação teórica deste trabalho.

b) Método para Avaliação de Risco: apresenta o método de avaliação de risco, define os critérios de avaliação de risco a serem utilizados no restante do trabalho com base no *framework* 4A, apresenta a especificação do protótipo para cálculo de risco a ser utilizado para os testes e o plano de testes proposto.

c) Análise dos Resultados: apresenta a análise dos resultados obtidos com a execução do plano de testes.

d) Conclusões: apresenta as conclusões obtidas com base nos resultados, destacando as contribuições e fornecendo sugestões para a continuidade deste trabalho.

# 2 REVISÃO BIBLIOGRÁFICA

## 2.1 Introdução

Nesta seção serão apresentados, de forma crítica, trabalhos representativos da pesquisa sobre diferentes técnicas de classificação de documentos e métodos de avaliação de desempenho destas técnicas, bem como será estabelecida uma ligação entre estes e um método a ser proposto que possibilite a sua comparação do ponto de vista de análise de risco.

No universo dos modelos de aprendizado de máquina, há inúmeras técnicas de classificação de dados baseadas em padrões. Apesar do propósito similar, nem todas as técnicas são apropriadas para todos os tipos de dados de entrada. Para este trabalho, foram selecionadas duas técnicas que aparecem com frequência em trabalhos de pesquisa, como em Marinai, Gori e Soda (2005) e em Manevitz e Malik (2002) quando o assunto é classificação de documentos ou imagens, a saber:

a) Redes Neurais Artificiais (ANN, ou *Artificial Neural Networks*);

b) Máquinas de Vetor Suporte (SVM, ou *Support Vector Machines*);

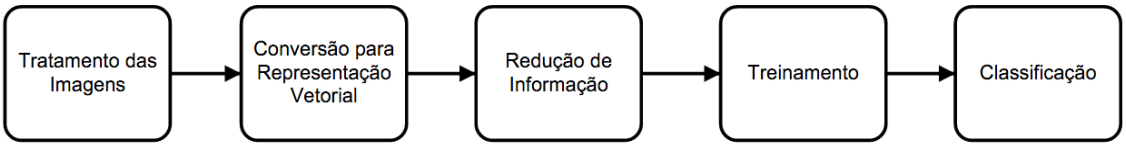
Redes Neurais Artificiais e Máquinas de Vetor Suporte são dois modelos computacionais de aprendizado bastante difundidos e estudados, com usos nas mais diversas áreas, como engenharia, agricultura e economia, entre outras.

Em um contexto de aprendizado supervisionado, o funcionamento externo dos dois modelos é similar: há uma fase de treinamento, em que são submetidas amostras de dados classificadas manualmente, que serão utilizadas como exemplos pelo modelo para a identificação de padrões existentes e geração de uma função de classificação. Uma vez que o modelo esteja treinado, novos dados são submetidos para esta função para serem classificados.

No caso de classificação de imagens, que podem variar em termos de resolução e qualidade, é necessário aplicar algumas técnicas para a melhoria da qualidade da imagem, como correção de distorção (*deskewing*) e limiarização (*thresholding*), e redução de informação, como Análise de Componentes Principais (PCA, ou *Principal Component Analysis*), para extrair as características da imagem que são significantes para o modelo de classificação.

Em todos os trabalhos analisados, observa-se uma sequência de passos comum, conforme mostrado na figura 1.

Figura 1 – Processo típico de classificação de imagens.




Fonte – Elaborado pelo autor

As fases mostradas na figura 1 podem ser descritas conforme abaixo:

a) Tratamento da imagem (correção de distorções e limiarização, delimitação da área de interesse a ser processada, entre outros);
b) Conversão da imagem em uma representação vetorial de uma dimensão;
c) Redução de informação na área escolhida;
d) Treinamento do modelo de classificação de dados, com imagens pré-classificadas manualmente;
e) Classificação de dados desconhecidos, a partir do modelo de classificação treinado.

Para este trabalho, será adicionado um novo passo de conversão da imagem para uma representação em tons de cinza (*grayscale*), como forma de eliminar qualquer informação de cor, que tipicamente não é presente em documentos de negócio mas que pode ser introduzida incorretamente durante a digitalização. Com a eliminação de informação de cor, a imagem será reduzida para aproximadamente um terço do tamanho, devido à conversão da representação RGB que utiliza 24 bits por *pixel* para a representação em tons de cinza, que requer apenas 8 bits por *pixel*. Esta redução de tamanho simplifica o posterior processamento da imagem e não compromete a forma dos elementos visuais do documento que será utilizada pelos modelos de classificação.

Apesar das similaridades externas, o funcionamento interno das Redes Neurais e das Máquinas de Vetor Suporte é significativamente diferente. Nas seções a seguir serão apresentadas estas técnicas em detalhe, com os respectivos exemplos de situações onde seu uso é indicado, de acordo com os trabalhos selecionados.

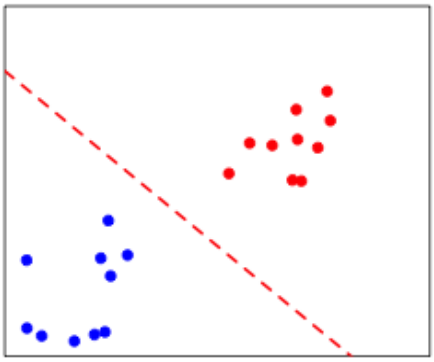
## 2.2 Redes Neurais Artificiais

Rede Neural Artificial é um modelo computacional inspirado no funcionamento do sistema nervoso de seres vivos, que possui a capacidade de aquisição e retenção do conhecimento (SILVA; SPATTI; FLAUZINO, 2010). Pode ser definida pela junção de um conjunto de unidades de processamento e matrizes de interconexões, que

funcionam de forma inspirada nos neurônios e sinapses de um cérebro biológico, respectivamente. Há diversos tipos de Redes Neurais Artificiais que compartilham os mesmos princípios mas que possuem objetivos distintos como classificação, regressão e agrupamento, e cujo treinamento pode ser supervisionado, como nos casos de classificação e regressão, ou não-supervisionado, como em casos de agrupamento. Neste trabalho será utilizado o *Perceptron* de Múltiplas Camadas (PMC), que possui bom desempenho para classificação de documentos, de acordo com Marinai, Gori e Soda (2005).

As Redes Neurais Artificiais possuem um longo histórico, com diversos altos e baixos: da sua criação nos anos 1940 à evolução até os modelos *Perceptron* e *Adaline* no final dos anos 1950, que possibilitam a classificação binária de dados cuja separação seja linear, como exemplificado na figura 2.

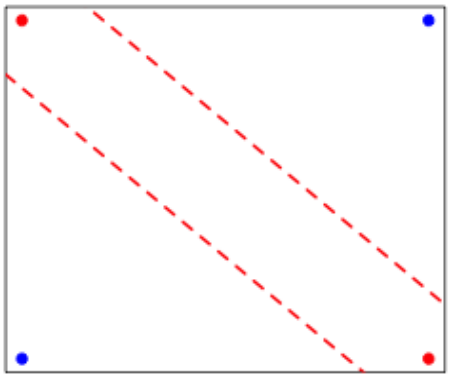
Figura 2 – Exemplo de dados linearmente separáveis.



Fonte – Elaborado pelo autor

Nos anos 1960, Minsky e Papert (1969) provaram que *Perceptrons* de uma única camada não eram capazes de classificar corretamente dados cujas classes fossem separáveis apenas não-linearmente, o que foi um grande revés para o desenvolvimento das Redes Neurais Artificiais na época, pois provocou uma diminuição do interesse pelo assunto até o final dos anos 1980, quando surgiu o modelo *Perceptron* de Múltiplas Camadas somado à técnica de *Backpropagation*. Este novo modelo solucionou o problema da separabilidade de classes não-lineares, o que despertou novo interesse no assunto, que se prolonga até os dias atuais. Na figura 3 é possível ver um exemplo de dados cuja separação é não-linear.

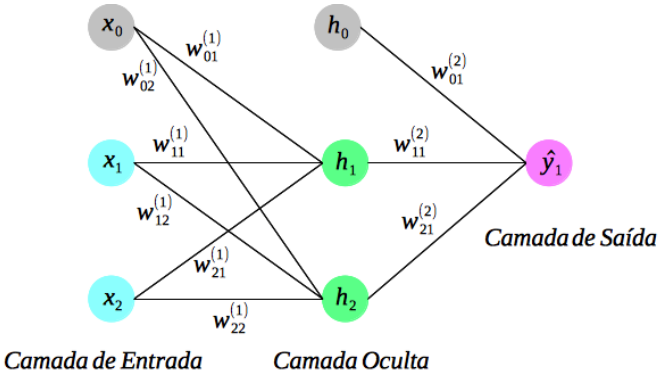
Figura 3 – Exemplo de dados não-linearmente separáveis.



Fonte – Elaborado pelo autor

Na figura 4 é possível ver o exemplo de um PMC criado para resolver o problema de aprendizado do operador lógico *XOR* (“ou-exclusivo”), que é não-linearmente separável, análogo ao exibido na figura 3. Este exemplo ilustra a arquitetura de um PMC mínimo e é útil para apoiar a fundamentação teórica do assunto.

Figura 4 – Rede neural XOR.




Fonte – Elaborado pelo autor

Este exemplo aceita como entrada tuplas do tipo $(x_1, x_2, y)$, em que $x_1$ e $x_2$ são os operandos de entrada para o operador *XOR*, e $y$ é o valor esperado de resposta, de acordo com a tabela 3.

Formalmente, um PMC possui ao menos três camadas, sendo uma de entrada, uma de saída, e uma ou mais camadas ocultas, intermediárias. A camada de entrada é um vetor de variáveis que representam os dados a serem classificados, indicados no exemplo pelos elementos $x_i$. A camada de saída é a representação vetorial das variáveis de resposta, ou seja, o valor calculado pelo PMC que indica a qual classe

Tabela 3 – Tabela-verdade do operador lógico *XOR*.

| $x_1$ | $x_2$ | $\hat{y}_1$ |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 0 |

Fonte – Elaborado pelo autor

determinado dado pertence. É indicada no exemplo por $\hat{y}_i$. As camadas ocultas têm a função de introduzir não-linearidade no modelo, possibilitando a classificação de dados com separação não-linear. No exemplo, tem-se somente uma camada oculta, indicada pelos elementos $h_i$. Autores como Heaton (2008) defendem que para a maior parte dos usos, apenas uma camada oculta é suficiente, e que camadas ocultas adicionais podem aumentar consideravelmente o tempo necessário para treinamento do PMC.

Todos os elementos, com exceção dos elementos da camada de entrada e dos elementos de viés, a serem definidos a seguir, estão associados a uma função de ativação, que juntamente com as camadas ocultas, são utilizadas para introduzir não-linearidade no modelo, e é o que possibilita o aprendizado com conjuntos de dados cuja separação é não-linear. Duas funções são comumente utilizadas para ativação, a função logística e a função tangente hiperbólica. Estas funções são utilizadas porque possuem imagem em um intervalo de valores conhecido (0 a 1 para logística, -1 a 1 para tangente hiperbólica), e porque são totalmente diferenciáveis, ou seja, suas derivadas de primeira ordem existem e são conhecidas em todos os pontos de seu domínio. Estas duas propriedades são importantes para manter consistência dos valores na rede neural.

A função logística é representada pela equação abaixo:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

A função tangente hiperbólica é representada pela equação abaixo:

$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2.2}$$

Todas as camadas com exceção da de saída possuem elementos de viés (*bias*), indicados na figura pelos elementos de índice 0 ($x_0$ e $h_0$). Estes elementos sempre possuem valor 1, e como não são calculados não possuem conexões de entrada. São importantes para permitir que a saída da função de ativação possa ser deslocada no eixo x. A relação do elemento de viés com a função de ativação pode ser considerada

similar à relação do intercepto $b$ em uma função linear $y = ax + b$, que permite que a reta inteira se desloque verticalmente, sem mudar a sua inclinação.

