Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Uso de Técnicas de Estatística e Redes Neurais para Previsão de Incêndio em Áreas Florestais

Max Reinheimer dos Santos Denig

26 de janeiro de 2018

Max Reinheimer dos Santos Denig

Uso de Técnicas de Estatística e Redes Neurais para Previsão de Incêndio em Áreas Florestais

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Prof. Dr. João Henrique Ferreira Flores

Porto Alegre
Novembro de 2017

Max Reinheimer dos Santos Denig

Uso de Técnicas de Estatística e Redes Neurais para Previsão de Incêndio em Áreas Florestais

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):_____
Prof. Dr. João Henrique Ferreira Flores, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Guilherme Pumi, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Porto Alegre
Novembro de 2017

Resumo

Este trabalho busca introduzir modelos computacionais, como Redes Neurais Aritificais (RNAs), na tarefa de prever incêndios em regiões secas, como a do Parque Nacional de Montesinho, Portugal. Aliado a isso, são utilizadas métodos estatísticos, como análise exploratória de dados, para reduzir a dimensão de entrada dos modelos e, assim, encontrar as mais relevantes variáveis de entrada para o modelo. O banco de dados utilizado neste trabalho consiste de 13 variáveis, sendo área queimada a variável resposta. Tratamos o banco de dados de três maneiras diferentes: utilizando o banco normalmente, retirando os zeros da variável resposta e tratando a variável resposta como binária. Foram testados três modelos: Regressão Linear e Logarítmica, Máquina de Vetor de Suporte (SVM em inglês) e Rede Neural de Função de Base Radial (RBF em inglês). Foram abordadas três propostas diferentes de variáveis de entrada para os modelos: todas as variáveis de entrada possíveis, apenas as selecionadas pelo algoritmo *Stepwise Backward* e apenas as variáveis indicadoras dos trimestres. Dentre estas três propostas, a que leva apenas as variáveis selecionadas pelo algoritmo *Stepwise Backward* se saiu melhor. Todos os modelos atingiram níveis satisfatórios de precisão, com o modelo RBF mostrando alguns problemas ao tratar com o banco sem os zeros. No geral, o modelo SVM com as variáveis selecionadas pelo algoritmo *Stepwise Backward* apresentou os melhores resultados. Foi possível obter resultados semelhantes, senão melhores, aos alcançados por Cortez e Morais (2007) "*A Data Mining Approach to Predict Forest Fires using Meteorological Data*", tese de PhD, Departamento de Sistema de Informações, Universidade de Minho, Guimarães, Portugal.

Palavras-Chave: Incêndios Florestais, Portugal, Redes Neurais, Máquina de Vetor de Suporte, Função de Base Radial, Regressão Linear, Regressão Logística, Análise Exploratória de Dados.

Abstract

This work aims to introduce computational models, such as Aritifical Neural Networks (ANNs), in the task of predicting fires in dry regions, such as Montesinho National Park, Portugal. Allied to this, statistical methods, such as exploratory data analysis, are used to reduce the input dimension of the models and, thus, to find the most relevant input variables for the model. The database used in this work consists of 13 variables, with burned area as the response variable. We use three different approaches to treat the database: using the complete database, by removing all zeros from the response variable and by treating the response variable as binary. Three models were tested for each approach: Linear and Logistic Regression, Support Vector Machine (SVM) and Radial Base Function Network (RBF). Three different approaches of input variables for models were addressed: all the possible input variables, only the variables selected by the Stepwise Backward algorithm and only the dummy variables of the quarters. Of these three proposals, the one that takes only the variables selected by the algorithm Stepwise Backward performed better. All models reached satisfactory levels of accuracy, with the RBF model showing some problems when dealing with the database without the zeros. In general, the SVM model with the variables selected by the Stepwise Backward algorithm presented the best results. It was possible to obtain results similar, if not better, to those obtained by Cortez and Morais (2007) "*A Data Mining Approach to Predict Forest Fires using Meteorological Data*", PhD thesis, Department of Information Systems, University of Minho, Guimarães, Portugal.

Keywords: Forest Fires, Portugal, Neural Networks, Support Vector Machine, Radial Basis Function, Linear Regression, Logistic Regression, Exploratory Data Analysis.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Comentários Iniciais

Diferenciando-se de outros tipos de incêndio por sua grande extensão, velocidade e capacidade de devastação, o incêndio florestal trata-se da queima sem controle dos combustíveis vegetais disponíveis na região afetada. Talvez uma das catástrofes naturais de maior impacto em Portugal, os incêndios florestais vêm atingindo grandes extensões das regiões secas da Península Ibérica e, assim, preocupando autoridades locais devido à potencial ameaça à vidas, bens e ao meio ambiente.

Mais comum durante épocas de seca e calor, geralmente durante os verões, os incêndios florestais também dependem de fatores naturais para ocorrer. Sabendo disso, o presidente da Liga de Bombeiros Portugueses, Jaime Soares, relatou em 2015 que por volta de 75% dos incêndios florestais no país são de origem criminosa, atribuindo os restantes 25% a casos de negligência na reportagem do Jornal de Notícias em 2005. Alguns estudos dão suporte à afirmação de Soares, como, por exemplo, Lourenço (2011), onde é afirmado que mais de 90% das ignições dos incêndios em Portugal tem origem em atos humanos, negligentes ou intencionais.

Em Portugal o problema dos incêndios florestais está em pauta há anos, visto o prejuízo causado por esta catástrofe ao longo de sua história. Apesar de o governo português vir investindo em medidas de prevenção e combate a incêndios, ainda não foi possível conter o fogo e as áreas afetadas continuam a se estender por vários hectares anualmente. Como é possível observar na Figura 1.1, os anos de 2003 a 2005 foram especialmente prejudiciais para os portugueses, quando quase 900 mil hectares foram queimados ao longo destes três anos.

Com a crescente capacidade computacional disponível hoje em dia, é de interesse geral colocar em uso técnicas e algoritmos para tentar prever local, data, hora e extensão de incêndios com a maior precisão e confiabilidade possível. Os modelos de Redes Neurais Artificiais (RNAs) vêm acompanhando estes avanços na área de capacidade computacional, possibilitando seu uso para as mais diversas finalidades. Assim, já é uma realidade trabalhar com modelos mais complexos de forma mais prática, simples e eficiente, o que nos traz a possibilidade de trabalhar com problemas complexos de maneira que antes não era possível.

Redes Neurais Artificiais são modelos computacionais que foram concebidos tendo inspiração na forma de um ser humano reconhecer padrões e no funcionamento do cérebro ao realizar esta tarefa. Não é difícil encontrar trabalhos utilizando Redes Neurais para realizar previsões, visto que são modelos capazes de realiza-lás com razoável acuidade. Existem, no entanto, diversos tipos de Redes Neurais, as

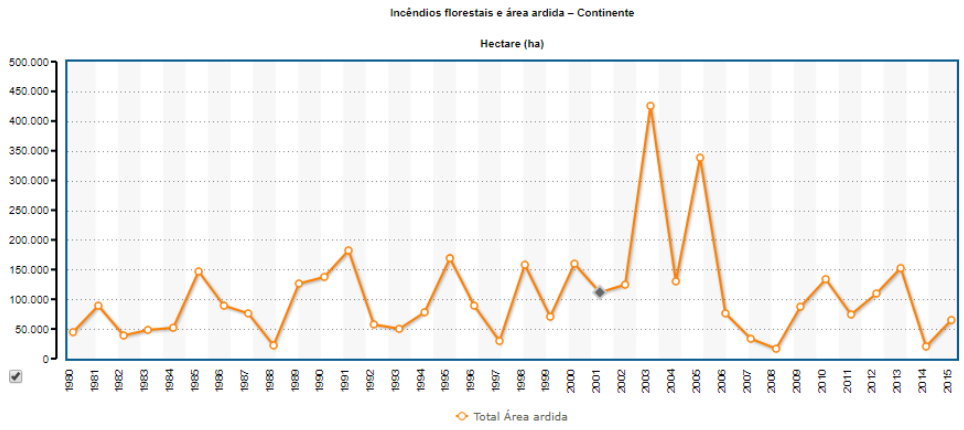

Figura 1.1: Área atingida por incêndios florestais em Portugal (em ha) anualmente no período de 1980 a 2015.
Fonte: de Dados Portugal Contenporâneo (2017).

quais podem ser classificadas em duas categorias: redes de aprendizado supervisionado e não-supervisionado.

Uma rede de aprendizado supervisionado recebe dados de entrada e de saída e, a partir disto, consegue aprender padrões, o que consiste na modificação dos pesos das conexões entre os neurônios de forma iterativa. A rede consegue fazer este aprendizado criando uma representação das informações no seu interior, o que isenta o usuário desta tarefa. Já uma rede de aprendizado não-supervisionado recebe apenas dados de entrada e tenta modificar os pesos de suas conexões sem os dados de saída correspondentes.

Mais especificamente, como citado em S. Haykin (2001) "*Redes Neurais: Princípios e prática*", o uso de redes neurais nos oferece as seguintes vantagens:

1. *Não-Linearidade.* Se uma Rede é formada por neurônios artificiais não-lineares, temos uma rede não-linear. Esta característica é bastante importante quando a relação entre os dados de entrada e os dados de saída, ou objetivos, é não linear.

2. *Mapeamento de Entrada-Saída.* Para redes de aprendizagem supervisionada, a fase de aprendizado consiste de apresentar para a rede um conjunto de dados de entrada e saída pareados chamados de *amostras de treinamento.* Com esse conjunto a rede modifica os seus pesos sinápticos de forma que minimize a diferença entre a resposta desejada e a resposta da rede. Assim, a rede aprende dos exemplos ao construir um *mapeamento de entrada-saída* para o problema considerado. Tal abordagem nos faz lembrar do estudo de *inferência estatística.*

3. *Adaptabilidade.* Uma rede treinada para um certo contexto pode facilmente adaptar seus pesos sinápticos para se acomodar a novos contextos, ou seja, ser retreinada para novas entradas.

4. *Resposta a Evidências.* Quando classificando padrões, uma rede pode fornecer informações não somente sobre qual padrão selecionar, mas também sobre a confiança na decisão tomada e, com essa informação, rejeitar padrões ambíguos.

5. *Informação Contextual.* O conhecimento é representado pela própria estrutura e estado de ativação de uma rede neural. Cada neurônio da rede é potencialmente afetado pela atividade de todos os outros neurônios na rede. Consequentemente, a informação contextual é tratada naturalmente pela rede neural.

6. *Implementação em Integração em Escala Muito Ampla (VLSI).* A natureza maciçamente paralela de uma rede neural a faz ser potencialmente rápida na computação de certas tarefas. Assim, uma rede neural pode ser implementada adequadamente utilizando tecnologia de integração em escala muito ampla, o que traz o benefício de capturar comportamentos bastante complexos de forma altamente hierárquica.

Vários trabalhos já integraram as ideias de Redes Neurais com previsão de incêndios florestais com o intuito de auxiliar na prevenção e redução de danos dos mesmos. Cortez e Morais (2007) "*A Data Mining Approach to Predict Forest Fires using Meteorological Data*", tese de PhD, Departamento de Sistema de Informações, Universidade de Minho, Guimarães, Portugal, é um exemplo disto. Neste trabalho foram comparados alguns modelos de Redes Neurais ao tentar prever incêndios florestais no Parque Nacional de Montesinho utilizando um banco de dados organizado durante o andamento do trabalho.

Há, também, uma vasta gama de ferramentas de análise estatística capazes de inferir informações a partir de um grande banco de dados, realizar previsões ou servirem de auxílio para outras técnicas. Temos, por exemplo, a regressão linear, que estima o valor esperado a partir de algumas variáveis de entrada utilizando uma equação linear. Existem, ainda, regressões não-lineares, que seguem o mesmo princípio, porém, a equação usada é não-linear, como a regressão logística, comumente utilizada para predizer variáveis categóricas. Pode-se utilizar, também, análise exploratória das variáveis independentes buscando encontrar um grupo otimizado de variáveis para serem incluídas no modelo.

1.2 Objetivos

1.2.1 Objetivo Principal

Este trabalho tem como objetivo principal conciliar o uso de ferramentas estatísticas com modelos computacionais, como de Redes Neurais Artificiais, para chegar a modelos capazes de prever a área afetada por incêndios florestais com acurácia e parcimônia similares ou superiores às obtidas por Cortez e Morais (2007) "*A Data Mining Approach to Predict Forest Fires using Meteorological Data*", tese de PhD, Departamento de Sistema de Informações, Universidade de Minho, Guimarães, Portugal. A introdução de métodos estatísticos, como o estudo de correlação e modelos simples de regressão, servirá para identificar variáveis importantes para o modelo e, assim, obter o conjunto de variáveis de entrada mais eficiente.

1.2.2 Objetivos Secundários

Os objetivos secundários deste trabalho incluem encontrar modelos capazes de predizer com acurácia a ocorrência e a área afetada por incêndios florestais, sendo

a área do incêndio abordada ao retirar os zeros da variável resposta, o evento do incêndio abordado pela variável resposta dicotomizada e as duas características sendo previstas utilizando o banco de dados original.

Também é um objetivo secundário constatar qual é o tipo de tratamento do banco de dados que gera melhores resultados para as suas finalidades, e, dentro destes, qual o melhor e mais parcimonioso modelo, dentre os propostos, para realizar tais predições.

1.3 Metodologia

Trabalhamos com a variável resposta, *area* (área queimada), de três maneiras diferentes. A primeira é usá-la normalmente, da mesma forma que aparece no banco de dados, a qual chamamos de **BDOrig** (Banco de Dados Original). A segunda forma é omitir as observações de áreas de grandeza irrelevante (menores do que 100 m^2), representadas por zero, para, então, criar um modelo, de forma a prever apenas a área do incêndio, e não sua ocorrência, chamada neste trabalho de **BDSZero** (Banco de Dados Sem Zeros). Já a terceira forma será dicotomizar a variável em 1, área maior do que zero, e área zero, 0, prevendo apenas a ocorrência de um incêndio potencialmente perigoso, a qual nomeamos neste trabalho **BDBin** (Banco de Dados Binário). O ponto de corte utilizado para fins de análises para o tratamento **BDBin** foi $0,5$.