O processo de treinamento do PMC possui duas fases, chamadas de Propagação Adiante (*Feedforward*) e Propagação Reversa (*Backward*), que são realizados de forma iterativa, sendo que em cada iteração ou época, o conjunto de amostras de treinamento é avaliado em sua totalidade. Na fase de Propagação Adiante, os valores dos elementos das camadas à frente da camada de entrada são calculados de acordo com a equação a seguir:

$$o_j = f_A(\sum_{i=0}^{n} x_i w_{ij}^{(k)}) \tag{2.3}$$

Em que $o_j$ é o j-ésimo elemento da camada seguinte de $m$ elementos, com $j = 1, ..., m$; $x_i$ o i-ésimo elemento da camada atual de $n$ elementos, com $i = 0, ..., n$; $f_A(x)$ é a função de ativação do j-ésimo elemento $o_j$; e $w_{ij}^{(k)}$ é o elemento da matriz de pesos $W^{(k)}$ indexado por $i$ e $j$, com $k = 1, ..., p-1, p$ = número de camadas. Explicado de outra forma, o valor de cada elemento da camada seguinte é calculado como sendo a aplicação da função de ativação sobre a soma dos valores da camada atual ponderados pela matriz de pesos existente entre as camadas atual e seguinte.

Uma vez obtidos os valores da camada de saída para todas as tuplas de amostra, deve-se calcular o erro do PMC para a amostra por meio da função de Erro Quadrático Médio (EQM), definida pela equação a seguir:

$$E = f_{EQM} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.4}$$

Em que $E$ é o erro calculado para a última camada $p$, $n$ é a quantidade de amostras, $y_i$ é o valor ideal, fornecido na i-ésima tupla de entrada, e $\hat{y}_i$ é o valor estimado para $y_i$, calculado pelo PMC.

Uma vez calculado o EQM em uma época, é iniciado o processo de Propagação Reversa, que se utiliza do Método do Gradiente para minimizar o erro. A Propagação Reversa tem início ao obter a derivada parcial da função de EQM em função da matriz de pesos imediatamente anterior à camada de saída, no exemplo, $W^{(2)}$, conforme a equação a seguir:

$$\nabla E^{(p)} = \frac{\partial E}{\partial w_{ij}^{(k)}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial i_j}\frac{\partial i_j}{\partial w_{ij}^{(k)}} \tag{2.5}$$

Em que $\nabla E^{(p)}$ é o gradiente do erro $E$ para a camada $p$, ou seja, o erro mínimo, $i_j$ é o valor da soma ponderada dos elementos da camada anterior à camada de saída

antes da aplicação da função de ativação. Uma vez obtida a derivada parcial em função de $w_{ij}^{(k)}$, para todo *ij* da matriz (*k*), os elementos $w_{ij}^{(k)}$ devem ser atualizados conforme a equação a seguir:

$$w_{ij}^{(k)*} = w_{ij}^{(k)} - \eta \frac{\partial E^{(p)}}{\partial w_{ij}^{(k)}} \tag{2.6}$$

Em que $\eta$ é a taxa de aprendizado, ou seja, uma constante que controla o tamanho do passo a ser dado a cada iteração do algoritmo. Valores grandes de $\eta$ fazem com que o treinamento do PMC seja mais rápido, e valores pequenos fazem com que o resultado do treinamento seja mais preciso. Tipicamente são utilizados valores entre 0 e 1.

Uma vez que os pesos da última matriz $W^{(2)}$ foram atualizados com base no EQM, é preciso atualizar os pesos da matriz $W^{(1)}$. No entanto, por se tratar de matriz de conexão com uma camada que não a de saída, não é possível calcular o seu EQM por não haver valores esperados de saída desta camada. Neste caso é preciso propagar as alterações ocorridas na matriz $W^{(2)}$ na hora de calcular $\nabla E^{(p-1)}$, o que se dá pela mesma equação de gradiente mostrada anteriormente, repetida aqui para clareza, mas com uma alteração no primeiro termo:

$$\nabla E^{(p-1)} = \frac{\partial E}{\partial w_{ij}^{(k)}} = \left( \sum_{k=1}^{n} \frac{\partial E}{\partial i_k} w_{kj}^{(2)} \right) \frac{\partial o_j}{\partial i_j} \frac{\partial i_j}{\partial w_{ij}^{(k)}} \tag{2.7}$$

Em que *n* é a quantidade de elementos da camada oculta, $\frac{\partial E}{\partial i_k}$ é a derivada parcial do EQM em função dos valores de saída da camada oculta, e $w_{kj}^{(2)}$ é a matriz de pesos entre as camadas oculta e de saída, já com os valores atualizados pelo gradiente executado anteriormente. Caso existam mais do que uma camada oculta, este passo deve ser repetido para que a propagação de erros ocorra até a primeira matriz de pesos, aquela que existe entre a camada de entrada e a camada oculta. Dessa forma está concluída a fase de Propagação Reversa, e portanto o algoritmo de *Backpropagation*.

Pode-se mencionar como características importantes das Redes Neurais Artificiais:

a) Capacidade de aprendizado: ao receber amostras de dados como entrada, o modelo é capaz de identificar os relacionamentos entre as diversas variáveis que compõem cada amostra e, assim, extrair os padrões existentes para uso futuro;

b) Capacidade de generalização: o modelo é capaz de reconhecer amostras de dados nunca antes vistas como sendo similares a algum dos padrões conhecidos;

c) Imunidade a padrões incorretos: a existência de alguns exemplos classificados incorretamente durante o treinamento não afeta de forma significativa o desempenho do modelo.

Como problemas inerentes às Redes Neurais, pode-se mencionar:

a) Demanda por capacidade computacional: o volume do conjunto de dados de treinamento deve caber na memória do computador;
b) Aprendizado com matriz de pesos aleatórios: uma seleção ruim de pesos pode afetar drasticamente o funcionamento do modelo, causando problemas durante o aprendizado e de desempenho durante a classificação.

Considerando as características mencionadas, Redes Neurais Artificiais são utilizadas largamente como ferramenta para identificação de padrões complexos.

Em Marinai, Gori e Soda (2005), são apresentados os diversos usos possíveis de Redes Neurais Artificiais para atividades de processamento de documentos digitalizados em formato de imagem, como por exemplo, para limpeza de caracteres quebrados ou sobrepostos, detecção de distorções, análise de *layout*, com o objetivo de extração de seções particulares de um documento, segmentação e reconhecimento de caracteres, verificação de assinaturas, e o objeto deste trabalho, que é a classificação de documentos.

As tentativas iniciais de se utilizar Redes Neurais para a classificação de documentos frequentemente dependiam do uso de linhas guia impressas nos documentos, como forma de simplificar a tarefa em termos computacionais. Abordagens mais recentes associam a tarefa de classificação à tarefa de análise de *layout*, que é um processo de estruturação do documento, para posterior extração de informações, comumente utilizando-se de Grafos e Redes Neurais Recursivas (GACEB; EGLIN; LEBOURGEOIS, 2011). No presente trabalho serão abordadas somente técnicas de classificação que não buscam estruturar os documentos, porque não se pretende extrair informações dos documentos, apenas classificá-los.

## 2.3 Máquinas de Vetor Suporte

Máquina de Vetor Suporte (SVM, ou *Support Vector Machine*) é um modelo computacional de aprendizado de máquina, proposto inicialmente por Vapnik (1995). Diferentemente das Redes Neurais, Máquinas de Vetor Suporte não utilizam uma abordagem inspirada na natureza, e sim teorias estatísticas.

Por definição, Máquinas de Vetor Suporte são classificadores binários, ou seja, seu resultado é indicar se um dado pertence ou não a uma determinada classe, diferentemente das Redes Neurais, que podem ser treinadas para classificar quantas

classes forem necessárias. Quando há a necessidade de se identificar múltiplas classes com SVM, são adotadas as seguintes estratégias, assumindo $n$ como o número de classes desejado (ANTHONY; GREGG; TSHILIDZI, 2007):

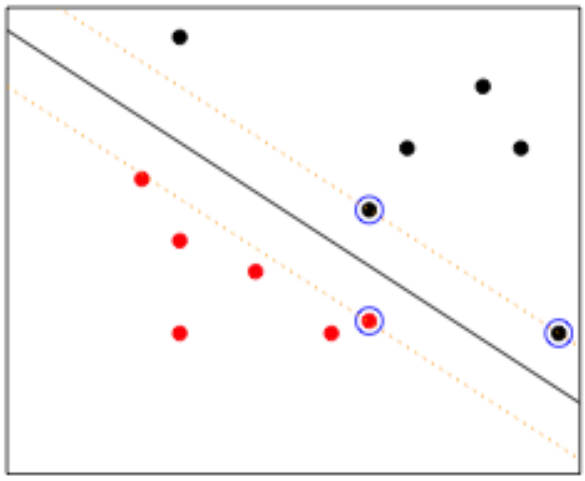
a) Um contra todos (1AA, de *one against all*): é criada uma SVM para cada classe, com exemplos positivos e negativos, totalizando $n$ máquinas;

b) Um contra um (1A1, de *one against one*): são criadas $n(n-1)/2$ máquinas, sendo uma para cada par de classes. A classe considerada vencedora no processo de classificação é aquela que tiver o maior número de votos para um determinado dado.

Estudos mostram que a estratégia 1A1 possui melhor desempenho de classificação quando comparada à estratégia 1AA, mas com um custo computacional maior, devido à necessidade de se treinar uma quantidade exponencialmente maior de máquinas.

Formalmente, o objetivo de uma Máquina de Vetor Suporte, enquanto classificador binário, ou seja, que permite a separação dos dados em duas categorias, é encontrar um hiperplano cuja posição permita separar os dados de treinamento em dois com a maior margem possível, chamado de hiperplano separador ótimo. Em situações onde os dados não são linearmente separáveis, é necessário mapear os dados para um novo espaço, o que é feito por meio de funções *kernel*. Uma vez encontrado um espaço onde os dados sejam separáveis, tem-se um problema de otimização que é resolvido com o uso multiplicadores de Lagrange. Os vetores suporte são os dados que estão mais perto do hiperplano separador ótimo, conforme mostrado na figura 5.

Figura 5 – Exemplo de hiperplano separador.



Fonte – Elaborado pelo autor

Máquinas de Vetor Suporte e Redes Neurais estão sujeitas ao fenômeno conhecido como sobreajuste (*overfitting*), situação que ocorre quando não há amostras suficientes de dados de treinamento para permitir que ocorra generalização no processo de classificação. Neste caso, o modelo só é capaz de classificar corretamente os dados de amostra utilizados no treinamento, e diz-se que o modelo "memorizou" os dados do conjunto de treinamento (LORENA; CARVALHO, 2007). Considera-se que Máquinas de Vetor Suporte são menos vulneráveis ao sobreajuste do que as Redes Neurais, quando utilizadas com dados de grandes dimensões, como imagens.

Pode-se mencionar como características importantes das Máquinas de Vetor Suporte:

a) Capacidade de generalização: o modelo é capaz de reconhecer amostras de dados nunca antes vistas como sendo similares a algum dos padrões conhecidos;

b) Robustez em grandes dimensões: Máquinas de Vetor Suporte são eficientes ao evitar o sobreajuste ao processar dados com grande quantidade de variáveis, como imagens.

c) Determinismo: diferentemente das Redes Neurais, Máquinas de Vetor Suporte não dependem de uma matriz de pesos aleatórios, que podem prejudicar a convergência da função objetivo.

Como problemas inerentes às Máquinas de Vetor Suporte, pode-se mencionar:

a) Demanda por capacidade computacional: o volume do conjunto de dados de treinamento deve caber na memória do computador;

b) Classificação binária: devido à característica de classificar dados de forma binária, são necessárias diversas Máquinas de Vetor Suporte para realizar a classificação de múltiplas classes, o que requer maior tempo de treinamento.

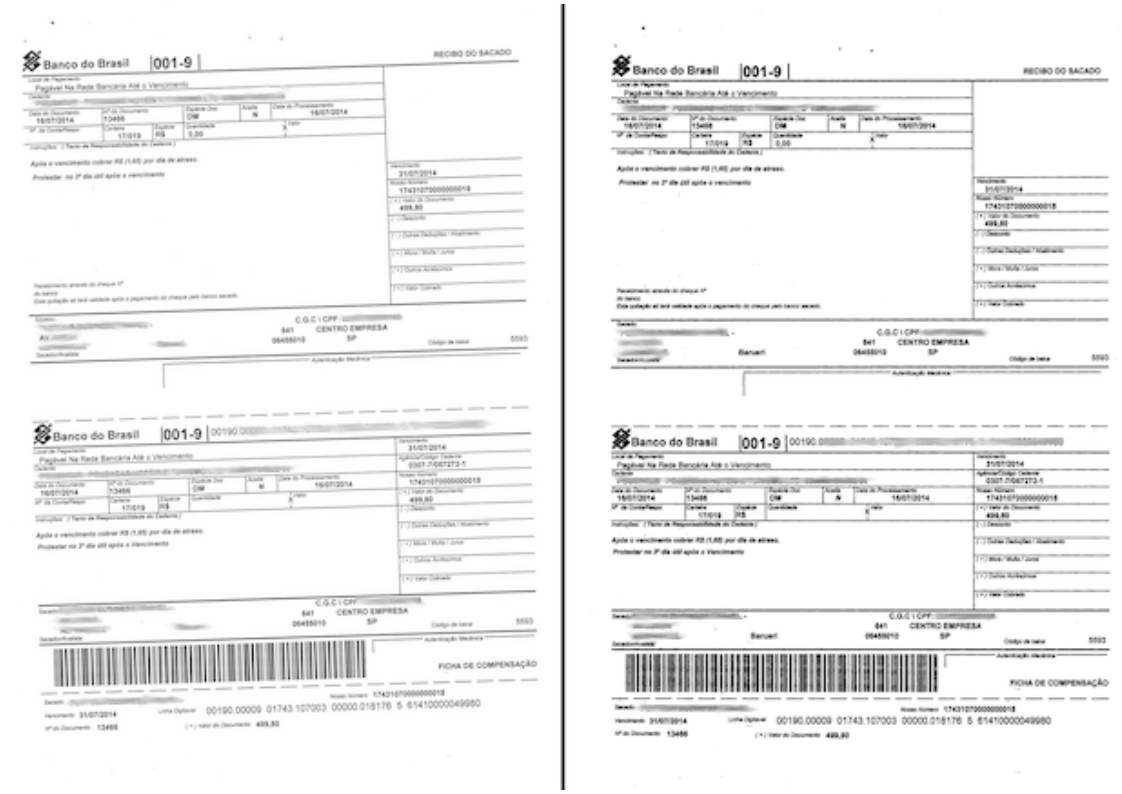
## 2.4 Pré-processamento de Imagens

Imagens de documentos digitalizados tipicamente apresentam defeitos inerentes ao processo de escaneamento, como distorções e ruídos. É necessário submetê-las a algumas etapas de tratamento, de modo que atinjam um padrão mínimo de qualidade antes de serem submetidas aos modelos de classificação. Em Marinai, Gori e Soda (2005) são mencionados frequentemente os métodos de detecção e correção de distorções e limiarização.

Um tipo de distorção comum que tem origem no processo de escaneamento de documentos é a inclinação da página. Mesmo que pequena, esta inclinação pode prejudicar gravemente a classificação de um documento. Um método comum utilizado

para a correção de distorções é a transformada de Hough (GONZALEZ; WOODS, 2006), que é utilizada para detectar padrões similares a linhas em uma imagem, e a partir dessas linhas identificar o ângulo de rotação. No caso de imagens de documentos, o texto, por ter um padrão regular, faz o papel das linhas. Uma vez que o ângulo seja detectado, uma operação inversa de rotação é realizada para eliminar a distorção. Na figura 6 é possível ver exemplo de uma imagem distorcida à esquerda, e a mesma imagem à direita com a distorção corrigida e com a limiarização aplicada.

Figura 6 – Exemplo de correção de distorção e limiarização.




Fonte – Elaborado pelo autor

O processo de limiarização (*thresholding*) é útil para a eliminação de ruído em imagens em tons de cinza, como pontos e marcas estranhos à imagem, introduzidos pelo *scanner* ou mesmo por defeitos no documento original em papel. A ideia da limiarização é simples: um valor de intensidade é escolhido como limiar (*threshold*) para determinar se um *pixel* é branco ou preto. *Pixels* com intensidade acima do limiar são considerados brancos, e abaixo do limiar são considerados pretos. O tratamento resulta na eliminação ou atenuação de pequenos defeitos, tornando a imagem mais pura para o modelo.

## 2.5 Redução de Informação

Em um contexto de análise de dados, imagens podem ser representadas digitalmente na forma de *bitmaps*, que são matrizes de *pixels* de diferentes intensidades, ou na forma de um histogramas, que representam a distribuição das diferentes intensidades dos *pixels* da imagem (CHAPELLE, 1998). *Bitmaps* podem ser considerados como uma representação mais próxima da realidade, já que é a representação que a visão humana compreende como uma imagem. Histogramas fornecem uma representação estatística das características de uma imagem, e podem ser úteis para o processamento de imagens coloridas.

Escolhida uma representação, se o número de amostras de treinamento não for significativamente maior do que a quantidade de variáveis dos dados a serem classificados, pode ocorrer sobreajuste. Uma solução para este problema é a redução de informação (*dimensionality reduction*), que consiste em identificar e extrair um conjunto de variáveis representativo e menor do que a quantidade de variáveis do dado original para ser submetido ao modelo de classificação. Além de ajudar a evitar o sobreajuste, dados associados com menor quantidade de variáveis requerem menos recursos computacionais para o seu processamento. Para este trabalho, que se concentra em imagens em tons de cinza, o processo de redução de informação será realizado com a representação do tipo *bitmap*.

Uma técnica comum para redução de informação é a análise de componentes principais (PCA, ou *Principal Component Analysis*) (ABDI; WILLIAMS, 2010), que consiste em identificar as variáveis que respondem pela maior parte da variância dos dados apresentados. A partir das variáveis analisadas são derivadas novas variáveis, chamadas de componentes principais, que têm como objetivo representar o dado original de forma mais sucinta. Devido a esta característica, PCA pode ser considerado um método eficaz de compressão de dados.

No caso de imagens em formato *bitmap*, é feita a conversão da representação matricial para uma representação vetorial, em que todas as linhas da matriz original são colocadas lado a lado, de modo a formar um grande vetor para cada imagem, e na qual cada elemento desse vetor, correspondente a um *pixel*, é considerado uma variável aleatória. Uma vez que todas as imagens do conjunto de dados estejam representadas dessa forma, o PCA é aplicado sobre todo o conjunto de dados de treinamento para identificar os componentes e os valores de média e variância das variáveis analisadas, para efeitos de normalização dos dados, ou seja, a transformação das variáveis originais em variáveis com média $\mu = 0$ e desvio padrão $\sigma = 1$. Calculados os componentes, o PCA os analisa para identificar quantos são necessários para representar a maior parte da variância dos dados, tipicamente 95%, e estes são os componentes principais, que serão utilizados posteriormente para treinamento dos

modelos de classificação.

Após o treinamento do modelo com os dados de treinamento reduzidos, os dados novos a serem submetidos ao modelo também devem ser reduzidos, mas desta vez com base nos parâmetros obtidos pelo PCA realizado com os dados de treinamento, de modo que a redução de informação e normalização seja consistente entre os conjuntos de dados de treinamento e novo.

## 2.6 Análise de Risco

Westerman e Hunter (2009) discutem a relação entre risco para o negócio da empresa com os riscos associados a uma operação de tecnologia de informação (TI). Os autores justificam a sua abordagem ao citar que, desde o início do século XXI, a TI se tornou peça fundamental na condução de boa parte dos negócios, devido principalmente ao surgimento de leis como Sarbanes-Oxley, que apesar de tratarem principalmente de regras para auditorias contábeis, falam extensamente sobre segurança de dados e detecção e prevenção a fraudes, que são dois assuntos que colocam a TI em evidência, e ainda responsabilizam a alta administração por fraudes ou falhas operacionais. Todos estes fatores contribuem para tornar a gestão de risco um tema relevante.

Risco é incerteza, e no caso da TI, riscos não são totalmente compreendidos pela alta administração. O trabalho cita o caso de uma companhia aérea, que postergou por diversas vezes a substituição de um sistema de informação crítico, até que este sistema falhou e causou enormes prejuízos. Este caso exemplifica a falta de compreensão dos riscos associados a uma decisão de TI.

Este artigo propõe uma linguagem comum para a gestão de risco, batizada de *framework* 4A. Cada “A” corresponde a uma categoria de risco fundamental:

a) Disponibilidade (*Availability*): qual o risco de um sistema ficar indisponível?

b) Acesso (*Access*): qual o risco de um sistema permitir acessos indevidos, causando vazamento de dados, por exemplo?

c) Precisão (*Accuracy*): qual o risco de um sistema fornecer informações incorretas, causando erros de contabilidade, por exemplo?

d) Agilidade (*Agility*): qual o risco de uma determinada atividade ou projeto não ser concluído dentro de um prazo legal, por exemplo?

O uso das quatro categorias busca tornar a incerteza do risco de TI mais gerenciável, ao auxiliar na conversão de problemas técnicos em problemas de negócio, e dessa forma apresentar os problemas em termos claros para a alta administração.

O artigo sugere que, para a implementação do *framework*, seja realizado um processo de identificação do perfil de risco da companhia, com questões a serem respondidas nos níveis executivo e operacional, para cada uma das quatro categorias de risco. Estas questões servem para melhorar a compreensão sobre a importância dos riscos de TI, e contribuem para reduzir a distância entre as diversas áreas e níveis de uma organização, em termos de risco.

O artigo apresenta três disciplinas fundamentais a serem desenvolvidas, como forma de implementar um processo de gestão de riscos:

a) Governança de riscos: processos e políticas que forneçam uma visão global dos riscos de TI;

b) Cultura de gestão de riscos (*risk-aware culture*): todos devem possuir conhecimento sobre riscos, e serem incentivados a discuti-los abertamente;

c) Fundação: infraestrutura, aplicações e equipe de suporte bem estruturados e gerenciados, com a mínima complexidade necessária.

Complexidade na fundação é o maior gerador de riscos de TI para uma empresa. A complexidade pode estar presente na forma de muitos sistemas diferentes, sistemas redundantes e com sobreposição de funcionalidades, muitos tipos diferentes de *hardware*, integrações e interações desconhecidas entre sistemas, sistemas com tecnologia considerada obsoleta, em que não há abundância de profissionais no mercado, entre outros. Tipicamente esta complexidade não foi planejada, e sim surgiu ao longo do tempo como forma de atender às necessidades da empresa. De acordo com o trabalho, empresas com uma fundação bem gerenciada e sem complexidade excessiva, apresentam risco menor do que as demais em todas as quatro categorias de risco mencionadas.