Trabalhando com a variável resposta *area*, notou-se uma forte assimetria em sua distribuição, onde a maioria de suas observações demonstram uma área queimada muito pequena. Devido a esta característica, a aplicação de uma transformação logarítmica foi necessária na variável resposta, *area*. Frequentemente utilizada em situações como esta, a transformação logarítmica $y = log(x+1)$ ameniza o problema da concentração em volta de valores pequenos, como visto na Figura 1.2.

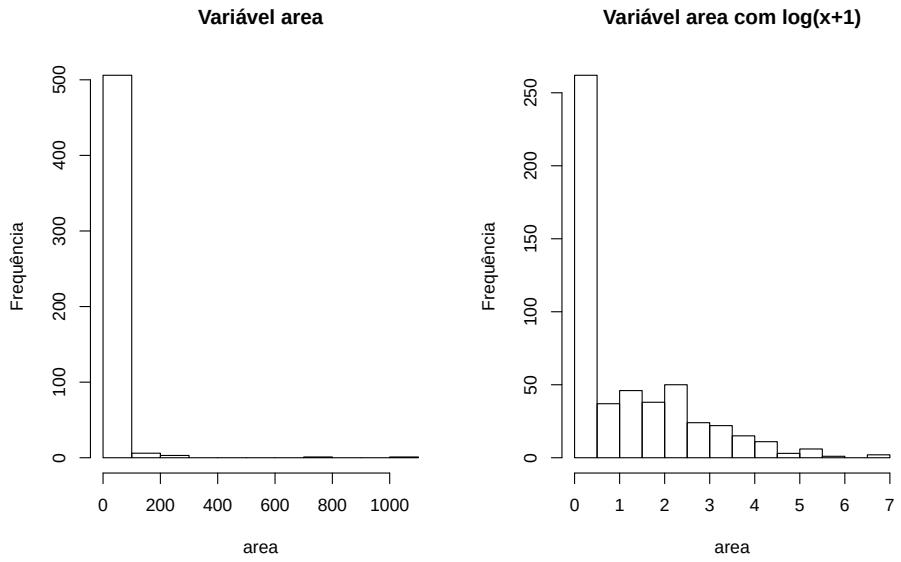

Figura 1.2: Variável *area* original e com a transformação logarítmica.

Para prever os incêndios, isto é, sua ocorrência e área atingida, usaremos três

técnicas para criar modelos capazes de tais predições: Regressão Linear, Regressão Logística, Máquina de Vetor de Suporte (SVM em inglês) e Função de Base Radial (RBF em inglês). Ao comparar os resultados, serão utilizados, para os casos **BDOrig** e **BDSZero**, a Raiz do Erro Quadrático Médio (RMSE em inglês) e o Desvio Absoluto Médio (MAD em inglês). Já para o tratamento do banco **BDBin**, utilizamos as medidas de Especificidade, Sensibilidade e Acurácia. Uma validação cruzada de 10 etapas proporcional em relação aos meses das observações foi utilizada para obter resultados mais generalizados e próximos dos valores reais dos Erros, RMSE e MAD.

Para realizar todos os testes, modelagens, cálculos, plotagem de gráficos e amostragens foi utilizado o ambiente de desenvolvimento integrado R R Core Team (2017) e uma variedade de pacotes disponíveis dentro desta ferramenta. Por sua praticidade, o RStudio RStudio Team (2015) foi a interface utilizada para o manuseio do R.

1.4 Delimitação

Este trabalho possui a ambição de trabalhar com dados referentes à incêndios florestais na área do Parque Nacional de Montesinho, ao nordeste de Portugal. Localizado na região nordeste de Portugal, o parque possui um clima supra-mediterrâneo, apresentando uma temperatura bastante variada, com concentrações em torno de 20 °C. Estes dados, disponibilizado em Cortez (2007b) , foram utilizados em vários trabalhos, inicialmente sendo coletados, organizados e analisados em (Cortez, 2007a). Temos exemplos de utilização deste banco de dados para análises em trabalhos como Kim (2009) e depois, ainda, em Shrivastava (2014).

Neste trabalho foram abordados quatro modelos: Regressão Linear, Regressão Logística, Máquina de Vetor de Suporte e Rede de Função de Base Radial. Os modelos de Regressão não serão aprofundados neste trabalho, dada sua notoriedade e vasta aplicação em vários campos, assim como não serão tratadas modificações estruturais nos modelos escolhidos. Os modelos serão utilizados, única e exclusivamente, de acordo com suas especificações, a serem tratadas no Capítulo 2, Seções 2.3 e 2.4.

2 Revisão Bibliográfica

2.1 Banco de Dados

O banco de dados utilizado neste trabalho é composto por 517 observações coletadas de Janeiro de 2000 a Dezembro de 2003. Duas fontes foram utilizadas para compor o banco: o inspetor responsável pelas ocorrências de incêndio em Montesinho, fornecendo informações de dia, mês, local e vegetação queimada no incêndio; e o Instituto Politécnico de Bragança, fornecendo informações meteorológicas coletadas por um centro meteorológico localizado no centro do parque de Montesinho.

A descrição das variáveis do banco de dados é vista a seguir, na Tabela 2.1.

Variável	Descrição
X	Coordenada do eixo X (de 1 a 9)
Y	Coordenada do eixo Y (de 1 a 9)
month	Mês do ano (de Janeiro a Dezembro)
day	Dia da semana (de Segunda a Domingo)
FFMC	Código de Boa Misutra de Combustível [1]
DMC	Código de Umidade de Duff
DC	Código de Seca
ISI	Índice de Propagação Inicial
temp	Temperatura (em °C)
RH	Umidade relativa do ar (em %)
wind	Velocidade do vento (em km/h)
rain	Volume de chuva (em mm/m^2)
area	Área queimada total (em ha)

Tabela 2.1: Descrição das Variáveis

As quatro primeiras variáveis são categóricas e se referem a características de espaço e tempo. Isto é, **X** e **Y** indicam o local do início do incêndio através de um *grid* de 8x8 da região, enquanto **month** e **day** indicam o mês e o dia da semana da observação. As quatro seguintes, **FFMC**, **DMC**, **DC** e **ISI**, são variáveis contínuas que representam quatro das seis componentes do *Fire Weather Index* (FWI), que é o sistema canadense para classificar risco de incêndio. Os três primeiros, **FFMC**, **DMC** e **DC** são componentes relativas à condição do solo e sua vegetação, ou seja, ao combustível do fogo. Já o **ISI** é a componente que relaciona estes fatores com a

[1] Tradução livre das descrições das variáveis **FFMC**, **DMC**, **DC** e **ISI**.

velocidade do vento. As outras duas componentes omitidas do *Fire Weather Index*, BUI e FWI, não foram incluídas no banco de dados, assim como feito por Cortez Cortez (2007a), pois são dependentes das outras quatro componentes. As quatro últimas variáveis são relacionadas ao clima. A variável contínua resposta, ***area***, nos dá a informação de quantos hectares foram queimadas pelo incêndio.

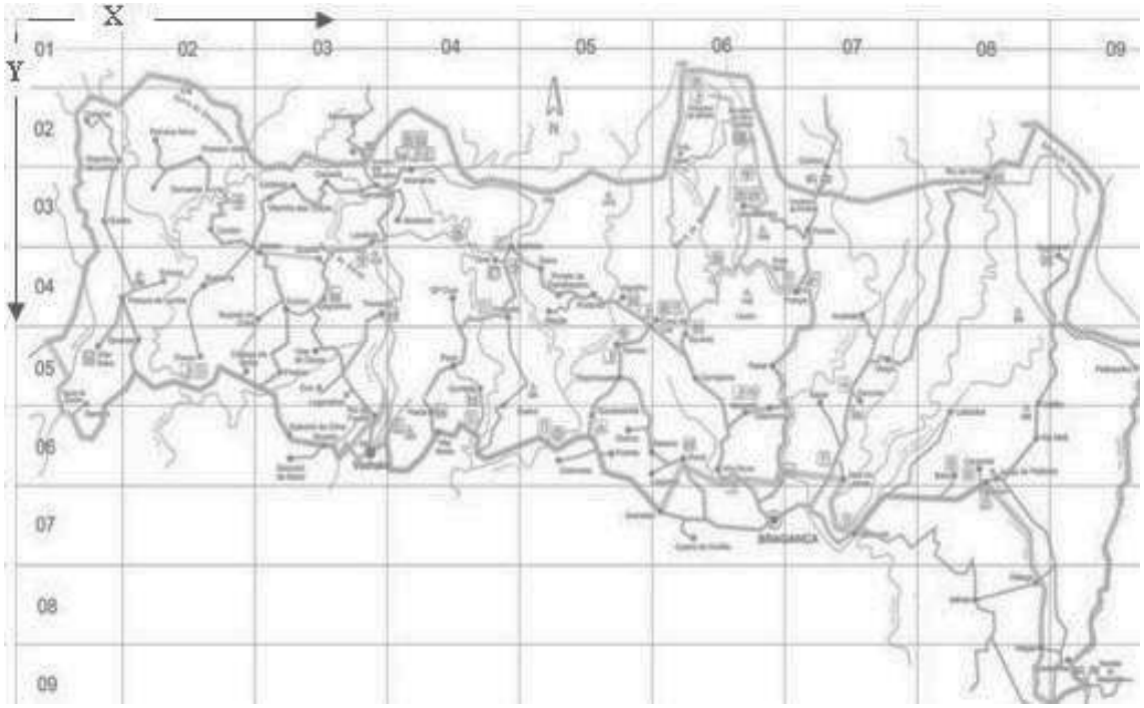

Figura 2.1: Mapa do Parque Nacional de Montesinho com o *grid* ilustrando as variáveis **X** e **Y**. Fonte: Cortez (2007a).

2.2 Análise de Regressão

A técnica de análise de regressão, amplamente conhecida não apenas na área da Estatística, como também em uma grande variedade de campos de estudo, consiste do estudo e análise da relação entre a variável dependente, a chamada variável resposta, e as variáveis independentes, também conhecidas como variáveis explicatórias. Utilizamos dois tipos de regressão neste trabalho: a Regressão Linear, para trabalhar com as variáveis resposta quantitativas; e a Regressão Logística, para trabalhar com a variável resposta binária.

Por não apresentarmos nada de novo quanto a utilização desta técnica, sendo inclusive utilizada como base de comparação e apoio a escolha de modelos, não nos aprofundaremos. Além disso, trata-se de modelos já bem conhecidos na literatura. Por estas razões, este tópico será apresentado apenas de forma superficial, de forma que a técnica de Regressão Linear, chamada assim por utilizar uma função linear para explicar a relação entre as variáveis, não será abordada.

2.2.1 Regressão Logística

Na Regressão Logística a variável resposta é comumente dicotômica, ou seja, apresenta o valor 1 para o acontecimento de um evento e 0 para a ausência do mesmo. É uma técnica que nos permite estimar a probabilidade associada à ocorrência de determinado evento ao modelar a relação entre a variável resposta, binária, e uma variedade de variáveis explicativas. Desta forma, um modelo de Regressão Logística adapta técnicas de Regressão Linear para determinar de uma superfície de separação entre duas classes, isto é, a ocorrência ou não de um evento.

Na Regressão Logística, costuma-se analisar dados distribuídos binomialmente, ou seja:

$$Y_i \sim Binom(p_i, n_i), \quad para \quad i = 1, \cdots m,$$

onde Y representa uma variável dependente que assume apenas dois valores e p_i as suas respectivas probabilidades.

Assim, seja $X_1, X_2, X_3, \cdots, X_k$ um conjunto de k variáveis independentes, o modelo de Regressão Logística pode ser escrito da seguinte forma:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}},$$

sendo a função $g : \mathbb{R} \to \mathbb{R}$ é:

$$g(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

onde β representa o vetor dos coeficientes $\{\beta_0, \beta_1, \cdots, \beta_k\}$, que são estimados a partir do conjunto de dados apresentado utilizando o método da máxima verossimilhança. Este método busca maximizar a probabilidade de a amostra ter sido observada ao escolher as combinações de β.

Dada um certo vetor β, ao variar os valores de X_i, observamos que a curva logística apresenta um comportamento probabilístico na forma de um S, característica que gera os seguintes resultados:

- Se $g(x) \to +\infty$, $P(Y = 1) \to 1$,
- Se $g(x) \to -\infty$, $P(Y = 1) \to 0$.

Assim, podemos estimar diretamente a probabilidade de ocorrência de um evento. Da mesma forma:

$$P(Y = 0) = 1 - P(Y = 1).$$

Quando utilizamos a Regressão Logística, a principal suposição é a de que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear, isto é:

$$\frac{P(Y = 1)}{P(Y = 0)} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k},$$

e, assim,

$$\ln\Big[\frac{P(Y = 1)}{P(Y = 0)}\Big] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

o que é conhecido como logito das probabilidades binomiais. Desta forma, é comum se interpretar e_i^{β} ao invés de β_i, dada o comportamento linear do logito.

Ao discriminar os grupos quando se está fazendo alguma análise utilizando um modelo logístico, é necessário utilizar uma regra de classificação, ou seja, um ponto de corte. É comum se utilizar:

$$\begin{cases} P(Y=1) > 0,5, \\ P(Y=0) \leq 0,5. \end{cases}$$

2.3 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM em inglês) é um método de aprendizado supervisionado de reconhecimento de padrões, utilizado para classificação e análise de regressão. Como qualquer método de aprendizado de máquina, a SVM aplica o princípio da indução, no qual obtém-se conclusões genéricas a partir de um conjunto particular de exemplos, chamado de conjunto de treino. Sendo uma rede de aprendizado supervisionado alimentada adiante (*feedforward*) universal, a SVM utiliza do conjunto de pares (saída-alvo) de entrada para criar uma representação, um modelo, capaz de produzir saídas aproximadas para um novo conjunto de entradas. Em sua essência, métodos de aprendizado de máquina são classificadores que utilizam dados de entrada para aprender um padrão separador e aplicá-lo a um novo conjunto de entrada.