A implementação de um processo de governança de riscos inicia-se com um dilema: as pessoas mais capazes de tomar decisões em alto nível, para toda a empresa, são as menos capazes de compreender os riscos em detalhe, e vice-versa. Políticas e métodos claros de gestão de risco podem auxiliar este processo, ao permitir que pessoas em todos os níveis hierárquicos sejam capazes de identificar e avaliar riscos nas suas áreas. Isto contribui para a criação de uma cultura de gestão de riscos, em que todos compreendem como suas decisões aumentam ou reduzem a exposição da empresa a riscos, e onde todos têm liberdade para discutir sobre riscos. É trabalho da alta administração incentivar este comportamento na empresa. Acredita-se que a discussão constante em todos os níveis facilita a identificação de novos riscos, antes considerados imprevisíveis.

O trabalho conclui tratando da impossibilidade de eliminação de todos os riscos de TI. O termo aqui é gestão: riscos bem geridos são reduzidos, e auxiliam a tomada

de decisão, como por exemplo, na identificação de projetos de atualização de sistemas que são prioritários.

O *framework* 4A serve como referência para a criação do método de análise de risco proposto neste trabalho, quanto à definição dos tipos básicos de riscos a serem analisados em um contexto de classificação de documentos, a saber:

a) Risco de Erro (Precisão): qual o risco do sistema classificar documentos incorretamente?

b) Risco de Indisponibilidade (Disponibilidade): qual o risco do sistema de classificação ficar indisponível?

c) Risco Temporal (Agilidade): qual o risco de atividades de classificação ou treinamento levarem mais tempo do que o previsto, e com isso causar atraso no processamento?

A categoria acesso não será abordada neste trabalho, por ser considerada uma particularidade de implementação da aplicação, que não é pertinente ao desempenho de modelos de classificação, assim como em termos de disponibilidade, não serão consideradas questões de infraestrutura, que são abordadas de forma mais ampla pelo *framework* 4A.

Para este trabalho será tratado como risco de indisponibilidade a demanda por memória para treinamento de um determinado modelo, considerando que uma demanda extrema pode tornar o sistema inutilizável, ou requerer maior investimento em *hardware*. Quanto à precisão, serão consideradas quais quantidades de exemplos de dados para cada categoria que possibilitam a classificação de dados com maior correção. Quanto à agilidade, será avaliado o desempenho em termos de tempo de treinamento e classificação para cada modelo.

## 2.7 Conclusão

Nesta seção foram apresentados trabalhos selecionados da área de classificação de documentos, particularmente os que tratam do uso de Redes Neurais Artificiais e Máquinas de Vetor Suporte para tal objetivo. Também foi apresentado o *framework* 4A, no qual é baseado o método para análise de risco a ser proposto neste trabalho. Na seção a seguir, será apresentado de forma detalhada o método para análise de risco, bem como o protótipo utilizado para implementá-lo.

# 3 MÉTODO PARA ANÁLISE DE RISCO

## 3.1 Introdução

Nesta seção será apresentado o método para análise de risco proposto, com seus critérios, fórmulas e os passos necessários para executá-lo, e o processo de classificação de documentos a ser utilizado durante a análise. Além disso, será apresentada a especificação de protótipo de *software* para calcular o risco de forma empírica, por meio da execução de diferentes modelos de classificação contra um conjunto de dados pré-definido e do armazenamento de resultados detalhados, que possibilitem a comparação entre modelos de acordo com os diferentes critérios propostos pelo método. Finalmente, será proposto um plano de testes para ser executado pelo protótipo, com o uso de bibliotecas de *software* para classificação de dados existentes.

O método proposto possui os seguintes passos, que são descritos detalhadamente na seção 3.4:

a) Coleta de Dados;

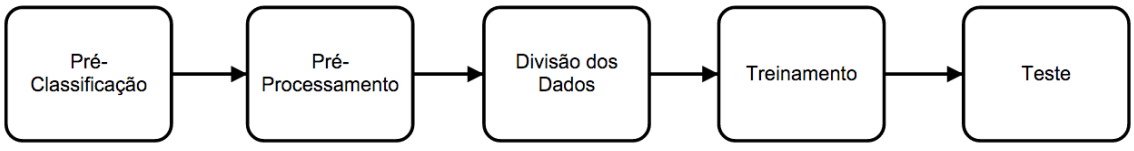
b) Cálculo do Risco Médio;

c) Cálculo do Risco Total Estimado.

## 3.2 Processo de Classificação de Documentos

No que se refere à coleta de dados, julgou-se necessário o uso de um processo padronizado de classificação de documentos que possibilite a comparação de diferentes modelos de classificação em condições de igualdade, e que minimize a possibilidade de que distorções sejam introduzidas nos dados coletados. Na bibliografia analisada foi identificado um conjunto de passos genéricos que são comumente utilizados em processos de classificação de documentos, e com base nestes passos foi proposto o processo a ser utilizado durante este trabalho, que está dividido nas fases mostradas na figura 7. Cabe notar o uso do termo “classificador” daqui em diante, para indicar uma instância de um modelo de classificação, treinada e pronta para uso.

Figura 7 – Processo de classificação de documentos proposto.




Fonte – Elaborado pelo autor

As fases mostradas na figura 7 podem ser descritas conforme abaixo:

a) Pré-classificação: coleta de imagens de documentos e identificação manual de acordo com seus respectivos tipos;

b) Pré-processamento: tratamento das imagens e redução de informação;

c) Divisão dos dados em conjuntos de dados de treinamento e teste;

d) Treinamento: criação dos classificadores com base no conjunto de dados de treinamento;

e) Teste: execução dos classificadores contra o conjunto de dados de teste, com o objetivo de aferir seu desempenho.

A fase de pré-classificação deve ser executada somente uma vez durante a preparação dos dados, e consiste em coletar um conjunto de imagens de documentos e identificá-las de acordo com seus respectivos tipos. Essa identificação prévia, também conhecida como etiquetação, tem o objetivo de permitir a avaliação do desempenho do classificador, por meio da comparação do resultado obtido com o resultado esperado.

A fase de pré-processamento tem como objetivos o melhoramento das imagens, por meio de correções de distorções e limiarização, e a redução de informação, por meio da conversão para tons de cinza. O algoritmo de Análise de Componentes Principais (PCA) é aplicado durante a fase de treinamento, quando o conjunto de dados de treinamento é avaliado com o objetivo de se encontrar um conjunto mínimo de variáveis que o represente adequadamente, de modo a utilizar um conjunto de dados reduzidos e assim reduzir a necessidade de capacidade computacional para o seu processamento.

A divisão dos dados tem como objetivo reservar uma parte dos dados pré-classificados somente para testes, de modo que não sejam utilizados durante o processo de treinamento. Este conjunto de dados não processados durante o treinamento será utilizado posteriormente na fase de teste. Esta divisão dos dados e execução de processos de treinamento e teste é chamada de validação cruzada (*cross-validation*) (JAMES et al., 2013), que serve para avaliar a capacidade de generalização do classificador, e consequentemente, estimar seu desempenho contra dados desconhecidos.

## 3.3 Métricas de Risco

De acordo com o *framework* FAIR (*Factor Analysis for Information Risk*) (The Open Group, 2013), dois fatores são fundamentais na caracterização de risco: a incerteza e a magnitude da perda. Destes fatores segue que o risco deve sempre ser representado na forma de probabilidade de perda, e deve possuir um componente de custo, que represente a magnitude da perda. Uma consequência direta desta relação

é que uma perda cuja magnitude seja mínima e probabilidade de ocorrer seja alta, representa um risco baixo, o mesmo valendo para a relação inversa, com perdas de grande magnitude mas improváveis.

A seguir são apresentados os critérios propostos de medição de probabilidade de perda para as três categorias de risco propostas (precisão, disponibilidade e agilidade), sendo que o critério de agilidade está dividido em agilidade de treinamento e de classificação.

- Risco de Erro (Precisão)

  Para avaliar a precisão dos classificadores, será utilizada a medida chamada de *F1-score* (RIJSBERGEN, 1974), que tem origem na área de Recuperação de Informação e trata de avaliar o percentual de documentos recuperados corretamente em uma busca. No contexto deste trabalho, esta medida será utilizada para determinar o percentual de documentos classificados corretamente.

  O *F1-score* é calculado de acordo com a equação a seguir:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precisao} \cdot \mathrm{Sensitividade}}{\mathrm{Precisao} + \mathrm{Sensitividade}} \tag{3.1}$$

  Em que precisão e sensitividade são definidos conforme as equações abaixo:

$$\mathrm{Precisao} = \frac{TP}{TP + FP} \tag{3.2}$$

$$\mathrm{Sensitividade} = \frac{TP}{TP + FN} \tag{3.3}$$

  Em que *TP*, de *true positives* ou positivos verdadeiros, é a quantidade de documentos classificados corretamente, *FP*, de *false positives* ou falsos positivos, é a quantidade de documentos classificados incorretamente como pertencendo a uma categoria, *FN*, de *false negatives* ou falsos negativos, é a quantidade de documentos classificados incorretamente como *não* pertencendo a uma categoria, e ainda, *TN*, de *true negatives* ou negativos verdadeiros, que não aparece na equação mas completa o conceito, é a quantidade de documentos classificados corretamente como *não* pertencendo a uma categoria.

  Tipicamente estas informações são dispostas em uma estrutura chamada matriz de confusão, conforme exemplificado na tabela 4.

  Na tabela 4 é possível ver um exemplo de matriz de confusão, em que as colunas indicam as quantidades corretas de documentos de cada categorias, e a linhas indicam as categorias estimadas pelos classificadores.

Tabela 4 – Exemplo de matriz de confusão.

| | **0** | **1** |
|---|---|---|
| 0 | 97 | 1 |
| 1 | 3 | 99 |

Fonte – Elaborado pelo autor

De acordo com o exemplo, ao se tomar como referência um problema de classificação binária análogo ao que será utilizado neste trabalho, em que é possível chamar as classes de 1 e 0, como positivo e negativo, tem-se os valores da tabela 5.

Tabela 5 – Valores da matriz de confusão.

| **Variável** | **Valor** |
|---|---|
| TN | 97 |
| FP | 3 |
| FN | 1 |
| TP | 99 |

Fonte – Elaborado pelo autor

O *F1-score* tem a propriedade importante de ter seu valor restrito ao intervalo $[0, 1]$, o que simplifica o cálculo do risco quando este é colocado ao lado de probabilidades, que por definição possuem seu valor restrito ao mesmo intervalo. Dado que para este trabalho o interesse é calcular o risco de erro de classificação, este é definido como $R_{prec} = 1 - F1$.

- Risco de Indisponibilidade (Disponibilidade)

  O risco de indisponibilidade $R_{mem}$ é definido como sendo a probabilidade que, durante a fase de treinamento de um classificador com $n$ amostras, a quantidade de memória utilizada *MemNec* exceda o consumo de memória médio observado *MemMed*, mais um percentual excedente da média *PctExced*, análogo a uma margem de segurança, a ser definido:

$$R_{mem} = P(\text{MemNec} > \text{MemMed} + \text{PctExced}) \tag{3.4}$$

- Risco Temporal de Treinamento (Agilidade)

  O risco temporal de treinamento $R_{trein}$ é definido como sendo a probabilidade que, durante a fase de treinamento de um classificador com $n$ amostras, o tempo necessário *TempoTrein* exceda o tempo médio observado *TempoMedTrein*, mais um percentual excedente da média *PctExced*, a ser definido:

$$R_{trein} = P(\text{TempoTrein} > \text{TempoMedTrein} + \text{PctExced}) \tag{3.5}$$

- Risco Temporal de Classificação (Agilidade)

  O risco temporal de classificação $R_{class}$ é definido como sendo a probabilidade de o tempo necessário *TempoClass* para classificar *d* documentos exceda o tempo médio observado *TempoMedClass*, mais um percentual excedente da média *PctExced*, a ser definido:

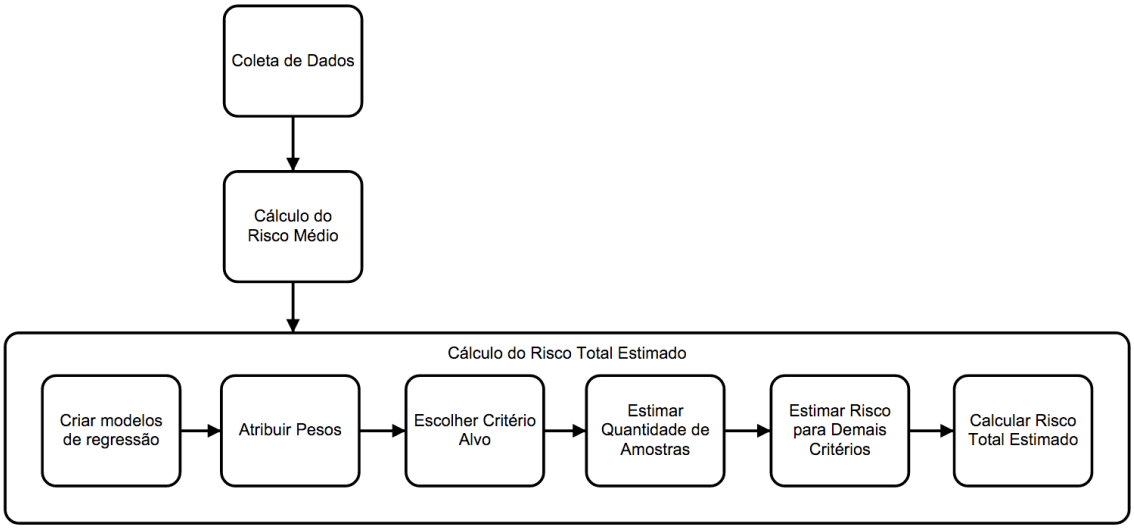
$$R_{class} = P(\text{TempoClass} > \text{TempoMedClass} + \text{PctExced}) \tag{3.6}$$

Para as variáveis *MemNec*, *TempoTrein* e *TempoClass*, assume-se a premissa de que são variáveis aleatórias que seguem distribuição T de Student com $n-1$ graus de liberdade.

## 3.4 Método

Uma vez definidas as métricas de risco e o processo de classificação a ser utilizado, tem-se a base para o método de cálculo de risco, conforme mostrado na figura 8.

Figura 8 – Método para cálculo de risco proposto.




Fonte – Elaborado pelo autor

Os passos mostrados na figura 8 são descritos a seguir:

- Coleta de Dados

  As etapas de treinamento e teste dos modelos de classificação devem ser executadas *m* vezes para cada valor *n* de amostras selecionado, e as informações

necessárias para o cálculo de risco (tempos de treinamento e classificação, consumo de memória e precisão) devem ser registradas. Os valores para *m* e *n* são definidos na seção 3.5;

- Cálculo do Risco Médio

  Com base nos dados coletados para cada valor *n* de amostras selecionado, devem ser calculadas as médias dos riscos temporais, de erro e de indisponibilidade para as *m* execuções realizadas, que são os riscos médios aferidos de um determinado classificador. Para determinar os riscos temporais e de indisponibilidade é preciso definir o percentual de desvio da média aceitável para cada um destes critérios. Neste trabalho será utilizado um valor inicial arbitrário de 5%, para efeito de testes;

- Cálculo do Risco Total Estimado

  Com base nos dados coletados e nos riscos médios aferidos, este trabalho propõe os seguintes passos para o cálculo do risco total estimado:

  a) Escolher um critério e um valor alvo desejado para o mesmo. Por exemplo, risco de erro de 5%;
  b) Estimar a quantidade de amostras necessárias para obter o valor alvo de risco do critério escolhido;
  c) Estimar o risco para os critérios restantes, com base na quantidade de amostras estimada no passo anterior;
  d) Atribuir pesos para cada um dos critérios de risco, de acordo com necessidades de negócio;
  e) Calcular o risco total estimado como sendo a média ponderada dos riscos estimados e seus pesos.

  O valor do critério escolhido como alvo, chamado $R_{alvo}$, representa um objetivo desejado pelo usuário do método, como por exemplo, um risco de erro de 5%, a partir do qual será estimada a quantidade de amostras necessária para seu atingimento, e baseado nessa quantidade, posteriormente serão estimados os demais riscos, chamados $\hat{R}_2$, $\hat{R}_3$ e $\hat{R}_4$ por meio de modelos de regressão.

  Devem ser criados os modelos de regressão listados nas tabelas 6 e 7, com o objetivo de estimar a quantidade de amostras necessárias para valores desejados de riscos temporais, de erro ou de indisponibilidade, e em seguida estimar os valores de riscos temporais, de erro ou de indisponibilidade com base na quantidade de amostras estimada na etapa anterior. São utilizados modelos lineares para a realização destas estimativas com o objetivo de se estimar o risco para quantidades de amostras *n* não testadas.

Tabela 6 – Modelos para estimação de quantidade de amostras.

| **Resposta** | **Preditor** |
|---|---|
| # amostras | Risco de Erro |
| # amostras | Risco Temporal de Treinamento |
| # amostras | Risco Temporal de Classificação |
| # amostras | Risco de Indisponibilidade |

Fonte – Elaborado pelo autor

Tabela 7 – Modelos para estimação dos demais critérios com base na quantidade de amostras.

| **Resposta** | **Preditor** |
|---|---|
| Risco de Erro | # amostras |
| Risco Temporal de Treinamento | # amostras |
| Risco Temporal de Classificação | # amostras |
| Risco de Indisponibilidade | # amostras |

Fonte – Elaborado pelo autor

Os pesos, a serem atribuídos pelo usuário do método, representam a importância relativa de cada critério para uma determinada situação, e podem ser interpretados como sendo o componente de custo do risco. Formalmente, os pesos são definidos como $w_{alvo}$, $w_2$, $w_3$ e $w_4$, que devem ser alocados de modo a somar 1, em que $w_{alvo}$ é o peso atribuído ao critério escolhido como alvo e para o qual se conhece o valor desejado, e $w_2$, $w_3$ e $w_4$ são os pesos para os demais riscos estimados. Definidas as variáveis, o risco total estimado é calculado como sendo a média ponderada dos riscos e seus pesos, e pode ser expresso de acordo com a equação a seguir:

$$\rho_{modelo} = R_{alvo} w_{alvo} + \hat{R}_2 w_2 + \hat{R}_3 w_3 + \hat{R}_4 w_4$$

Ao se considerar que há um conjunto de variáveis em questão, propõe-se que este risco total estimado seja apreciado apenas em termos relativos, quando comparado com um outro classificador avaliado pelo mesmo método e com os mesmos critérios.

## 3.5 Plano de Testes

O plano de testes descrito a seguir tem por objetivo definir como estes devem ser realizados, bem como os parâmetros a serem utilizados durante os testes, de modo que os resultados sejam comparáveis. A primeira fase do teste é a de pré-processamento dos documentos, que, a partir dos documentos pré-classificados manualmente, irá criar um banco de dados de imagens com todos os tratamentos corretivos descritos

aplicados. Esta fase deve ser executada uma única vez e as imagens armazenadas poderão ser utilizadas em todos os testes posteriores.

Uma vez criado o banco de dados de imagens pré-processadas, os testes serão executados *m* vezes, conforme os passos descritos a seguir:

a) Selecionar *n* amostras de cada categoria para treinamento;

b) Selecionar *d* amostras de cada categoria para teste de classificação;

c) Treinar um classificador com *n* amostras, para cada modelo selecionado;

d) Classificar *d* amostras de cada categoria em cada classificador;

e) Calcular precisão, consumo de memória e tempos de treinamento e teste para cada classificador.

Com base nestes passos, serão coletados os dados necessários para o cálculo de risco descrito na seção anterior.

Abaixo, os parâmetros considerados para o plano de testes, com seus respectivos valores:

a) Modelos de classificação: *Perceptron* de Múltiplas Camadas e Máquina de Vetor Suporte;

b) Categorias de documentos: Notas Fiscais e Boletos bancários;

c) Quantidade de documentos de cada categoria: 300;

d) Quantidade de documentos *d* a serem utilizados na fase de teste: 100 de cada categoria;

e) Quantidades de amostras *n*, a serem utilizadas na fase de treinamento: de 20 a 200 de cada categoria, variando de 20 em 20;

f) Quantidade de execuções *m* das fases de treinamento e teste, para cada quantidade de amostras *n*: 40;

g) Resolução das imagens após o tratamento: 128 x 180 pixels, resultando em um vetor de 23040 variáveis por imagem.

A quantidade de documentos de cada categoria foi determinada com base na disponibilidade de dados de produção para a realização do trabalho. Este conjunto de dados será dividido de modo que 100 documentos de cada categoria serão utilizados exclusivamente para testes de classificação, e os restantes utilizados para treinamento. A subdivisão dos dados de treinamento em conjuntos de 20 a 200 amostras para cada categoria busca atender ao objetivo de observar a influência do tamanho do conjunto de dados de treinamento sobre o desempenho do modelo de classificação, descrito no plano de testes. A quantidade de execuções proposta foi escolhida devido à

necessidade de ser ter um conjunto de dados suficientemente grande para possibilitar a premissa de normalidade, com $n \geq 30$.

Este plano de testes será executado pelo protótipo, que será definido na seção seguinte.

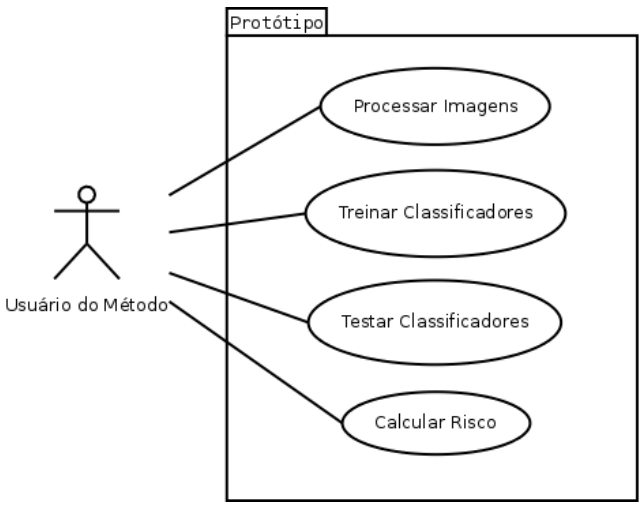
## 3.6 Protótipo de *Software* para Cálculo de Risco

O protótipo é o meio pelo qual serão executados os testes, coletados os resultados e calculados os riscos para os modelos de classificação em análise. Nesta seção será apresentada a especificação do protótipo, dividida em casos de uso, estruturas de dados e diagramas de sequência.

- Casos de Uso

  Na figura 9, tem-se os casos de uso identificados para o protótipo.

Figura 9 – Modelo de casos de uso.