Proposta por Vapnik Vapnik (1992) a SVM tem como idéia principal a construção de um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima, segundo Haykin Haykin (2001). Haykin ainda afirma que a máquina de vetor se suporte é uma implementação do método de minimização estrutural de risco, princípio indutivo baseado no fato de que a taxa de erro de uma SVM sobre dados de teste é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis. Em consequência disto, as máquinas de vetores de suporte possuem a característica única de apresentar um bom desempenho de generalização em problemas de classificação de padrões mesmo sem incorporar conhecimento do domínio do problema.

De extrema importância para o método da SVM, os vetores de suporte estão no cerne da construção do algoritmo de aprendizagem da máquina, estando presentes no núcleo do produto interno entre um vetor de suporte e um vetor retirado do espaço de entrada. Estes vetores de suporte são um subconjunto dos dados de treino extraídos pelo algoritmo. Dependendo de como este núcleo de produto interno foi gerado, pode-se construir diferentes máquinas de aprendizagem.

Em outras palavras, o que uma SVM faz é encontrar uma linha de separação, chamada hiperplano, entre dados de dois padrões. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classe. A SVM primeiro classifica as classes corretamente e depois em função dessa restrição define a distância entre as margens.

2.3.1 Máquina de Vetor de Suporte para Padrões Linearmente Separáveis

Considerando uma amostra de treino $(\mathbf{x_i}, d_i)_{i=1}^{N}$, onde $\mathbf{x_i}$ é o padrão de entrada para a *i-ésima* observação da amostra e d_i é a resposta correspondente (conjunto

saída-alvo), assumimos que a classe representada pelo subconjunto $d_i = +1$ e a classe representada pelo subconjunto $d_i = -1$ são linearmente separáveis, isto é, podemos utilizar uma reta para separar os padrões(classes). Esta reta , ou superfície de decisão, chamada *hiperplano*, que realiza tal separação é definida pela equação

$$\mathbf{w^T x_i} + b = 0, \tag{2.1}$$

onde $\mathbf{x} \in \mathbb{R}^p$ é um vetor de entrada, $\mathbf{w} \in \mathbb{R}^p$ é um vetor de peso ajustável, $\mathbf{w^T}$ é seu vetor transposto e $b \in \mathbb{R}$ é o viés (*bias*) associado. Assim, podemos dizer que

$$\begin{aligned} \mathbf{w^T x_i} + b \geq 0 \quad se \quad d_i = +1, \\ \mathbf{w^T x_i} + b < 0 \quad se \quad d_i = -1. \end{aligned}$$

Tendo um dado vetor peso $\mathbf{w}$ e *bias* b, a separação entre o hiperplano definido pela Equação (2.1) e o ponto dado mais próximo é definido como *margem de separação*. Temos como objetivo agora encontrar o hiperplano particular para qual a margem de separação é máxima. Este hiperplano particular é dito *hiperplano ótimo*, como ilustrado na Figura 2.2.

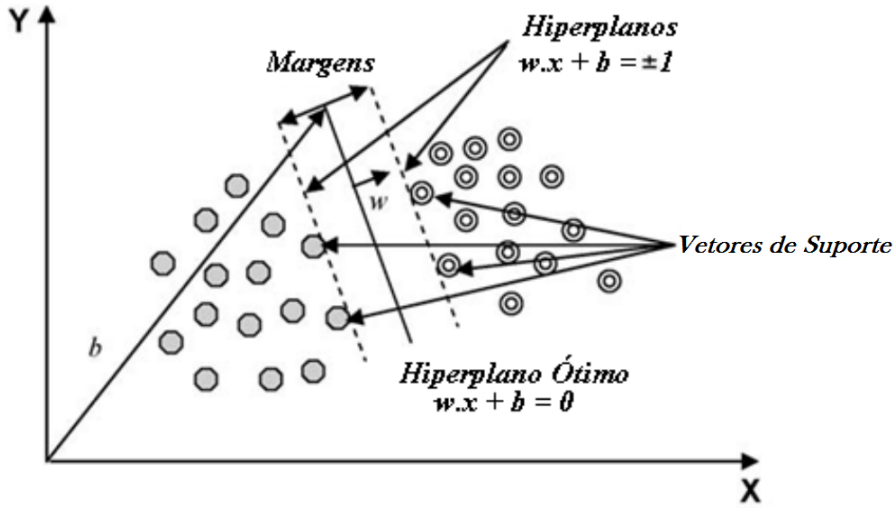

Figura 2.2: Representação gráfica do hiperplano ótimo e suas margens.
Fonte: Kavzoglu (2009).

Consideremos que $\mathbf{w_0}$ e b_0 representam valores ótimos do vetor peso e do *bias*. Assim, reescrevemos a Equação (2.1) para obter a definição do hiperplano ótimo

$$\mathbf{w_0^T x} + b_0 = 0.$$

Podemos descrever a distância de $\mathbf{x}$ até o hiperplano pela função discriminante $g : \mathbb{R}^k \to \mathbb{R}$

$$g(x) = \mathbf{w_0^T x} + b_0,$$

expressando $\mathbf{x}$ de maneira que

$$\mathbf{x} = \mathbf{x_p} + r\frac{\mathbf{w_0}}{||\mathbf{w_0}||},$$

onde $\mathbf{x_p}$ é a projeção normal de $\mathbf{x}$ sobre o hiperplano ótimo e r é a distância algébrica desejada, r vira positivo se x estiver no lado positivo do hiperplano ótimo e negativo se estiver no lado negativo. Dado, por definição, que

$$g(\mathbf{x_p}) = 0,$$

temos que

$$g(\mathbf{x}) = \mathbf{w_0^T x} + b_0 = r||\mathbf{w_0}||,$$

logo

$$r = \frac{g(x)}{||\mathbf{w_0}||}. \tag{2.2}$$

Como o objetivo é encontrar os parâmetros w_0 e b_0 para o hiperplano ótimo, devemos nos voltar para as restrições

$$\begin{aligned} \mathbf{w_0^T x} + b_0 \geq 0 \quad se \quad d_i = +1, \\ \mathbf{w_0^T x} + b_0 < 0 \quad se \quad d_i = -1, \end{aligned}$$

as quais serão sempre válidas se a Equação (2.1) for válida, ou seja, se os padrões forem linearmente separáveis. Assim, os pontos particulares $(\mathbf{x_i}, d_i)$ que satisfazem a primeira ou a segunda linha da restrição acima com sinal de igualdade são chamados de *vetores de suporte*, de onde vem o nome das máquinas de vetores de suporte. Estes vetores são os pontos dados que se encontram mais próximos da superfície de decisão e são, portanto, os de mais difícil classificação. Eles representam papel fundamental na localização do hiperplano ótimo. Considerando um vetor de suporte $\mathbf{x^{(s)}}$ associado a $d^{(s)} = +1$, temos

$$g(\mathbf{x^{(s)}}) = \mathbf{w_0^T x^{(s)}} \pm b_0 \pm 1 \quad se \quad d^{(s)} = \pm 1.$$

Aliando isto à Equação (2.2) vemos que

$$r = \frac{g(\mathbf{x^{(s)}})}{||\mathbf{w_0}||},$$

que nos leva a

$$\begin{aligned} \frac{1}{||\mathbf{w_0}||} \quad se \quad d^{(s)} = +1, \\ -\frac{1}{||\mathbf{w_0}||} \quad se \quad d^{(s)} = -1. \end{aligned}$$

Supondo que ρ represente o valor ótimo da margem de separação entre dois padrões que fazem parte do conjunto de treino, então

$$\rho = 2r = \frac{2}{||\mathbf{w_0}||}, \tag{2.3}$$

o que nos afirma que maximizar a margem de separação entre padrões igual a minimizar a norma euclidiana do vetor peso w. Podemos dizer, então, que o vetor peso w_0 fornece a máxima separação possível entre observações positivas e negativas.

2.3.2 SVM para Padrões Não Linearmente Separáveis

Segundo, Lorena Lorena (2007), "em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis. Isso se deve a diversos fatores, entre eles a presença de ruídos e outliers nos dados ou à própria natureza do problema, que pode ser não linear". Este problema acontece quando há pontos que se encontram dentro da região de separação, seja no lado correto ou não da superfície de decisão, o que pode comprometer a confiança do modelo SVM. Quando um ponto destes ocorre, dizemos que a margem de separação é suave, e a chamamos de rígida quando não ocorre.

Seja a amostra de treino $(\mathbf{x_i}, d_i)_{i=1}^{N}$ para $x_i \in \mathbb{R}^{m_0}$, vemos que na inequação

$$d_i(\mathbf{w^T x_i} + b) \geq 1 - \xi_i, \qquad i = 1, 2, \cdots, N, \tag{2.4}$$

é introduzido um novo conjunto de variáveis escalares não negativas $\{\xi_\mathbf{i}\}_{i=1}^{N}$, chamado de *variáveis soltas*, as quais medem o desvio de um ponto da condição ideal de separação de padrões. Seja $\boldsymbol{\xi}$ um vetor constituído por $\{\xi_1, \xi_2, \cdots, \xi_N\}$, então, se

1. $0 \leq \xi_i \leq 1$, o ponto $\mathbf{x_i}$ está dentro da região de separação, mas no lado correto da superfície de decisão;

2. $\xi_i > 1$, o ponto $\mathbf{x_i}$ está do lado errado da superfície de decisão.

Vetores de suporte se caracterizam como pontos que satisfazem a Equação (2.4), independentemente de ξ_i ser maior que zero. Nota-se que se uma observação apresentar $\xi_i > 0$ e for deixado de fora do conjunto de treino, o hiperplano não muda, ou seja, os vetores de suporte são definidos de maneira igual para casos linearmente separáveis e não separáveis.

Como um erro no conjunto de treino é indicado por $\xi_i > 1$, a soma dos ξ_i representa um limite no número de erros de treinamento Lorena (2007). Minimizamos, então, o erro sobre os dados de treino ao reformular a Equação (2.3) para definir uma função objetivo $\phi : \mathbb{R} \to \mathbb{R}$ dada por

$$\phi(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i,$$

onde se otimiza ϕ em relação a $\mathbf{w}$ e $\{\xi_{i=1}^{N}\}$, restrita à Equação (2.4).

A constante C é um termo que controla a complexidade da máquina e o número de pontos não-separáveis. Assim, esta constante é um parâmetro de regularização que impõe um peso à minimização dos erros no conjunto de treino em relação à minimização da complexidade do modelo. O parâmetro C é selecionado pelo usuário, seja experimentalmente, com uma espécie grosseira de reamostragem, ou analiticamente, estimando a dimensão de Vapnik-Chervonenki, pelo seguinte teorema que Haykin Haykin (2001) cita

Teorema 1. *Seja D o diâmetro da menor esfera contendo todos os vetores de entrada* $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}$ *de dimensão* m_0*. O conjunto de hiperplanos ótimos descrito pela equação*

$$\mathbf{w_0}^T\mathbf{x_i} + b_0 = 0$$

tem uma dimensão de Vapnik-Chervonenki, h, limitada acima por

$$h \leq \min\left\{\left\lceil \frac{D^2||\mathbf{w_0}||^2}{4} \right\rceil, m_0\right\} + 1$$

onde o sinal $\lceil . \rceil$ *representa o menor inteiro maior que ou igual ao número abrangido por ele e* m_0 *é a dimensionalidade do espaço de entrada.*

Vemos que o problema de otimização é quadrático. Sua solução envolve a introdução de uma função Lagrangiana e tornando suas derivadas parciais nulas. Tem-se como resultado o problema dual de encontrar os multiplicadores de Lagrange, α_i, variáveis não-negativas, que maximimizam a função objetivo

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x_i}^T \mathbf{x_j},$$

com as restrições

1. $\sum_{i=1}^{N} \alpha_i d_i = 0,$
2. $0 \leq \alpha_i \leq C, \quad \forall i = 1, 2, \cdots, N,$

onde C é um parâmetro especificado pelo usuário.

A solução ótima para o vetor peso $\precapprox$ é dada por

$$w_0 = \sum_{i=1}^{N_s} \alpha_{0,i} d_i \mathbf{x_i},$$

onde N_s é o número de vetores de suporte. Para determinar os valores ótimos de *bias* virá das condições de Kunh-Tucker, definidas por

$$\alpha_i[d_i(\mathbf{w}^T\mathbf{x_i} + b) - 1 + \xi_i] = 0, \quad i = 1, 2, \cdots, N, \tag{2.5}$$

$$\mu_i \xi_i = 0, \quad i = 1, 2, \cdots, N. \tag{2.6}$$

Na Equação (2.6) os μ_i representam multiplicadores de Lagrange que foram introduzidos para forçar a não-negatividade das variáveis soltas $\xi_i, \forall i$. A derivada da função lagrangiana no ponto de sela para o problema primordial em relação à variável solta ξ_i é zero, produzindo Haykin (2001)

$$\alpha_i + \mu_i = C. \tag{2.7}$$

Juntando as Eqs. (2.6) e (2.7), observa-se que

$$\xi_i = 0, \quad se \quad \alpha_i < C.$$

Para calcular o *bias* ótimo, toma-se qualquer ponto $(\curvearrowleft\!\!\sqsupset, d_i)$ do conjunto de treino para o qual $0 < \alpha_{0,i} < C$ e, assim, $\xi_i = 0$, e aplica-se este ponto na Equação (2.5) Haykin (2001).

2.4 Redes de Função de Base Radial

Segundo Haykin Haykin (2001), "aprender é equivalente a encontrar uma superfície, em um espaço multidimensional, que forneça o melhor ajuste para os dados de treinamento, com o critério de melhor ajuste sendo medido em sentido estatístico". Uma rede neural de Função de Base Radial (RBF em inglês) consiste em um modelo neural multicamadas alimentadas adiante (*feedforward)*, de aprendizado supervisionado capaz de analisar padrões complexos e resolver problemas não-linearmente separáveis. Uma rede RBF dá uma abordagem de problema de ajuste (aproximação) de curva às redes neurais.

A arquitetura de uma RBF é dividida em três camadas: a camada de entrada, que faz a interface dos padrões com a rede; a segunda camada realiza o mapeamento não-linear do espaço de entrada para o espaço oculto, o qual é, comumente, de alta dimensionalidade; e a camada de saída, que fornece a resposta da rede ao padrão apresentado. Segundo o teorema de Cover (Cover, 1965), "um problema complexo de classificação de padrões disposto não linearmente em um espaço de alta dimensionalidade tem maior probabilidade de ser linearmente separável do que em um espaço de baixa dimensionalidade". Essa é razão de ser frequente a alta dimensionalidade da camada oculta de uma rede RBF.