Fonte – Elaborado pelo autor

A seguir uma breve descrição de cada um dos casos de uso exibidos no diagrama:

a) Processar Imagens: aplica tratamentos ao lote de imagens selecionado (converte imagens para tons de cinza, redimensiona para tamanho padrão, corrige distorções e aplica limiarização) e grava arquivo de texto com registros delimitados por vírgula (CSV), com uma imagem por linha e um *pixel* por coluna;

b) Treinar Classificadores: treina modelos de classificação de dados com base em imagens pré-processadas;

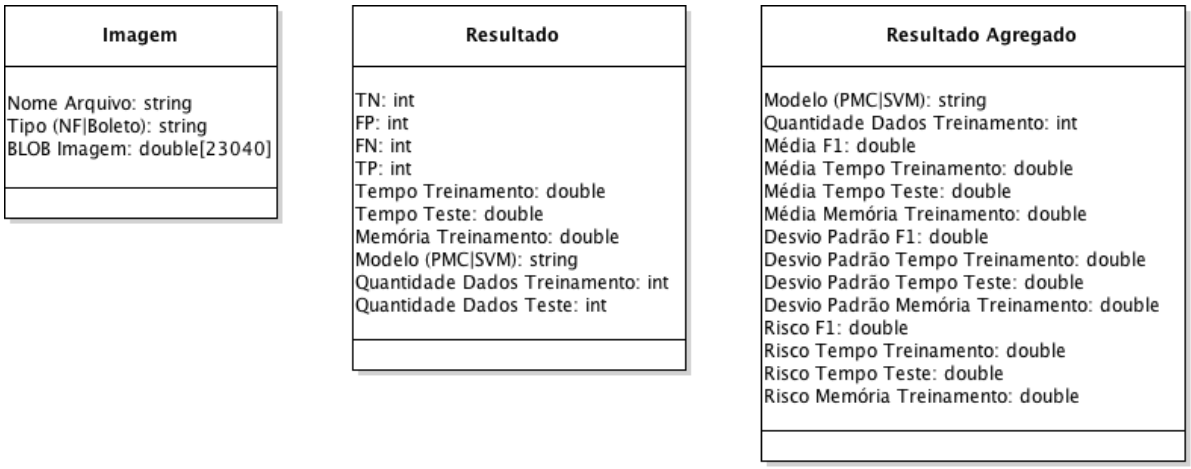
c) Testar Classificadores: testa modelos de classificação de dados com base em imagens pré-processadas;

d) Calcular Risco: calcula risco com base nos resultados obtidos nos processos de treinamento e teste dos modelos de classificação, somados à escolha de critério alvo de risco e atribuição de pesos aos critérios de risco pelo usuário do método.

- Estruturas de Dados

  Na figura 10 estão descritas as estruturas de dados a serem utilizadas pelo protótipo. Estas estruturas são implementadas na forma de *data frames*, que são estruturas existentes na linguagem R análogas a tabelas, no sentido de que cada linha da tabela representa uma instância do dado a ser utilizado, e as colunas representam os atributos do dado.

Figura 10 – Estruturas dos *data frames* utilizados.




Fonte – Elaborado pelo autor

A seguir uma breve descrição de cada uma das estruturas exibidas na figura 10:

a) Imagem: representa uma imagem pré-processada, transformada em vetor de *pixels*, com o nome do arquivo original e a categoria a que a imagem pertence;

b) Resultado: representa os resultados de uma iteração de treinamento e teste, com informações de precisão, tempos de treinamento e teste e consumo de memória para treinamento, bem como do modelo de classificação utilizado e das quantidades de amostras utilizadas para treinamento e teste;

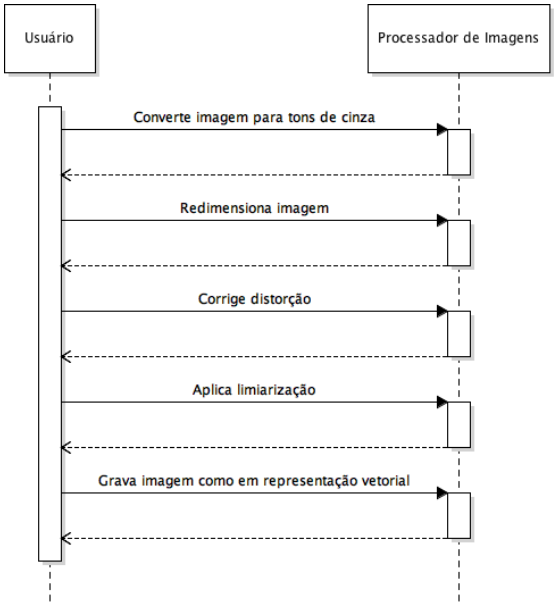
c) Resultado Agregado: representa o resultado agregado de várias iterações de treinamento e teste, agrupadas por modelo de classificação e

por quantidade de amostras utilizadas para treinamento, com as médias, desvios padrões e valores de risco médio para precisão, tempos de treinamento e teste e consumo de memória para treinamento.

- Diagramas de Sequência

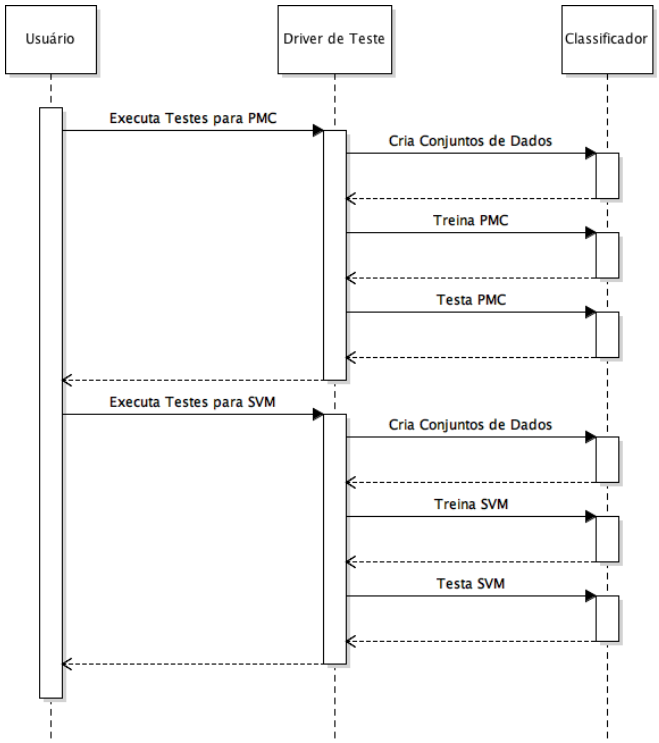
  Na figuras 11, 12 e 13 a seguir tem-se os diagramas de sequência correspondente aos casos de uso descritos na figura 9. É importante notar que aqui são descritas interações entre funções, e não entre objetos, como comumente observado em diagramas deste tipo.

Figura 11 – Diagrama de sequência do caso de uso “Processar Imagens”.



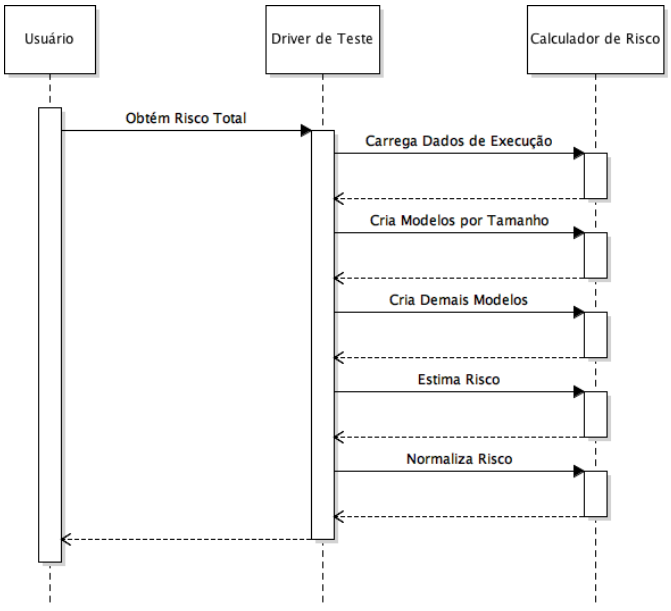

Fonte – Elaborado pelo autor

Figura 12 – Diagrama de sequência dos casos de uso “Treinar Classificadores” e “Testar Classificadores”.




Fonte – Elaborado pelo autor

Figura 13 – Diagrama de sequência do caso de uso “Calcular Risco”.




Fonte – Elaborado pelo autor

## 3.7 Conclusão

Nesta seção foi apresentado o método para análise de risco proposto, bem como o processo de classificação e o plano de testes a serem avaliados. Além disso, o protótipo de *software* a ser utilizado foi especificado. Na seção a seguir, será apresentada a análise dos resultados obtidos com a execução dos testes.

# 4 ANÁLISE DOS RESULTADOS

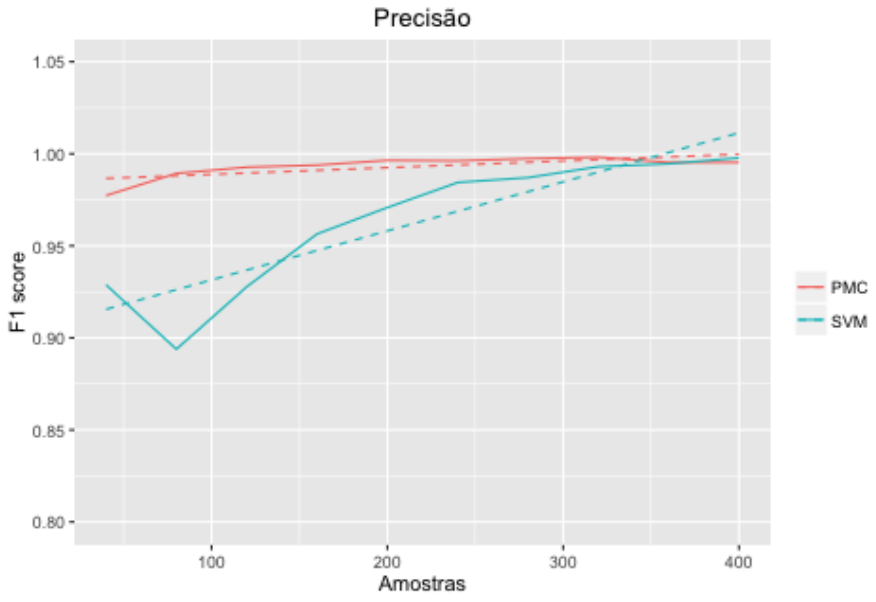
## 4.1 Introdução

Nesta seção será apresentada a análise dos dados gerados por meio de simulação para cada classificador avaliado de modo a evidenciar o comportamento observado, e também os resultados obtidos com a aplicação do método de análise de risco proposto na forma de cenários fictícios, criados para simular situações reais de produção que explorem os critérios de precisão, agilidade e disponibilidade.

## 4.2 Análise Exploratória de Dados

Na figura 14 é possível observar o desempenho aferido dos modelos de classificação em termos de precisão, de acordo com a quantidade de amostras de treinamento fornecidas.

Figura 14 – Gráfico de precisão média.