As funções de base radial, inicialmente utilizadas na solução de problemas de interpolação multivariada real, são funções sobre números reais cujos valores dependem apenas da distância a partir de algum ponto. As somas de funções de base radial, como, por exemplo, a função gaussiana, são tipicamente usadas para aproximar funções, o que possibilita seu uso como funções da camada oculta de redes RBF. Como já notado na seção anterior, o problema de classificação é relativamente simples uma vez que os padrões são linearmente separáveis. Dito isto, podemos abordar a operação de uma rede RBF como um classificador de padrões estudando a separabilidade de padrões.

2.4.1 Teorema de Cover sobre Separabilidade de Padrões

Suponhamos que χ represente um conjunto de N padrões, ou seja, vetores, $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}$ para $\mathbf{x_i} \in \mathbb{R}^{m_1}$ onde cada um deles pertence à uma classe χ_1 ou χ_2. Se existe uma superfície de uma família de superfícies que separe os pontos da classe χ_1 dos pontos da classe χ_2, então esta dicotomia é separável em relação à esta família de superfícies. Para cada $\mathbf{x} \in \chi$ definimos uma função $\{\phi_i(\mathbf{x}) | i = 1, 2, \cdots, m_1\}$ de forma que $\phi : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1}$, assim como

$$\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \cdots, \phi_{m_1}(\mathbf{x})]^T$$

Se existe um padrão $\mathbf{x}$ em um espaço de entrada de dimensão m_0, o vetor $\phi(\mathbf{x})$ mapeia os pontos no espaço de entrada de dimensão m_0 para outros pontos em um novo espaço de dimensão m_1. Neste caso, $\phi_i(x)$ é chamada uma *função oculta.* Correspondentemente, o espaço abrangido pelo conjunto de funções ocultas $\{\phi_i(\mathbf{x})\}_{i=1}^{m_1}$ é dito *espaço oculto* (Haykin, 2001).

Uma dicotomia $\{\chi_1, \chi_2\}$ de χ é dita separável por ϕ se existir um vetor $\mathbf{w} \in \mathbb{R}^{m_1}$ para o qual podemos escrever (Cover, 1965)

$$\mathbf{w^T}\phi(\mathbf{x}) > 0, \quad \mathbf{x} \in \chi_1,$$
$$\mathbf{w^T}\phi(\mathbf{x}) < 0, \quad \mathbf{x} \in \chi_2.$$

Com o hiperplano definido pela equação

$$\mathbf{w^T}\phi(\mathbf{x}) = 0,$$

que descreve a superfície de separação no espaço ϕ. Sua imagem inversa é definida por

$$\mathbf{x} : \mathbf{w^T}\phi(\mathbf{x}) = 0,$$

o que também define a superfície de separação no espaço de entrada.

Considerando-se uma classe de mapeamentos oriundos de uma combinação linear de produtos de r coordenadas vetoriais do padrão, as superfícies de separação associadas serão referidas como *variedades racionais de ordem r*. Uma variedade racional de ordem r em um espaço de dimensão m_0 é descrita por uma equação de grau r que envolve as coordenadas do vetor de entrada $\mathbf{x}$, de forma que

$$\sum_{0 \leq i_1 \leq i_2 \leq \cdots \leq i_r \ m_0} a_{i_1 i_2} x_{i_1} x_{i_2} \cdots x_{i_r} = 0 \tag{2.8}$$

onde x_0 é fixo em uma unidade para expressar a equação de forma homogênea e x_i corresponde à i-ésima componente do vetor de entrada $\mathbf{x}$ e $a_{i_1 i_2}$ representam variáveis auxiliares para $a_{i_1 i_2} \in \mathbb{R}$. Um produto de ordem r das componentes x_i de $\mathbf{x}$ é chamado de *monômio*. Para um espaço de entrada de dimensionalidade m_0 existem

$$\frac{(m_0 - r)!}{m_0! r!}$$

monômios na Equação (2.8). Esta equação, Equação (2.8), é capaz de descrever uma série de superfícies de separação, como, por exemplo, hiperplanos (variedades racionais de primeira ordem), quádricas (variedades racionais de segunda ordem) e hiperesferas (quádricas com algumas restrições para os coeficientes). De forma geral, a separabilidade linear implica a separabilidade esférica que, por sua vez, implica a separabilidade quádrica. O inverso, porém, não é válido.

Considere os padrões $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}$ independentemente escolhidos, de acordo com uma probabilidade imposta ao espaço de entrada. Suponha, agora, que todas as dicotomias possíveis $\chi = \{\mathbf{x_i}\}_{i=1}^{N}$ têm mesma probabilidade. Se $P(N, m_1)$ representa a probabilidade de um dicotomia particular escolhida aleatóriamente é separável por ϕ, onde a classe de superfícies de separação escolhida tem m_1 graus de liberdade, então, segundo Cover (1965), dizemos que

$$P(N, m_1) = \frac{1}{2} \sum_{m=0}^{m_1 - 1} \binom{N-1}{m} \tag{2.9}$$

Podemos observar que, em um ambiente probabilístico, a separabilidade de um conjunto de padrões pode ser considerado um evento aleatório que depende da dicotomia e da distribuição dos padrões no espaço de entrada. Vemos que na Equação (2.9) as superfícies das unidades ocultas consideradas estão em forma binomial e, assim, diferente das comumente utilizadas em redes RBD, porém, é válido notar que o

conteúdo essencial da equação tem aplicabilidade geral. Isto é, quanto mais alta for a dimensão de m_1 do espaço oculto, mais próximo da unidade será a probabilidade $P(N, m)$ Haykin (2001). Resumidamente, o teorema de Cover sobre a separabilidade de padrões engloba dois ingredientes básicos Cover (1965): a formulação não-linear da função oculta definida por $\phi_1(\mathbf{x}), \phi_1 : \mathbb{R}^{m_1} \rightarrow \mathbb{R}$, onde $\mathbf{x}$ é o vetor de entrada e $i = 1, 2, \cdots, m_1$; a alta dimensionalidade do espaço oculto comparado com o espaço de entrada, a qual é determinada pelo valor atribuído a m_1.

2.4.2 Teorema de Micchelli

O Teorema de Micchelli Micchelli (1986) diz que

Considere que $\{x_i\}_{i=1}^{N}$ seja um conjunto de pontos distintos em $\mathbb{R}^{m_0}$. Então, a matriz de interpolação $\mathbf{\Phi}$, $N \times N$, cujo o ji-ésimo elemento é $\phi_{ji} = \phi(||\mathbf{x_j} - \mathbf{x_i}||)$, não é singular.

Seja ϕ uma função de base radial, então, dentre as funções de base radial cobertas pelo teorema de Micchelli, as mais relevantes para redes RBF são:

- *Funções Gaussianas:*

$$\phi(r) = exp\Big(-\frac{r^2}{2\varphi^2}\Big), \quad para \quad \phi > 0 \quad e \quad r \in \mathbb{R} \tag{2.10}$$

- *Multiquádricas:*

$$\phi(r) = (r^2 + c^2)^{1/2}, \quad para \quad c > 0 \quad e \quad r \in \mathbb{R} \tag{2.11}$$

- *Multiquádricas Inversas:*

$$\phi(r) = \frac{1}{(r^2 + c^2)^{1/2}}, \quad para \quad c > 0 \quad e \quad r \in \mathbb{R} \tag{2.12}$$

Independente dos valores do tamanho N dos pontos de dados ou da dimensão m_0 dos vetores $\mathbf{x}$, tudo que é exigido para que estas funções de base radial listadas acima sejam não-singulares é que os pontos $\{x_i\}_{i=1}^{N}$ devem ser distintos. As funções multiquadráticas inversas e as gaussianas compartilham uma característica interessante: são funções localizadas, isto é, $\phi(r) \rightarrow 0$ quando $r \rightarrow \infty$. Por sua vez, as funções multiquadráticas são não-localizadas, pois $\phi(r)$ se torna ilimitada quando $r \rightarrow \infty$. É notável o fato de que uma matriz de interpolação $\mathbf{\Phi}$ baseada nas multiquadráticas é não-singular, e, assim, adequada para o uso de uma rede RBF.

3 Resultados

3.1 Introdução aos Resultados

Neste capítulo serão apresentados os resultados obtidos pelos modelos propostos, de forma que podemos compará-los não somente entre eles, mas avaliar suas performances contra os modelos do trabalho original Cortez (2007a). Serão explicitados neste capítulo, também, decisões tomadas ao longo do desenvolvimento do trabalho e as razões das mesmas, aliadas à resultados que lhes dão suporte.

Antes de testarmos tais modelos, foi necessário criar algumas variáveis indicadoras (*dummies*) para as variáveis de entrada que se classificavam em categorias, isto é, criar variáveis novas indicadoras para cada categoria de cada variável. Por exemplo, para a variável ***day***, foi necessário criar uma variável indicadora para cada um dos dias da semana (7 categorias) que assume o valor 1 se a observação (linha) aconteceu no dia da semana indicado pela indicadora e 0 caso contrário. O mesmo foi feito para a variável ***month***, para que, assim, pudéssemos melhor avaliar o impacto destas variáveis no modelo utilizando regressão linear.

Ao tentarmos pela primeira vez utilizar a técnica de validação cruzada de 10 etapas encontramos o problema da distribuição da variável ***month***. A variável se mostrou bastante assimétrica, como visto na Tabela 3.1.

Mês	Janeiro	Fevereiro	Março	Trimestre 1
Frequência	2	20	54	76
Mês	Abril	Maio	Junho	Trimestre 2
Frequência	9	2	17	28
Mês	Julho	Agosto	Setembro	Trimestre 3
Frequência	32	184	172	388
Mês	Outubro	Novembro	Dezembro	Trimestre 4
Frequência	15	1	9	25

Tabela 3.1: Distribuição da Variável ***month***.

Esta característica marcante da variável provavelmente advém do fato de que acontecem mais incêndios em meses mais quentes, durante o verão português, que começa em Junho e termina em Setembro. Porém, essa característica nos traz problemas na hora de realizar a validação cruzada de 10 etapas: cada uma das 10 etapas necessita possuir pelo menos uma observação de cada categoria das variáveis categóricas, ou seja, necessita apresentar pelo menos uma observação de cada mês.

Como vimos na Tabela 3.1, vários meses apresentam menos de 10 observações, o que inviabiliza a realização da validação cruzada com tantas etapas assim.

Para contornar este impedimento utilizamos exatamente o fato da distribuição da variável dos meses ser bastante concentrada no verão: agrupamos os meses em trimestres. Com a criação de quatro variáveis indicadoras para os trimestres, pudemos não apenas obter mais do que 10 observações para cada categoria, mas também simplificamos uma variável de acordo com sua distribuição, obtendo modelos mais parcimoniosos.

No entanto, dada a natureza aleatória da validação cruzada, ocorreram sorteios dos grupos para as 10 etapas em que um ou mais grupos ficaram sem nenhuma observação de algum trimestre. Necessitou-se, então, elaborar um algoritmo para gerar grupos com números proporcionais de observações de cada trimestre, isto é, cada um dos 10 grupos seria sorteado tendo a certeza de que o número de observações de cada trimestre seria proporcional ao número de observações de cada trimestre no banco de dados. Assim, realizamos uma validação cruzada de 10 etapas proporcional aos quatro trimestres.

Para comparação, variamos o tipo de tratamento do banco de dados. Utilizamos o **BDOrig**, o banco de dados original com o logaritmo natural aplicado à variável resposta acrescida de um, o **BDSZero**, o mesmo, porém, sem os zeros, e o **BDBin**, que é o caso binário, isto é, observações de área zero foram consideradas 0, e o resto, 1.

As métricas de erro avaliadas são, para o caso **BDBin**, Sensibilidade, Especificidade e Área Abaixo da Curva ROC, enquanto para os outros dois casos de tratamento da variável ***area***, a Raiz do Erro Quadrático Médio (RMSE em inglês) e o Desvio Absoluto Médio (MAD em inglês) das previsões. Vamos comparar, quando possível, os erros obtidos neste trabalho com os melhores erros gerados por Cortez (2007a), onde o melhor MAD foi obtido por um modelo SVM, enquanto o menor RMSE foi oriundo de um modelo *Naive*. O modelo *Naive* é utilizado em previsões e é comumente empregado na tarefa de comparação com outros modelos por apresentar a maior taxa de custo-efetividade. É um modelo simples em que as previsões são apenas o valor anterior da série. Para obtermos os resultados das métricas de erro de cada um dos modelos a seguir realizamos, assim como feito por Cortez (2007a), 30 repetições da validação cruzada de 10 etapas proporcional aos quatro trimestres.

3.2 Modelo Inicial

O modelo inicial foi criado utilizando todas as variáveis como entrada para cada um dos modelos propostos. Isto engloba a retirada da variável ***month*** e a inclusão das variáveis indicadoras dos trimestres. Sejam $\delta(X)_i$ e $\delta(Y)_i$, para $i = 1, ..., 9$, as variáveis indicadoras que representam, respectivamente, os níveis das variáveis categóricas X e Y. Assim, o modelo é:

$$
\begin{aligned}
area = & \beta_0 + \sum_{i=1}^{9} \beta_i \delta(X)_i + \sum_{j=1}^{9} \beta_{9+j} \delta(Y)_j + \beta_{19} tri1 + \beta_{20} tri2 + \beta_{21} tri3 + \\
& \beta_{22} segunda + \beta_{23} terca + \beta_{24} quarta + \beta_{25} quinta + \beta_{26} sabado + \\
& \beta_{27} FFMC + \beta_{28} DMC + \beta_{29} DMC + \beta_{30} DC + \beta_{31} ISI + \beta_{32} temp + \\
& \beta_{33} RH + \beta_{34} wind + \beta_{35} rain
\end{aligned}
$$

Nota-se que a variável indicadora do quarto trimestre, referente ao inverno português, ou seja, com menor número de ocorrência de incêndios de tamanho relevante, assim como a do dia de sexta-feira foram incluídos na média para evitar problemas de colinearidade. Em seguida observamos os resultados gerados por este modelo.