Fonte – Elaborado pelo autor

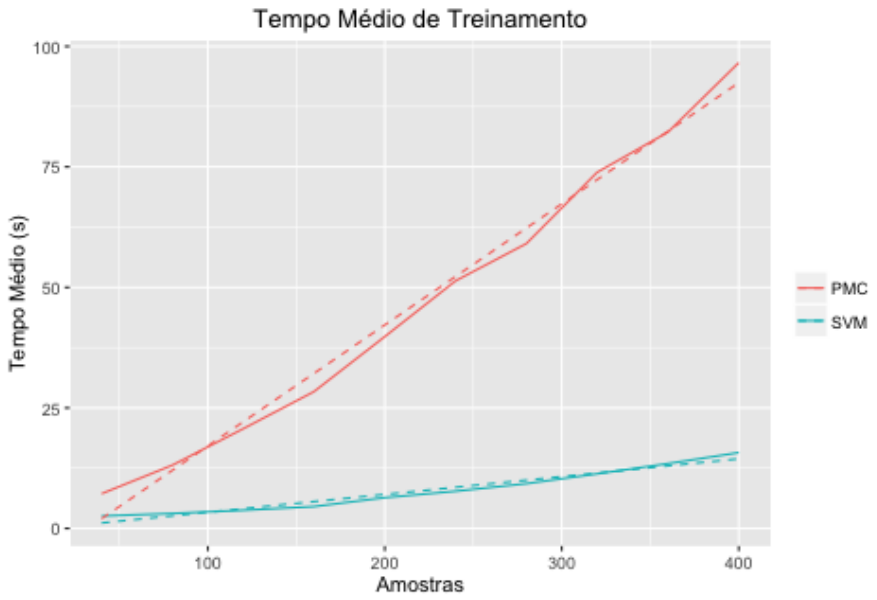
Observa-se que o *Perceptron* de Múltiplas Camadas (PMC) precisou de menor número de amostras do que a Máquina de Vetor Suporte (SVM) para atingir precisão próxima a 1, ou 100%. Causa estranheza ainda o fato de que em alguns casos o desempenho da SVM, na média, caiu com o incremento do número de amostras. Com base nestas informações é possível afirmar que para o caso em análise, de classifi-

cação de documentos, o PMC em média necessita de número menor de amostras de treinamento do que a SVM para obter valores de precisão próximos do ideal de 100%. A linha tracejada representa a reta de regressão linear calculada para os dados apresentados.

É possível observar que para SVM, devido à não-linearidade dos dados, a reta extrapola precisão de 100% para quantidades de amostras maiores do que 350, o que precisa ser ajustado para evitar que sejam utilizados valores incorretos no cálculo de risco.

Na figura 15 é possível observar o desempenho aferido dos modelos de classificação em termos de tempo necessário para treinamento, de acordo com a quantidade de amostras de treinamento fornecidas.

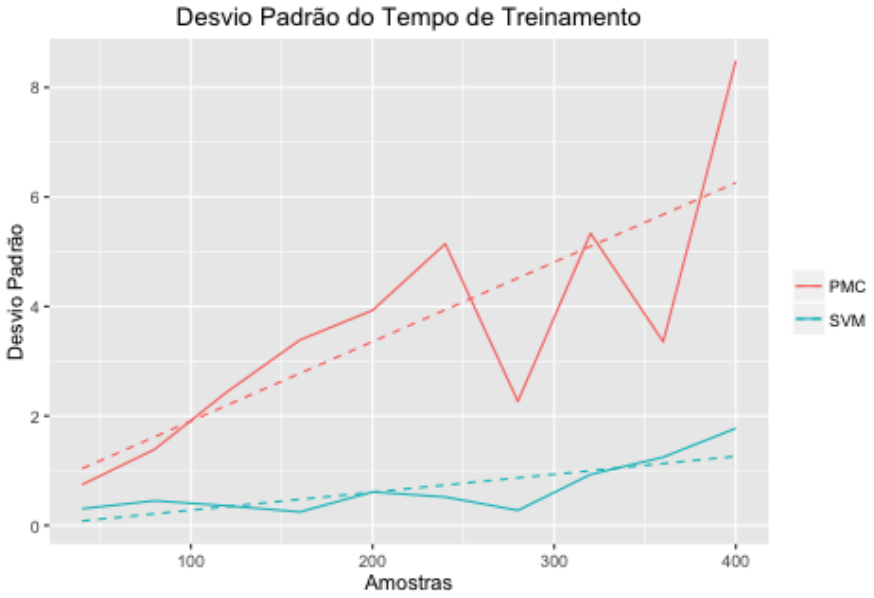
Figura 15 – Gráfico de tempo de treinamento médio.




Fonte – Elaborado pelo autor

Observa-se que o tempo de treinamento aumenta linearmente de acordo com o aumento do número de amostras, mas no caso do PMC, o aumento é significativamente mais acentuado quando comparado com o da SVM. Além disso, na figura 16, que mostra a relação entre desvio padrão do tempo de treinamento e quantidade de amostras, é possível verificar que a variabilidade do tempo de treinamento do PMC é maior do que o da SVM, o que implica menor capacidade de estimar corretamente o tempo necessário para treinamento de um PMC.

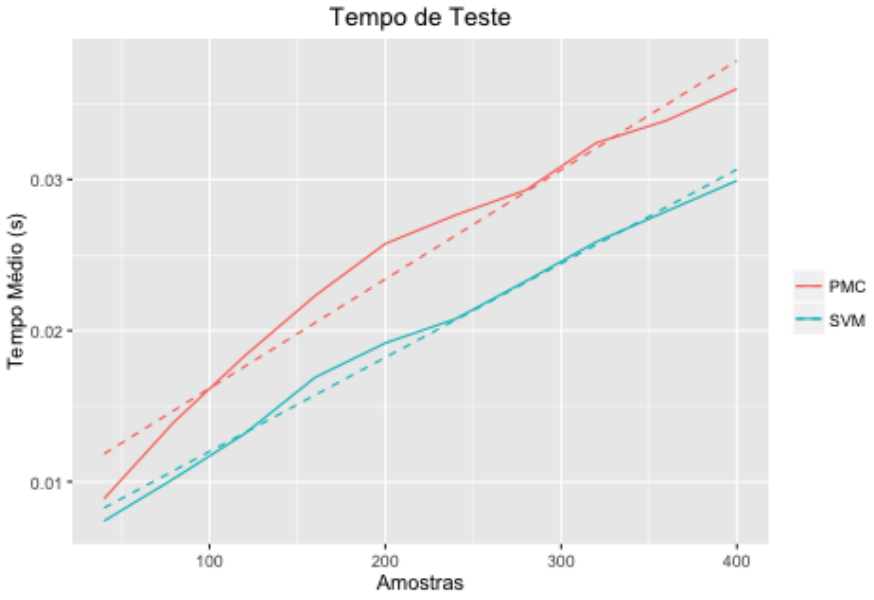
Figura 16 – Gráfico de desvio padrão do tempo de treinamento.




Fonte – Elaborado pelo autor

Na figura 17 é possível observar o desempenho dos modelos de classificação em termos de tempo necessário para classificação de um conjunto de novas imagens com tamanho fixo, de acordo com a quantidade de amostras de treinamento fornecidas.

Figura 17 – Gráfico de tempo de teste médio.

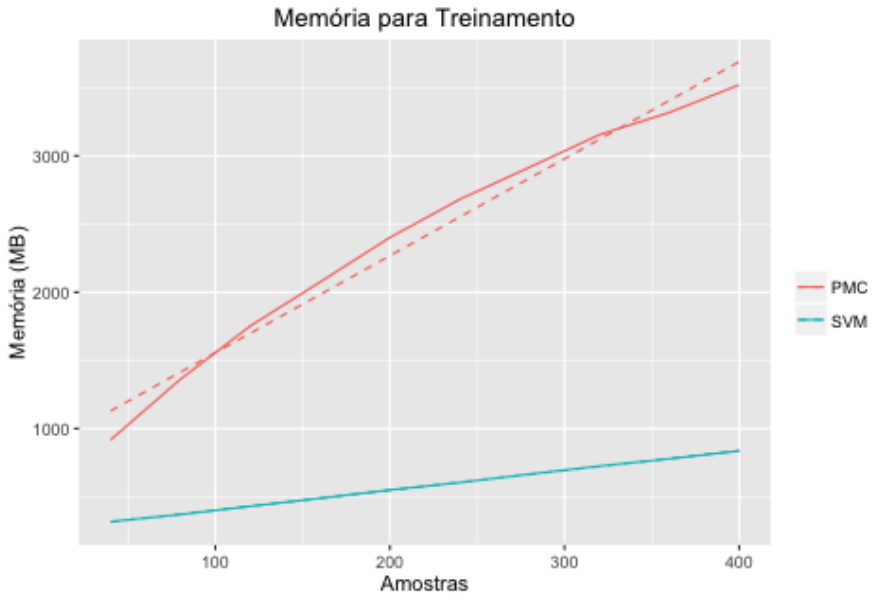



Fonte – Elaborado pelo autor

Os tempos de teste de classificação obtidos para PMC e SVM são equivalentes, com pequena vantagem para a SVM, e observou-se também correlação entre tempo de teste e quantidade de amostras de treinamento.

Na figura 18 é possível observar o desempenho dos modelos de classificação em termos da quantidade de memória necessária para treinamento, de acordo com a quantidade de amostras de treinamento fornecidas.

Figura 18 – Gráfico de consumo médio de memória.




Fonte – Elaborado pelo autor

Similarmente ao observado na figura 15, os consumos de memória para PMC e SVM possuem crescimento linear de acordo com o aumento do número de amostras, mas no caso do PMC, o aumento é mais acentuado quando comparado com o da SVM.

## 4.3 Cenários de Risco

Conforme detalhado na seção 3.4, o método proposto requer que o usuário decida quais pesos atribuir a cada um dos critérios de risco, e que escolha um critério alvo e um valor para o critério escolhido. Diante da grande quantidade de combinações possíveis para estes parâmetros, são propostos cenários para evidenciar o funcionamento do método em situações diversas. Em todos os cenários, busca-se determinar dentre os modelos de classificação avaliados qual possui o menor risco total estimado, de acordo com os parâmetros escolhidos. A seguir são listados os cenários propostos:

a) Cenário 1: Usuário busca risco de erro de 1% e dá peso maior para tempo de treinamento.

b) Cenário 2: Usuário busca risco de erro de 1% e dá peso maior para disponibilidade.

c) Cenário 3: Usuário busca risco de indisponibilidade de 5% e dá peso maior para precisão.

d) Cenário 4: Usuário busca risco de indisponibilidade de 5% e dá peso maior para tempo de treinamento.

e) Cenário 5: Usuário busca risco temporal de treinamento de 5% e dá peso maior para precisão.

f) Cenário 6: Usuário busca risco temporal de treinamento de 5% e dá peso maior para tempo de classificação;

g) Cenário 7: Usuário busca risco temporal de classificação de 5% e dá peso maior para tempo de treinamento;

h) Cenário 8: Usuário busca risco temporal de classificação de 5% e dá peso maior para disponibilidade.

O objetivo destes cenários é o de penalizar os modelos para os quais o valor do critério de maior peso seja maior do que o do outro modelo em comparação. Por exemplo, no cenário 1 o objetivo é penalizar modelos que requerem maior tempo para treinamento para atingir a precisão desejada, dessa forma favorecendo modelos cujo treinamento é mais rápido.

De acordo com cada um dos cenários propostos, foi determinado um conjunto de pesos para cada critério de risco, listados na tabela 8.

Tabela 8 – Pesos por critério para cada cenário.

| Cenário | Precisão | Tempo Treinamento | Tempo Classificação | Disponibilidade |
|---|---|---|---|---|
| 1 | 0.2 | 0.4 | 0.2 | 0.2 |
| 2 | 0.2 | 0.2 | 0.2 | 0.4 |
| 3 | 0.4 | 0.2 | 0.2 | 0.2 |
| 4 | 0.2 | 0.4 | 0.2 | 0.2 |
| 5 | 0.4 | 0.2 | 0.2 | 0.2 |
| 6 | 0.2 | 0.2 | 0.4 | 0.2 |
| 7 | 0.2 | 0.4 | 0.2 | 0.2 |
| 8 | 0.2 | 0.2 | 0.2 | 0.4 |

Fonte – Elaborado pelo autor

Uma vez definidos os critérios alvo e os pesos, é possível calcular o risco total para cada modelo em cada cenário. Na tabela 9 é possível ver o risco total calculado para os dois modelos avaliados em cada um dos cenários propostos, e qual o modelo escolhido com base no menor risco.

Tabela 9 – Risco total por cenário.

| Cenário | Risco Total PMC | Risco Total SVM | Modelo Escolhido |
|---|---|---|---|
| 1 | 0.1961 | 0.1694 | SVM |
| 2 | 0.1490 | 0.1194 | SVM |
| 3 | 0.1129 | 0.1390 | PMC |
| 4 | 0.1570 | 0.1849 | PMC |
| 5 | 0.0927 | 0.1241 | PMC |
| 6 | 0.1503 | 0.1873 | PMC |
| 7 | 0.0308 | 0.1294 | PMC |
| 8 | 0.0308 | 0.0926 | PMC |

Fonte – Elaborado pelo autor

Pode-se ver que, de modo geral, o modelo SVM é o preferido nos cenários em que os critérios de maior peso são tempo de treinamento e indisponibilidade, e o PMC é o preferido nos demais cenários, em que os critérios de maior peso são erro e tempo de classificação. Esses resultados são consistentes com os observados no início desta seção.

## 4.4 Conclusão

Nesta seção foram descritos os resultados obtidos com a execução dos modelos de classificação por meio do protótipo, bem como exemplos do funcionamento do método de análise de risco por meio de cenários. Na seção a seguir serão apresentadas as conclusões do trabalho, como contribuições, dificuldades encontradas durante sua elaboração e sugestões para continuidade.

# 5 CONCLUSÕES

Nesta seção serão apresentadas as conclusões deste trabalho, no que se refere ao atingimento dos objetivos propostos, resultados obtidos, dificuldades encontradas, contribuições e sugestões para pesquisas futuras.

Na seção 2 foram detalhados os principais modelos de classificação de documentos encontrados na literatura, a saber, *Perceptron* de Múltiplas Camadas e Máquinas de Vetor Suporte, e na seção 3 foi detalhado o método proposto para a análise de risco destes modelos. Os dados obtidos com a execução do plano de teste proposto na seção 3 foram analisados na seção 4.

Foram propostos três objetivos na seção introdutória deste trabalho, recapitulados a seguir:

a) Propor um método para avaliação de risco de modelos de classificação de documentos digitalizados encontrados na literatura;

b) Comparar, segundo o método de avaliação de risco a ser proposto, o desempenho de bibliotecas de *software* existentes de Redes Neurais e Máquinas de Vetor Suporte para classificação de documentos, por meio da criação de um protótipo e testes com dados reais, ou seja, não fabricados especificamente para os testes;

c) Identificar, dentre as técnicas de classificação de documentos escolhidas, as que possuem o menor risco associado em termos de disponibilidade, precisão e agilidade.

O método de avaliação de risco, objeto deste trabalho, foi proposto na seção 3, e foi colocado em prática conforme resultados mostrados e analisados na seção 4. Foi comparado, com base em um universo de 600 documentos reais de negócio, o desempenho das bibliotecas *nnet* e *e1071*, utilizadas por meio da biblioteca *caret*, e que implementam os modelos *Perceptron* de Múltiplas Camadas (PMC) e Máquinas de Vetor Suporte (SVM), respectivamente. Dentre os modelos avaliados e para o conjunto de dados utilizado, pode-se concluir qual modelo possui menor risco médio associado de acordo com a tabela 10, que mostra o risco médio dos dois modelos quando treinados com 280 amostras.

Estes números confirmam os dados evidenciados na seção 4, em que foi observada alta precisão de classificação, o que confirma que esses modelos são adequados para a classificação de documentos, conforme indicado na literatura pesquisada. O PMC leva vantagem nos quesitos de precisão e agilidade de classificação, e a SVM

Tabela 10 – Risco médio por modelo e critério.

| **Critério** | **Valor PMC** | **Valor SVM** |
|---|---|---|
| Risco de Erro | 0.002603 | 0.012955 |
| Risco Temporal de Treinamento | 0.095946 | 0.044770 |
| Risco Temporal de Classificação | 0.306829 | 0.340366 |
| Risco de Indisponibilidade | 0.010908 | 0.000970 |

Fonte – Elaborado pelo autor

nos quesitos de agilidade de treinamento e disponibilidade, para o conjunto de dados avaliado. É importante ressaltar que o PMC mostrou consistentemente maior consumo de memória e maior tempo necessário para treinamento do que a SVM, o que pode ser uma restrição em situações em que os recursos computacionais são escassos. No entanto, cabe mencionar que em situações em que são necessárias mais do que duas classes, o consumo de memória observado pode ser favorável ao PMC devido à necessidade de se treinar mais de uma SVM, já que esta realiza apenas classificação binária por definição.

A linguagem R, por possuir vasta biblioteca de funções estatísticas e estruturas de dados nativas próprias para análise de dados, foi um fator acelerador no processo de desenvolvimento do protótipo. No entanto, uma dificuldade encontrada foi a lentidão no processamento de dados com grande quantidade de variáveis, que fez com que a execução dos testes levasse cerca de 11 horas, mesmo com o uso da biblioteca *parallel* (R Core Team, 2015), que possibilita o uso de vários núcleos de processamento e atualmente é parte da distribuição padrão da linguagem R, versão 3.3.1, de 21/06/2016. Este tempo possivelmente pode ser reduzido com ajustes finos no código atual.

Os testes foram executados em um computador *MacBook Pro* com processador Intel i7-2720QM de 4 núcleos e frequência de 2.2 GHz e 16 gigabytes de memória RAM, com o sistema operacional OS X versão 10.11, codinome “El Capitan”.

Uma outra dificuldade encontrada foi a criação do conjunto de dados para a realização dos testes. O autor partiu de um conjunto de cerca de 13.000 imagens de vários tipos distintos, com o objetivo de selecionar apenas 600 de dois tipos diferentes, 300 de cada tipo. Com o auxílio do algoritmo *k-means* foi possível realizar automaticamente uma pré-separação das imagens em dez tipos distintos, e posteriormente realizar a separação manual dos tipos desejados em poucos minutos.

Acredita-se que o método, ao cumprir os objetivos propostos, contribui para a popularização do uso de técnicas de classificação automatizadas nas empresas. Adicionalmente, acredita-se que a terminologia adotada por este método pode fazer o papel de ponte entre times técnicos e de negócio, por se utilizar de termos facilmente compreensíveis pelos últimos.

Como sugestão de continuidade deste trabalho, acredita-se ser possível generalizar o método proposto para permitir seu uso na comparação de quaisquer modelos de classificação em termos de risco, para qualquer situação em que uma análise de risco seja necessária. Acredita-se ainda ser possível a generalização do método proposto para qualquer modelo de aprendizado para o qual exista um parâmetro de precisão tangível, como é o caso em modelos de regressão. Em termos de aplicabilidade, com a disseminação de dispositivos portáteis como *smartphones* e equipamentos de automação residencial, entre outros, em que a capacidade computacional é restrita e em que se deseja colocar cada vez mais funcionalidade, tal generalização pode ser uma ferramenta útil para se determinar os modelos ótimos a serem utilizados.

# REFERÊNCIAS


ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010.

ANTHONY, G.; GREGG, H.; TSHILIDZI, M. Image classification using svms: One-against-one vs one-against-all. **CoRR**, vol abs/0711.2914, 2007.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **The Journal of Machine Learning Research**, JMLR.org, v. 3, p. 993–1022, 2003.

BORKO, H.; BERNICK, M. Automatic document classification. **Journal of the ACM (JACM)**, ACM, v. 10, n. 2, p. 151–162, 1963.

BRASIL. **Código Civil Brasileiro**. 2002. <http://www.planalto.gov.br/ccivil_03/leis/2002/l10406.htm>. Acesso em: 06 jun. 2015.

Business Week. **The Office of the Future**. 1975. <http://www.bloomberg.com/bw/stories/1975-06-30/the-office-of-the-futurebusinessweek-business-news-stock-market-and-financial-advice>. Acesso em: 06 ago. 2016, via Internet Archive.

CHAPELLE, O. Support vector machines for image classification. École Normale Superieure de Lyon, 1998.

CHEN, N.; BLOSTEIN, D. A survey of document image classification: Problem statement, classifier architecture and performance evaluation. **International Journal of Document Analysis and Recognition (IJDAR)**, Springer, v. 10, n. 1, p. 1–16, 2007.

DATAFOLHA. **Consulta de Cargos e Salários**. 2015. <http://datafolha1.folha.com.br/empregos/salarios>. Acesso em: 06 jun. 2015.

FENACON. **Guarda e Manutenção de Documentos Fiscais**. 2010. <http://www.fenacon.org.br/usuarios/arquivos%5Cpublicacoes%5CGUIA_DE_PRAZOS_FENACON_versao_2010.pdf>. Acesso em: 06 jun. 2015.

GACEB, D.; EGLIN, V.; LEBOURGEOIS, F. Classification of business documents for real-time application. **Journal of Real-Time Image Processing**, Springer Berlin Heidelberg, v. 9, n. 2, p. 329–345, 2011. ISSN 1861-8200.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing (3rd Edition)**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 013168728X.

HEATON, J. **Introduction to Neural Networks with Java**. [S.l.]: Heaton Research, Inc., 2008.

ImageMagick Studio LLC. **ImageMagick 7.0.1-6**. 2016. <http://www.imagemagick.org/Magick++>. Acesso em: 06 ago. 2016.

JAMES, G. et al. **An Introduction to Statistical Learning**. [S.l.]: Springer, 2013. v. 112.

JONES, J. An introduction to factor analysis of information risk (fair). **Norwich Journal of Information Assurance**, v. 2, n. 1, p. 67, 2006.

KARATZOGLOU, A. et al. kernlab – an S4 package for kernel methods in R. **Journal of Statistical Software**, v. 11, n. 9, p. 1–20, 2004.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

LYMAN, P.; VARIAN, H. R. **How Much Information?** 2003. <http://groups.ischool.berkeley.edu/archive/how-much-info-2003/execsum.htm#summary>. Acesso em: 11 jun. 2015.

MANEVITZ, L. M.; YOUSEF, M. One-class svms for document classification. **The Journal of Machine Learning Research**, JMLR.org, v. 2, p. 139–154, 2002.

MARINAI, S.; GORI, M.; SODA, G. Artificial neural networks for document analysis and recognition. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, v. 27, n. 1, p. 23–35, 2005.

MEYER, D. et al. **e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien**. 2015. <https://CRAN.R-project.org/package=e1071>. Acesso em: 10 mar. 2016.

MINSKY, M.; PAPERT, S. Perceptrons. MIT press, 1969.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. <https://www.R-project.org/>. Acesso em: 10 mar. 2016.

RIJSBERGEN, C. J. V. Foundation of evaluation. **Journal of Documentation**, MCB UP Ltd, v. 30, n. 4, p. 365–373, 1974.

SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. **Redes Neurais Artificiais para Engenharia e Ciências Aplicadas - Curso Prático**. São Paulo: Artliber, 2010. ISBN 9788588098534.

The Open Group. **Open Group Standard for Risk Analysis (O-RA)**. [S.l.], 2013. 50 p.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. New York: Springer, 1995. ISBN 9781475732641.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. <http://www.stats.ox.ac.uk/pub/MASS4>. ISBN 0-387-95457-0.

WESTERMAN, G.; HUNTER, R. **Developing a Common Language About IT Risk Management**. Cambridge: Massachusetts Institute of Technology (MIT) - Center for Information Systems Research (CISR), 2009.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. 2009. <http://ggplot2.org>. Acesso em: 06 ago. 2016.

WING, M. K. C. from J. et al. **caret: Classification and Regression Training**. 2016. <https://CRAN.R-project.org/package=caret>. Acesso em: 06 ago. 2016.