3.2.1 Tratamento BDOrig

Os resultados referentes aos modelos utilizando o tratamento do banco de dados **BDOrig**, juntamente com os melhores resultados obtidos no trabalho original, seguem na Tabela 3.2:

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	13,2918	13,0512	13,1085	12,86 (SVM)
RMSE	50,1691	50,4768	50,4323	63,70 (*Naive*)

Tabela 3.2: Resultados do Modelo Inicial - **BDOrig**.

Nota-se que, em termos do MAD, tanto o modelo SVM quanto o RBF apresentaram menores erros em relação ao modelo de Regressão. O menor MAD obtido, do modelo SVM, mostrou uma melhora de apenas pouco menor de 2% em relação ao modelo de Regressão. No caso do RMSE, ambos os modelos SVM e RBF mostraram um desempenho pior do que o de Regressão, apesar de que as diferenças estejam abaixo da ordem de 1%, ou seja, insignificantes.

Em relação aos melhores resultados obtidos no trabalho original, observamos que o nosso modelo SVM apresenta um valor de MAD um pouco maior, diferença de menos do que 2%. O RMSE, no entanto, mostra que todos os três modelos propostos se mostram bastante melhores, chegando a obter uma melhoria no RMSE de mais de 20%.

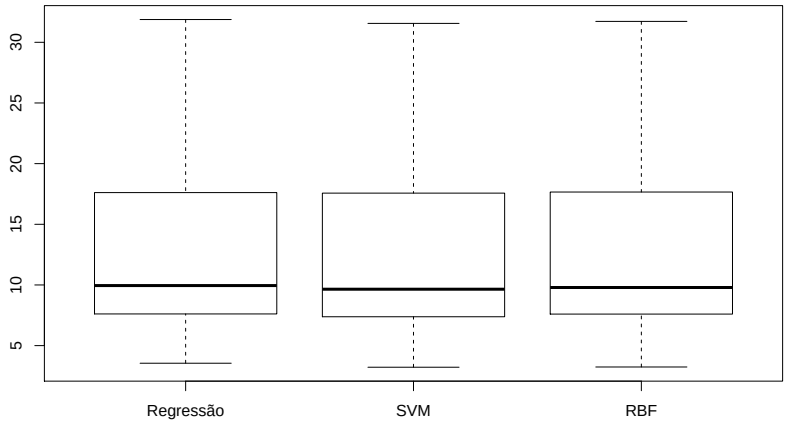

Figura 3.1: *Boxplots* da métrica de erro MAD obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

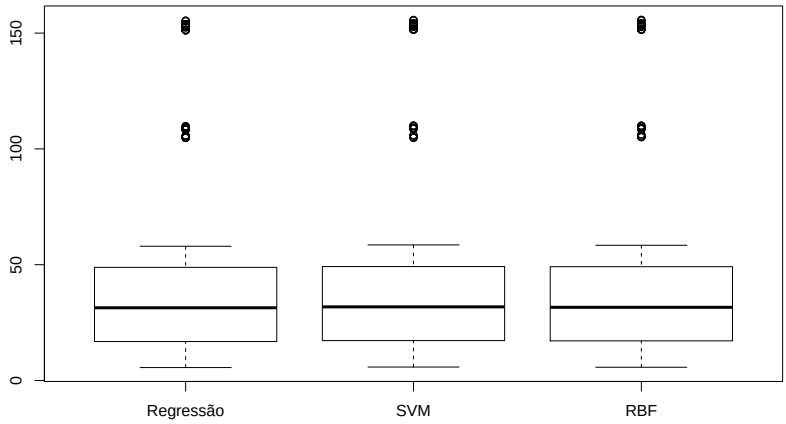

Figura 3.2: *Boxplots* da métrica de erro RMSE obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

Podemos observar através das Figuras 3.1 e 3.2 a similaridade dos três modelos ao verificar os seus erros utilizando o tratamento do banco **BDOrig**. Não é possível encontrar um destaque claro nos *boxplots*.

3.2.2 Tratamento BDSZero

Na tabela a seguir vemos as métricas de erro dos modelos ao utilizar o tratamento do banco de dados **BDSZero**, e, também, os melhores resultados obtidos no trabalho original:

Vemos que, tanto para o MAD quanto para o RMSE, a regressão apresentou o melhor resultado, apesar de a diferença ser bastante pequena em ambos os casos. Destaca-se o desempenho ruim do modelo RBF neste tratamento, chegando o modelo

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	1,0627	1,0760	1,2493	12,86 (SVM)
RMSE	1,3781	1,3971	1,6754	63,70 (*Naive*)

Tabela 3.3: Resultados do Modelo Inicial - **BDSZero**.

de Regressão a apresentar um RMSE até 17% melhor do que o RBF. Não podemos comparar estes resultados com os obtidos no trabalho original pois utilizamos um banco de dados diferente do utilizado no artigo, devido à remoção de valores.

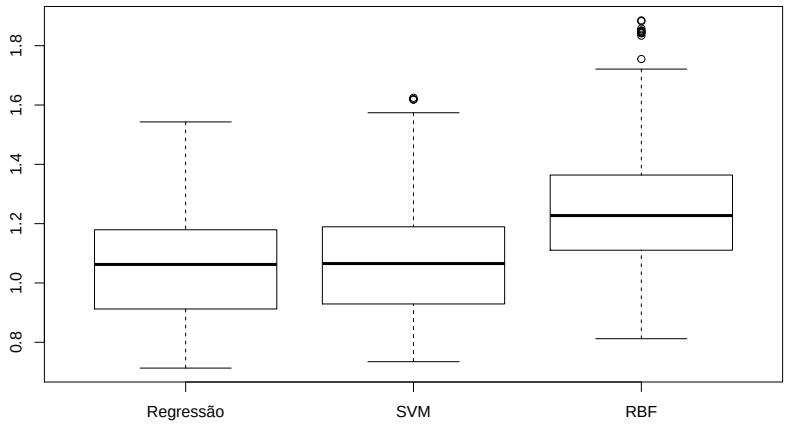

Figura 3.3: *Boxplots* da métrica de erro MAD obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

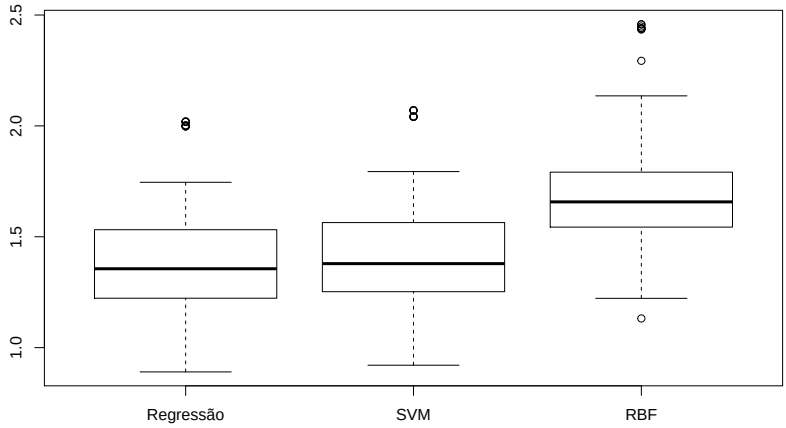

Figura 3.4: *Boxplots* da métrica de erro RMSE obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

Os gráficos nas Figuras 3.3 e 3.4 ilustram o mal desempenho do modelo RBF no tratamento do banco **BDSZero**, enquanto os modelos de Regressão e SVM se mos-

tram bastante similares, ainda que o modelo de Regressão se mostre marginalmente mais apto neste caso.

3.2.3 Tratamento BDBin

Apresentamos, agora, os resultados obtidos pelos três modelos utilizando o tratamento do banco de dados **BDBin**:

Métrica de Erro	Regressão	SVM	RBF
Sensibilidade	0,6427	0,7379	0,8976
Especificidade	0,3903	0,3644	0,1717
Área Abaixo da Curva ROC	0,5165	0,5511	0,5221

Tabela 3.4: Resultados do Modelo Inicial - **BDBin**.

O modelo RBF apresenta uma melhora na Sensibilidade de quase 40% em relação ao modelo de Regressão, enquanto para a Especificidade, o modelo de Regressão ainda apresenta o melhor resultado, chegando a ter uma melhora de mais de 125% em relação ao modelo RBF. No meio termo se encontra o modelo SVM, apresentando, tanto para Sensibilidade e Especificidade, valores razoáveis, e ainda mostrando a maior Área Abaixo da Curva ROC.

No geral, o modelo SVM se portou melhor no caso **BDBin** com o modelo original. Não há comparação com o trabalho de Cortez e Morais, visto que eles não trataram a variável resposta de forma binária.

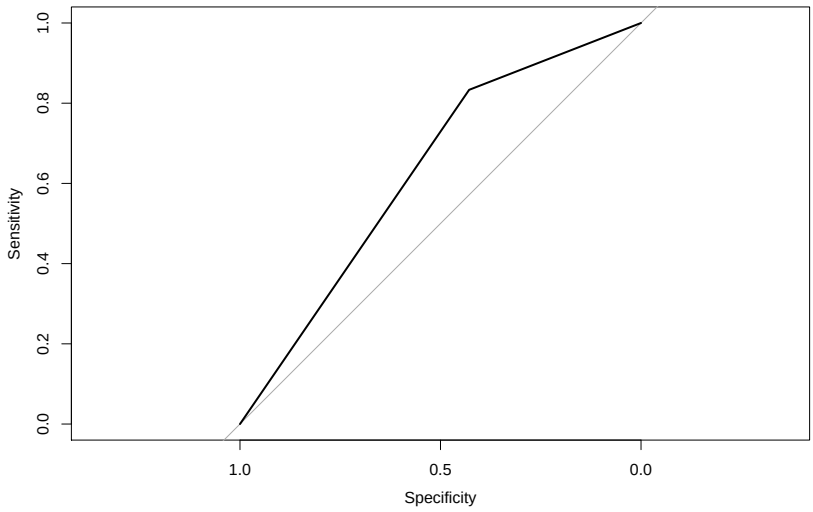

Figura 3.5: Curva ROC do modelo de Regressão.

Vemos na Figura 3.5 a melhor das três curvas obtidas pelos modelos, gerada pelo modelo de Regressão. Observa-se que a Sensibilidade está razoável, porém, a Especificidade deixa bastante a desejar. Isto influencia nosso modelo no sentido de que a sua capacidade de predizer corretamente a não ocorrência de um incêndio está prejudicada.

3.3 Redução da Dimensão de Entrada do Modelo

Existem várias técnicas de redução da dimensão de entrada do modelo, o que, em geral, resulta em melhora na acurácia do modelo, além de gerar um modelo mais parcimonioso. No entanto, vários métodos, como, por exemplo, os da área de Análise Multivariada, são indicadas para quando se trabalha com um banco de dados com muitas variáveis, na ordem das dezenas. Assim, dado o número relativamente pequeno de variáveis no banco de dados que estamos utilizando, uma análise exploratória acaba por se mostrar mais eficaz.

3.3.1 Análise Preliminar das Variáveis

Iniciamos a análise observando o comportamento das variáveis de entrada com a variável resposta através de um gráfico de dispersão aliado da linha de Regressão deste variável com a variável resposta, ***area***. Notou-se, inicialmente, que a distribuição das variáveis **ISI** e **FFMC**, pareciam se concentrar em direções opostas uma a outra, como visto na Figura 3.6.

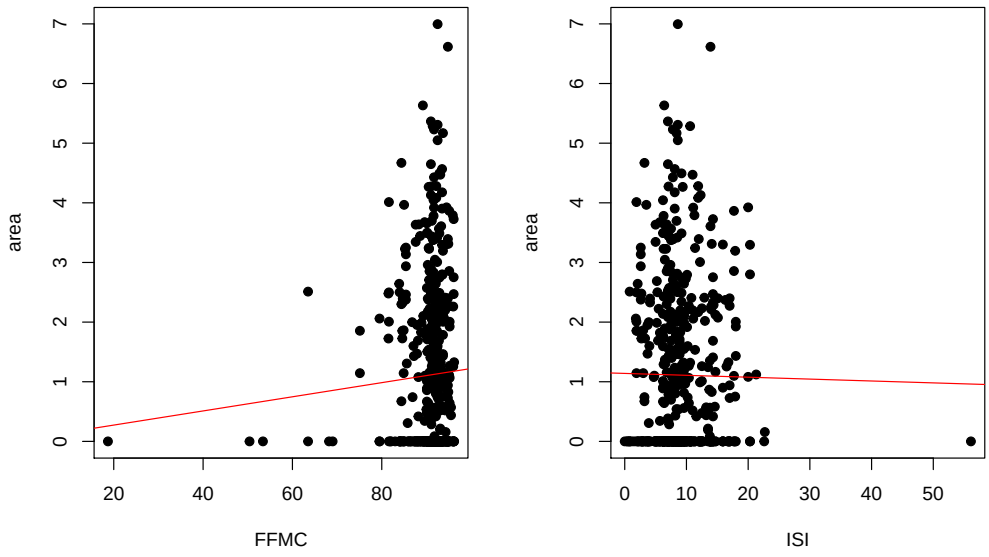

Figura 3.6: Comportamento das Variáveis **FFMC** e **ISI** com ***area***.

Esse comportamento levantou a possibilidade de se juntar estas duas variáveis em um uma através do produto das mesmas. Criamos, assim, a variável **ISIFFMC**, que é o produto das duas variáveis, **FFMC** e **ISI**. Observamos na Figura 3.7, no entanto, que a nova variável não aparenta apresentar nenhum padrão de comportamento linear com a variável ***area***.

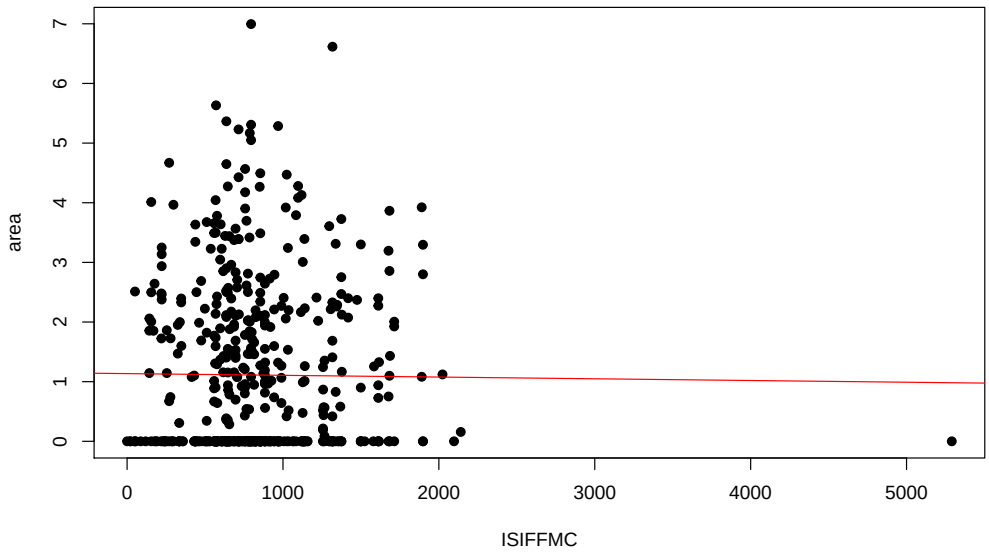

Figura 3.7: Comportamento da Variável **ISIFFMC** com ***area***.

Em seguida, para ampliar a variabilidade das variáveis **X** e **Y**, que apresentam apenas valores discretos de 1 a 9, criamos a variável **XY**, que foi gerada a partir da concatenação das duas variáveis. Isto é, para um conjunto $(X, Y) = (5, 3)$ temos a variável $XY = 53$. Assim, temos 81 combinações que resultam em 81 níveis para a variável categórica XY.

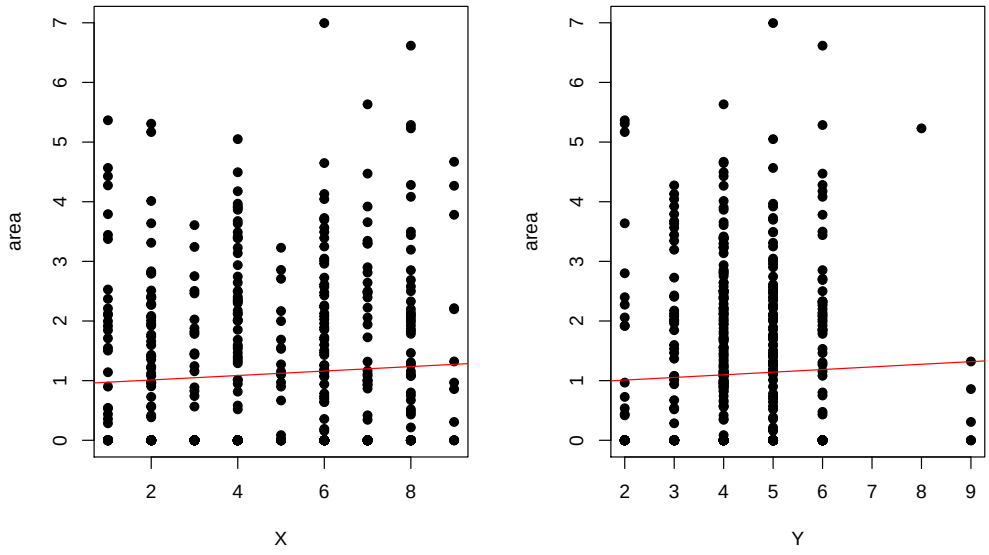

Figura 3.8: Comportamento das Variáveis **X** e **Y** com ***area***.

Vemos na Figura 3.9 que, apesar de melhorar o problema da variabilidade, a variável recém criada não solucionou totalmente este problema. A variável **XY**

melhora levemente a correlação com a variável resposta em relação a correlação da variável **X**, aumentando de 0.06199 para 0.06221, o que é insignificante além de, ainda, ser uma correlação bastante baixa.

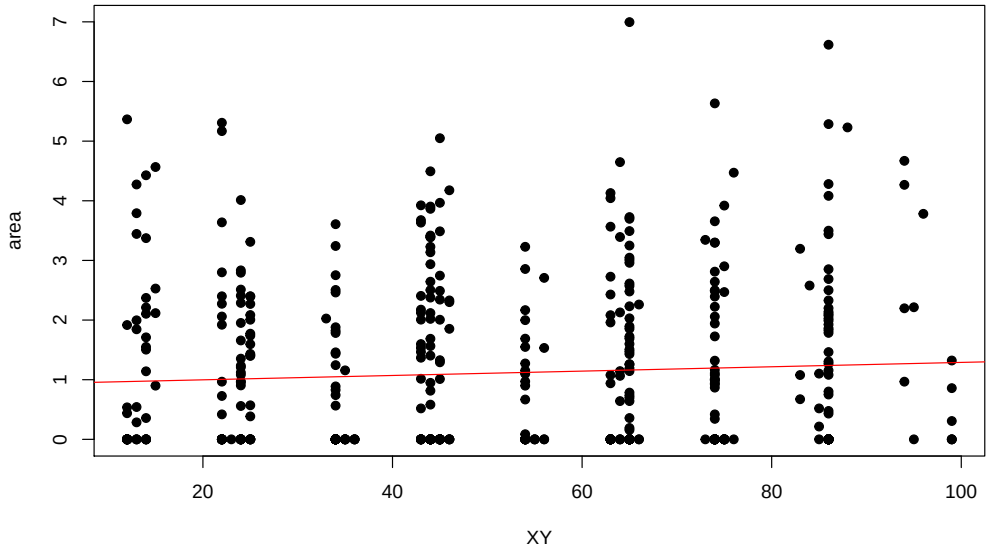

Figura 3.9: Comportamento da Variável **XY** com ***area***.

Observamos, também, o comportamento da variável **RH**, que se mostrou potencial de apresentar relação com a variável resposta. Para encontrar essa relação, tentamos criar a variável inversa dela, **RH_Inv**.

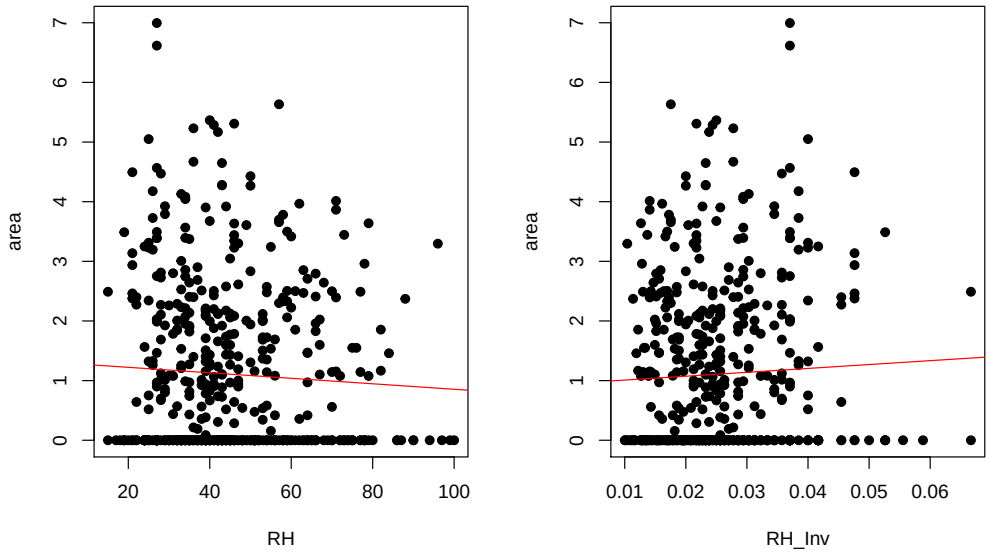

Figura 3.10: Comportamento das Variáveis **RH** e **RH_Inv** com ***area***.

Como pode-se ver na Figura 3.10, não há grandes mudanças no comportamento desta variável.

3.3.2 Análise de Regressão Linear Stepwise

Com estas variáveis novas em mãos, testamos um modelo de Regressão Linear utilizando um algoritmo *Stepwise Backward*, que seleciona o modelo a ser utilizado através do Critério de Informação de Akaike (AIC em inglês) Akaike (1987). O algoritmo calcula o AIC do modelo com todos as variáveis de entrada, e, uma por uma, retira cada variável de entrada e avalia se o modelo produziu um valor de AIC menor. Caso a redução no critério esteja presente, o algoritmo exclui definitivamente do modelo esta variável. Caso o AIC aumente, a variável fica no modelo, pois a retirada dela prejudica o desempenho do modelo segundo o Critério de Informação de Akaike. Este processo se repete até o algoritmo chegar a um modelo em que a remoção de nenhuma variável presente resulta em um menor AIC do modelo.

Seja o $\eta(XY)_i$ a variável indicadora que representa os 81 níveis da variável categórica XY. Assim, tendo o nível 81 da variável XY como base, o modelo que o algoritmo obteve ao final foi:

$$area = \beta_0 + \sum_{i=1}^{80} \beta_i \eta(XY)_i + \beta_{82} tri1 + \beta_{83} RH + \beta_{84} wind$$

O que revela que a criação das variáveis dos trimestres e da concatenação XY surtiram efeito, enquanto as novas variáveis **RH_Inv** e **ISIFFMC** não foram importantes para o modelo.

3.4 Modelo Reduzido

Testamos, então, o modelo indicado pelo algoritmo *Stepwise Backward* para procurar um modelo mais parcimonioso, ao mesmo tempo em que buscamos maior acurácia nas previsões. O modelo testado nesta Seção é composto das variáveis de entrada **XY**, **tri1**, **RH** e ***wind***, isto é, como explanado na Subseção 3.3.2:

$$area = \beta_0 + \sum_{i=1}^{80} \beta_i \eta(XY)_i + \beta_{82} tri1 + \beta_{83} RH + \beta_{84} wind$$

Em seguida veremos os resultados obtidos em cada um dos casos de tratamento do banco de dados.

3.4.1 Tratamento BDOrig

Na Tabela 3.5 podemos comparar os resultados dos três modelos propostos utilizando o tratamento do banco de dados **BDOrig**, além dos melhores resultados obtidos no trabalho original:

O modelo SVM, que apresenta o melhor MAD dentre os modelos propostos nesta situação, se mostra quase 1,5% maior do que o do trabalho original. Vemos mais uma vez os valores de RMSE dos três modelos propostos neste trabalho menores do

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	13,2322	13,0473	13,1175	12,86 (SVM)
RMSE	50,1632	50,4387	50,4245	63,70 (*Naive*)

Tabela 3.5: Resultados do Modelo Reduzido - **BDOrig**.

que o obtido por Cortez (2007a), sendo o do modelo de Regressão o menor de todos. O modelo SVM, apesar do menor MAD, apresentou o maior valor de RMSE dos três, enquanto o menor foi obtido pelo modelo de Regressão. Dito isto, as diferenças não passam de 2% entre os erros dos três modelos.

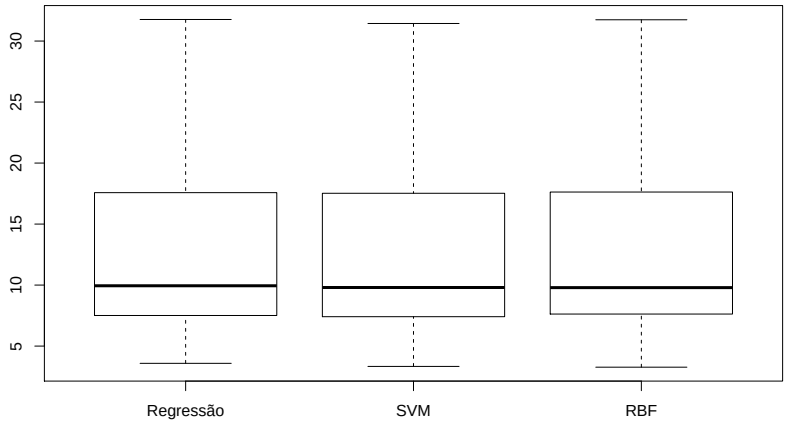

Figura 3.11: *Boxplots* da métrica de erro MAD obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

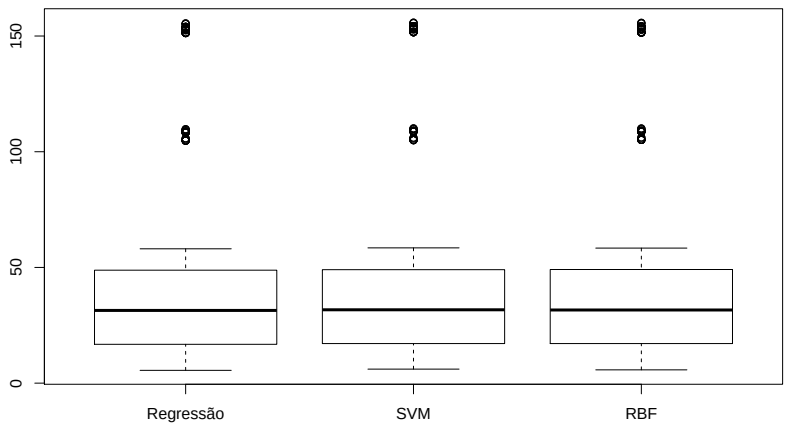

Figura 3.12: *Boxplots* da métrica de erro RMSE obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

Novamente as Figuras 3.11 e 3.12 não revelam grandes diferenças entre os erros MAD e RMSE dos três modelos.

3.4.2 Tratamento BDSZero

Na Tabela 3.6 estão as métricas de erro utilizando o tratamento do banco de dados **BDSZero**, juntamente com os melhores resultados obtidos no trabalho original:

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	1,0538	1,0464	1,2496	12,86 (SVM)
RMSE	1,3606	1,3430	1,6696	63,70 (*Naive*)

Tabela 3.6: Resultados do Modelo Reduzido - **BDSZero**.

Novamente vemos o modelo RBF não gerar bons resultados com o tratamento do banco de dados **BDSZero** em relação aos modelos SVM e de Regressão, que se portam melhores sem os zeros na variável resposta, chegando a apresentar um MAD 18% maior do que o modelo de Regressão. Apesar de mostrar erros similares ao modelo de Regressão, o modelo SVM foi o mais preciso neste caso. Mais uma vez não compararemos estes valores dos erros com os produzidos no trabalho original, pois utilizamos um bancos de dados diferentes, devido à remoção de valores.

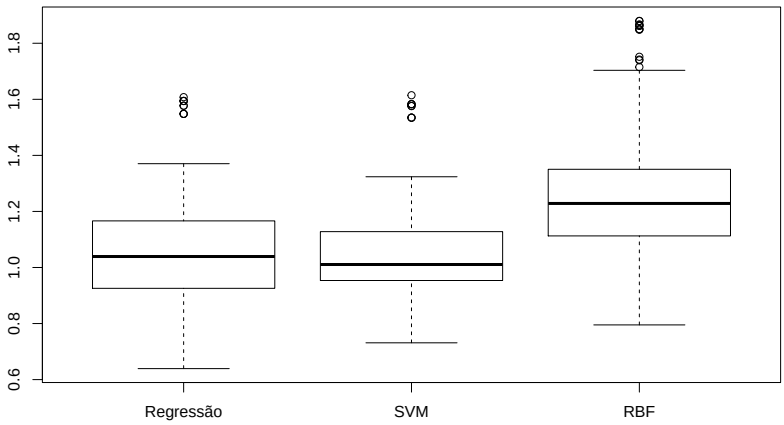

Figura 3.13: *Boxplots* da métrica de erro MAD obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

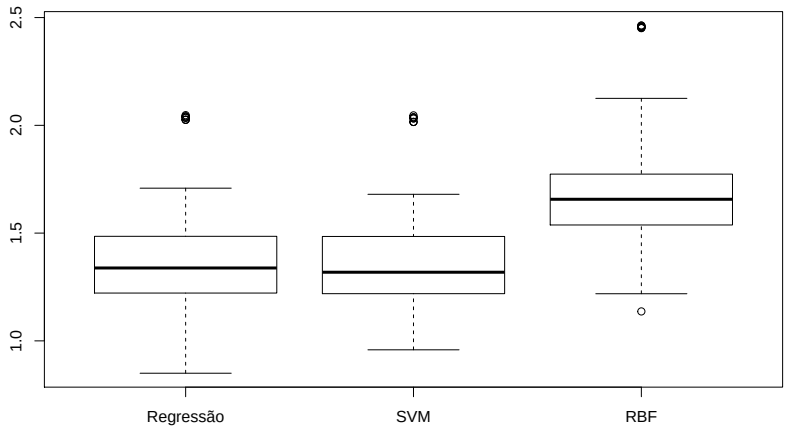

Figura 3.14: *Boxplots* da métrica de erro RMSE obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

Como ilustrado pelas Figuras 3.13 e 3.14, é confirmada mais uma vez a perda de desempenho do modelo RBF ao tratar com a variável ***area*** sem os zeros. É possível observar também erros levemente menores do modelo SVM em relação ao de Regressão.

3.4.3 Tratamento BDBin

Vemos a seguir os resultados obtidos utilizando o tratamento do banco de dados **BDBin**:

Métrica de Erro	Regressão	SVM	RBF
Sensibilidade	0,7040	0,8529	0,5774
Especificidade	0,3447	0,2798	0,4959
Área Abaixo da Curva ROC	0,5244	0,5596	0,5367

Tabela 3.7: Resultados do Modelo Reduzido - **BDBin**.

Aqui o modelo SVM se destaca em relação à Sensibilidade e Área Abaixo da Curva ROC, enquanto o modelo RBF mostra uma Especificidade quase 80% maior do que a do modelo SVM. O modelo RBF, apesar de mostrar bom desempenho na questão da Especificidade comparado aos outros modelos, nas outras métricas não se mostra tão eficaz, chegando a apresentar uma Sensibilidade mais de 30% menor do que o modelo SVM

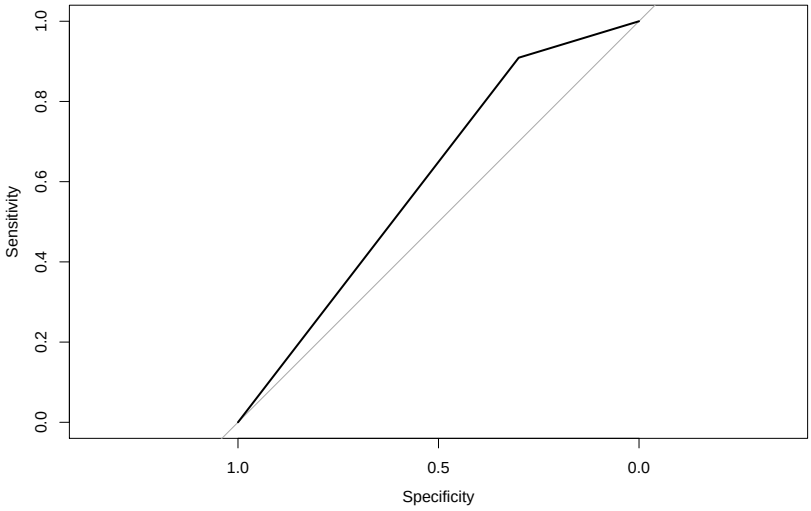

Figura 3.15: Curva ROC.

Novamente a melhor curva foi obtida pelo modelo de Regressão que, apesar de apresentar o melhor desempenho, não se mostra tão diferentes dos outros modelos. A situação é muito parecida com a do modelo Inicial, alcançando altos valores de Sensibilidade e baixos de Especificidade.

3.5 Modelo Trimestres

Procurando chegar a um modelo mais parcimonioso ainda, observamos as variáveis de entrada presentes no modelo anterior e o conjunto de variáveis relativas aos trimestres parecem ser a melhor escolha, dada sua natureza temporal e distribuição bastante assimétrica, com concentrações no terceiro trimestre. Assim, decidimos testar um modelo apenas com as variáveis indicadoras dos trimestres, representado por:

$$area \sim tri1 + tri2 + tri3$$

Tendo, assim, o quarto trimestre introduzido na média. A seguir veremos os resultados obtidos por este modelo.

3.5.1 Tratamento BDOrig

Os resultados apresentados a seguir são referentes aos três modelos propostos utilizando o tratamento do banco de dados **BDOrig**, juntamente com os melhores resultados obtidos no trabalho original:

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	13,2245	13,0979	13,1597	12,86 (SVM)
RMSE	50,1990	50,5531	50,3109	63,70 (*Naive*)

Tabela 3.8: Resultados do Modelo Trimestres - **BDOrig**.

Observa-se que o modelo de Regressão apresentou o pior desempenho em relação ao MAD, enquanto mostrou o melhor RMSE. Dito isto, as diferenças não são grandes, tendo o modelo SVM uma melhora no MAD menor que 1% sobre o modelo de Regressão, o qual mostrou um RMSE quase 1% menor do que o do modelo SVM.

Já em relação aos resultados obtidos no trabalho original, novamente o RMSE apresentado pelos nossos modelos são bastante menores do que o obtido por Cortez (2007a), enquanto os valores de MAD obtidos se apresentaram muito similares ao do trabalho original, ainda que o modelo SVM, o de menor erro, se mostre quase 2% menos preciso.

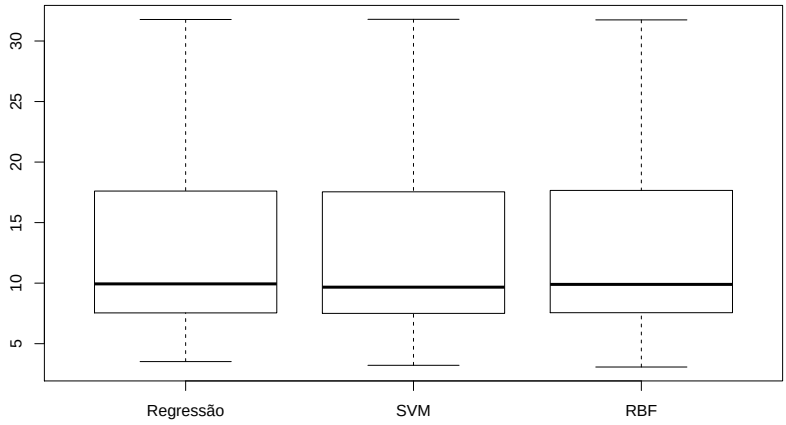

Figura 3.16: *Boxplots* da métrica de erro MAD obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

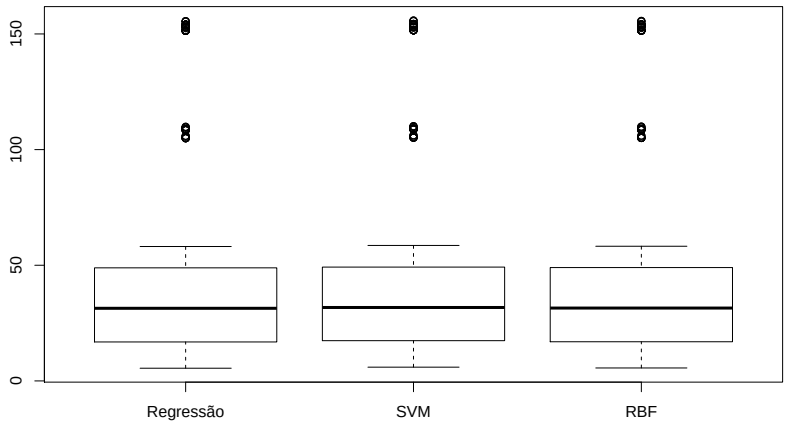

Figura 3.17: *Boxplots* da métrica de erro RMSE obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

Não é possível notar diferenças significativas nos *boxplots* dos erros MAD e RMSE, ilustrados nas Figuras 3.16 e 3.17.

3.5.2 Tratamento BDSZero

Na Tabela 3.9 estão as métricas de erro utilizando o tratamento do banco de dados **BDSZero**, juntamente com os melhores resultados obtidos no trabalho original:

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	1,0341	1,0397	1,1578	12,86 (SVM)
RMSE	1,3528	1,3413	1,4979	63,70 (*Naive*)

Tabela 3.9: Resultados do Modelo Trimestres - **BDSZero**.

Neste caso se destaca negativamente o modelo RBF, apresentando um MAD quase 12% maior do que o modelo de Regressão, enquanto mostra um RMSE 11% maior do que o do modelo SVM. Já os modelos de Regressão e SVM não se mostram significativamente diferentes, com diferenças de MAD e RMSE menores que 1%. Novamente não comparamos estes resultados com os obtidos no trabalho original devido à alteração do banco de dados.

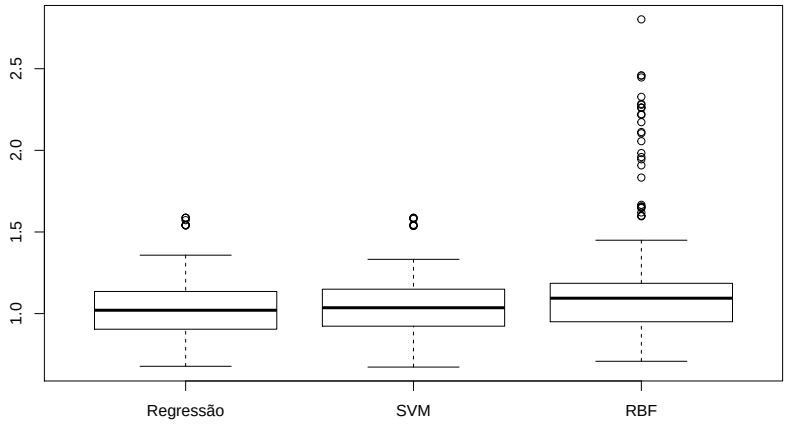

Figura 3.18: *Boxplots* da métrica de erro MAD obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

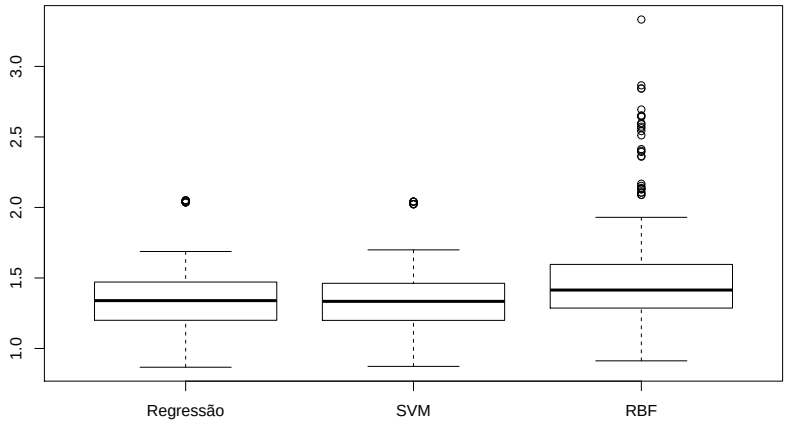

Figura 3.19: *Boxplots* da métrica de erro RMSE obtida nas 30 repetições da validação cruzada em 10 etapas proporcionais ao trimestre.

. É notável, mais uma vez, o pior desempenho do modelo RBF no caso **BDS-Zero**. As Figuras 3.16 e 3.17 mostram, também, a pequena vantagem que o modelo SVM leva sobre o modelo de Regressão.

3.5.3 Tratamento BDBin

Vemos a seguir os resultados obtidos utilizando o tratamento do banco de dados **BDBin**:

Métrica de Erro	Regressão	SVM	RBF
Sensibilidade	0,8422	0,8347	0,9185
Especificidade	0,2335	0,2324	0,1105
Área Abaixo da Curva ROC	0,5378	0,5336	0,5145

Tabela 3.10: Resultados do Modelo Trimestres - **BDBin**.

Apesar de o modelo RBF apresentar a melhor medida de Sensibilidade, até 10% maior do que o modelo SVM, mostrou uma Especificidade muito baixa, chegando a ser mais de 50% menor do que a Especificidade apresentada pelo modelo de Regressão. As medidas de Área Abaixo da Curva ROC se mostraram similares neste caso, com o maior valor sendo do modelo de Regressão, menos de 5% maior do que do modelo RBF, o menor dos três. O modelo de Regressão, no geral, se destacou positivamente, apresentando maiores valores com a exceção da Sensibilidade, porém, as diferenças para o modelo SVM não são grandes, todas menores do que 1%.

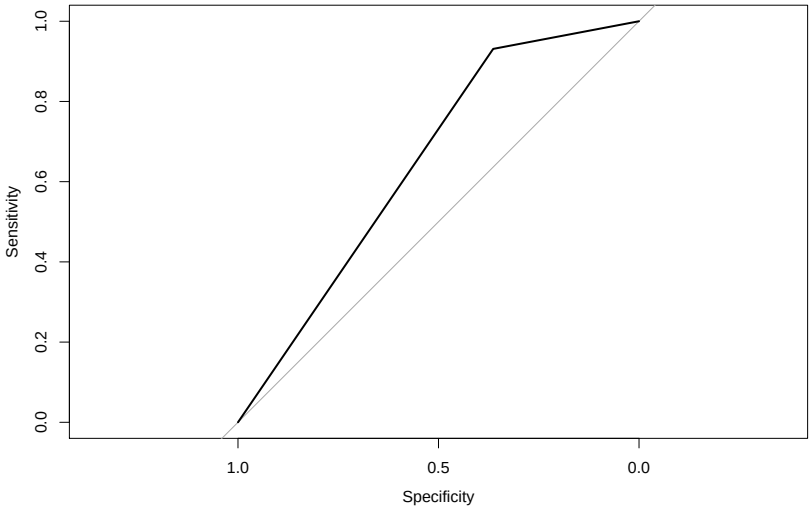

Figura 3.20: Curva ROC.

Na Figura 3.20 observamos o mesmo comportamento dos outros tipos de tratamento do banco de dados: melhor curva do modelo de Regressão, baixa Especificidade e Sensibilidade alta.

3.6 Melhores Resultados

Nesta Seção iremos analisar e buscar os melhores resultados, a fim de buscar o melhor modelo.

3.6.1 Tratamento BDOrig

Para o caso do tratamento do banco de dados **BDOrig**, vemos na Tabela 3.11 que o modelo que apresentou melhores resultados foi o Reduzido, apenas com as variáveis de entrada **XY**, **tri1**, **RH** e ***wind***. Dentro desta proposta de variáveis de entrada, os modelos SVM e de Regressão se sobressaíram, com o modelo SVM apresentando um MAD menor enquanto o modelo de Regressão aparece com o menor RMSE.

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	13,2322	**13,0473**	13,1175	12,86 (SVM)
RMSE	**50,1632**	50,4387	50,4245	63,70 (*Naive*)

Tabela 3.11: Melhores Resultados de **BDOrig**. Modelo Reduzido.

3.6.2 Tratamento BDSZero

Utilizando o tratamento do banco de dados **BDSZero** o modelo que levava apenas as variáveis indicadoras dos trimestres como entrada se destacou, apresentando os menores valores. Vemos na Tabela 3.12 que o desempenho do modelo RBF não

é bom quando a variável resposta, ***area***, não apresenta os zeros. Novamente os modelos SVM e de Regressão geram erros bastante próximos, com o SVM mostrando um RMSE menor, enquanto o de Regressão traz um MAD menor.

Métrica de Erro	Regressão	SVM	RBF	Original
MAD	**1,0341**	1,0397	1,1578	12,86 (SVM)
RMSE	1,3528	**1,3413**	1,4979	63,70 (*Naive*)

Tabela 3.12: Melhores Resultados de **BDSZero**. Modelo Trimestres.

3.6.3 Tratamento BDBin

Ao avaliar os resultados do tratamento **BDBin** na Tabela 3.13, nota-se que os melhores resultados foram mais distribuídos entre os modelos. A proposta de variáveis de entrada Reduzida, isto é, com apenas as quatro variáveis selecionadas pelo algoritmo, se destacou apresentando os melhores resultados de Especificidade e Área Abaixo da Curva ROC com os modelos RBF e SVM, respectivamente. Já a maior Sensibilidade foi obtida pelo modelo RBF com apenas as variáveis indicadoras dos trimestres como entrada. O modelo Inicial, com todas as variáveis de entrada, não apresentou nenhum destaque.

O modelo SVM Reduzido apresentou o maior valor de Área Abaixo da Curva ROC, enquanto a melhor Sensibilidade foi obtida pelo modelo RBF com apenas as variáveis indicadores dos trimestres. Já a maior Especificidade foi obtida pelo modelo Reduzido RBF. Vale salientar que, exceto pela Especificidade, a proposta de variáveis de entrada reduzida apresentou as melhores métricas no geral.

Inicial	Regressão	SVM	RBF
Sensibilidade	0,6427	0,7379	0,8976
Especificidade	0,3903	0,3644	0,1717
Área Abaixo da Curva ROC	0,5165	0,5511	0,5221
Reduzido	Regressão	SVM	RBF
Sensibilidade	0,7040	0,8529	0,5774
Especificidade	0,3447	0,2798	**0,4959**
Área Abaixo da Curva ROC	0,5244	**0,5596**	0,5367
Trimestres	Regressão	SVM	RBF
Sensibilidade	0,8422	0,8347	**0,9185**
Especificidade	0,2335	0,2324	0,1105
Área Abaixo da Curva ROC	0,5378	0,5336	0,5145

Tabela 3.13: Melhores Resultados de **BDBin**.

4 Conclusão

Ao observar os resultados que obtivemos, podemos dizer que atingimos o objetivo de trabalhar com os modelos propostos neste trabalho e gerar resultados similares aos atingidos no trabalho original, por Cortez (2007a), com apenas diferenças marginais em relação ao erro MAD, e resultados superiores em relação ao erro RMSE. Acredito que, a partir dos testes realizados, pode-se dizer que os melhores modelos encontrados neste trabalho são capazes de praticar previsões de incêndios dentro do escopo do banco de dados, isto é, dentro do Parque Natural de Montesinho ou em regiões de similares características com razoável acurácia.

Analisando os melhores resultados alcançados pelos modelos propostos por este trabalho, pode-se especular que, em geral, os modelos de Regressão e SVM apresentaram os melhores desempenhos, mesmo quando olhamos para o caso **BDBin**, que foi onde o modelo RBF se destacou. É importante salientar o bom desempenho geral do modelo SVM que, apesar de revelar erros próximos aos do modelo de Regressão, em poucos casos mostrou um desempenho ruim ao fazer as predições.

Em relação às propostas de variáveis de entrada do modelo, fica claro que, para os casos **BDOrig** e **BDSZero**, o modelo Original, ou seja, com todas as variáveis de entrada possíveis, apresenta o pior desempenho. Utilizando o banco de dados original, **BDOrig**, se destaca o modelo Reduzido, onde apenas entram as variáveis especificadas pelo algoritmo *Stepwise Backward*. Já no tratamento do banco de dados **BDSZero**, ao se retirar os zeros, vemos que o modelo que leva como entrada as variáveis indicadoras dos trimestres se sai melhor, além de o modelo RBF apresentar problemas para lidar com este tipo de banco de dados. No caso do **BBin**, onde a variável resposta é tratada como binária, os melhores modelos não ficaram tão evidentes, visto que, para uma métrica de erro, um modelo se sobressai, enquanto, para outra, não funciona tão bem.

Utilizando o tratamento do banco **BDSZero**, notamos a dificuldade em comparar os resultados dos três modelos aqui propostos com os erros atingidos por Cortez (2007a) , visto a diferença de magnitude entre os mesmos. Este fenômeno provavelmente se dá pela ausência dos zeros na variável resposta, ***area***, o que encurta muito as distâncias dos pontos reais para os preditos, pois não há mais uma sequência grande de zeros no banco de dados. Essa diminuição da distância acarreta em baixos somatórios dos erros. Esta característica do modelo RBF ficou evidente ao observar os *boxplots* dos erros, ilustrados, principalmente, nas Figuras 3.3, 3.4, 3.13 e 3.14.

Nota-se ainda um comportamento peculiar da Especificidade no tratamento **BDBin**, apresentando baixos valores em todas as situações, nunca apresentando um

valor maior do que $0,5$, observado na Tabela 3.13. É interessante ressaltar que esse comportamento se acentua ao dar como entrada para o modelo as variáveis indicadoras dos trimestres, não atingindo um valor de Especificidade maior do que $0,24$ em nenhum dos modelos, como visto na Tabela 3.10.

Esse desempenho peculiar da Especificidade é mais perceptível ao cruzar a Especificidade com o modelo RBF, onde chega a atingir o maior valor no modelo Reduzido, e os menores nas outras propostas de variáveis de entrada. A Especificidade tem essa característica melhor observada ao examinar todas as curvas ROC, nas Figuras 3.5, 3.15 e 3.20, que chegam perto de serem uma reta, isto é, aleatoriedade na predição, visto que mostra, apenas, uma Sensibilidade relativamente alta. A baixa Especificidade acarreta em um modelo capaz de predizer com acurácia a ocorrência de um incêndio, porém, tende a errar ao predizer a ausência do mesmo.

Dada a busca de um modelo parcimonioso capaz de realizar as predições de incêndio dentro de todas as propostas de modelos e tratamentos do banco de dados, é possível conjecturar que, no geral, o modelo SVM com a dimensão de entrada reduzida é o que apresentou melhores resultados neste trabalho. O modelo SVM Reduzido apresenta o menor erro MAD no caso **BDOrig**, enquanto para o **BDSZero** alcança valores muito próximos aos ótimos. Já quando o tratamento da variável resposta é binário, **BDBin**, o modelo SVM com entrada reduzida não atinge nenhum destaque, porém, no geral, obtém valores bons em todas as métricas. Em termos de parcimônia, o modelo Reduzido é bastante econômico em tamanho de entrada, perdendo apenas para o modelo dos trimestres. Ainda assim, em termos gerais de parcimônia e acurácia, o modelo SVM Reduzido se mostrou o melhor deste trabalho.

O comportamento muito similar de todos os modelos pode ter ocorrido pelo excesso de zeros na variável ***area***, o que a tornaria uma variável mista. Isto é, no momento em que existe um excesso de, neste caso, zeros, a variável apresenta um comportamento discreto em grande parte de sua análise, enquanto, em outras instâncias, se mostra contínua. Esta característica pode ter influenciado nos resultados e causado a pouca distinção entre os modelos. Outra característica que pode ter levado aos resultados obtidos é baixa dimensionalidade de entrada dos modelos, fazendo com que modelos mais complexos como o SVM e o RBF não se mostrem mais eficientes do que modelos mais simples como os de Regressão.

Este trabalho buscou aliar modelos computacionais à ferramentas estatísticas para encontrar os melhores modelos capazes de predizer o comportamento de incêndios florestais em regiões supra-mediterrâneas. Em projetos futuros devemos buscar alternativas na construção de modelos para previsão de incêndios em regiões secas, como a abordada neste trabalho. Também se buscará em posteriores pesquisas se aprofundar em questões abordadas nas seções de Resultados, como a queda de desempenho do modelo RBF quando se retira os zeros da variável resposta, ou o comportamento abaixo do esperado da métrica de Especificidade e, por fim, explorar modelos capazes de incorporar a característica mista da variável ***area***.

Referências Bibliográficas

Akaike, H. (1987). *Factor analysis and AIC. Psychometrika*, 52:317–332.

Cortez, P. e Morais, A. (2007a). *A Data Mining Approach to Predict Forest Fires using Meteorological Data*. PhD thesis, Department of Information Systems, University of Minho, Guimarães, Portugal.

Cortez, P. e Morais, A. (2007b). UCI machine learning repository. (http://archive.ics.uci.edu/ml). University of California, Irvine, School of Information and Computer Sciences, 2007.

Cover, T. M. (1965). *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers*, EC-14:326–334.

da Conservação da Natureza, I. (2013). *Incêndios Florestais na Rede Nacional de Áreas Protegidas em 2013*. (http://www.icnf.pt/portal/florestas/dfci/Resource/doc/rel/2013/relatorio-dfci-ap-2013).

de Dados Portugal Contenporâneo, P. B. (2017). *Incêndios florestais e área ardida - Continente*. (http://www.pordata.pt/Portugal/Inc%C3%AAndios+florestais+e+%C3%A1rea+ardida+%E2%80%93+Continente-1192-9576).

de Notícias, J. (2015). *75% dos incêndios florestais são de origem criminosa*. (https://www.jn.pt/nacional/interior/75-dos-incendios-florestais-sao-de-origem-criminosa-4864388.html?id=4864388).

Haykin, S. (2001). *Redes Neurais: Princípios e prática*. Bookman, Porto Alegre, RS, 2ª edição.

Hosmer, D. e Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons, New York, EUA.

Kavzoglu, T. e Colkesen, I. (2009). *A Kernel Functions Analysis for Support Vector Machines for Land Cover Classification. International Journal of Applied Earth Observation and Geoinformation*, 11:352–359.

Kim, H. R. (2009). *Prediction of Forest Fires using Data Mining Methods*. Master's thesis, University of Western Ontario.

"Kovács, Z. L. (2002). *Redes Neurais Artificiais*. Editora Livraria da Física, São Paulo, SP.

Lorena, A. C. e Carvalho, A. C. P. L. F. (2007). *Uma Introdução às Support Vector Machines*. *Revista de Informática Teórica e Aplicada*, XIV(2).

Micchelli, C. A. (1986). *Interpolation of scattered data: Distance matrices and conditionally positive definite functions*. *Constructive Approximation*, 2:11–22.

Pereira, G. G. A. e Centeno, J. A. S. (2013). *Utilização de Support Vector Machine para classificação multiclasses de imagens Landsat TM+*. In *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto*, Foz do Iguaçu, PR, Brasil.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org/).

Ripley, B. D. (2008). *Pattern Recognition and Neural Networks*. Cambridge University Press, New York, EUA.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. (http://www.rstudio.com/).

Shrivastava, P. e Shukla, M. (2014). *Uses The Bagging Algorithm of Classification Method with WEKA Tool for Prediction Technique*. In *Proceedings of 16th IRF International Conference, 26th October 2014, ISBN: 978-93-84209-60-5*, Chennai, India.

Steinwart, I. e Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media, New York, EUA.

Vapnik, V. N. (1992). *The nature of Statistical learning theory*. Springer-Verlag, New York, EUA.