Escola Politécnica da Universidade de São Paulo
RENATA RAMOS RODRIGUES DE PAULA

Análise comparativa de modelos de estatística multivariada aplicados à previsão de níveis de poluentes atmosféricos

São Paulo
2017

Escola Politécnica da Universidade de São Paulo
RENATA RAMOS RODRIGUES DE PAULA

Análise comparativa de modelos de estatística multivariada aplicados à previsão de níveis de poluentes atmosféricos

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

São Paulo

2017

Escola Politécnica da Universidade de São Paulo
RENATA RAMOS RODRIGUES DE PAULA

Análise comparativa de modelos de estatística multivariada aplicados à previsão de níveis de poluentes atmosféricos

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

Área de Concentração:
Engenharia Química

Orientador:
Dr. Roberto Guardani

São Paulo

2017

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

de Paula, Renata
Análise comparativa de modelos de estatística multivariada aplicados à previsão de níveis de poluentes atmosféricos / R. de Paula -- versão corr. -- São Paulo, 2017.
p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Química.

1.Estatística multivariada 2.Machine learning 3.Qualidade do ar 4.Ozônio I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Química II.t.

Agradecimentos

À minha família, em especial minha mãe, Teresa, e meu pai, Claudio, pelo apoio que me deram durante esta jornada, por me ensinarem a importância do conhecimento e por estarem sempre presentes no momentos mais difíceis.

Às minhas colegas Ana Maria e Cristhiane, por terem me ajudado em várias etapas do mestrado e por terem sido boas companhias nas tardes de trabalho.

Ao Professor Roberto Guardani por ter me dado a oportunidade de crescimento profissional, por ter me proporcionado a enriquecedora experiência de estagiar na Universidade de Toronto e a oportunidade de divulgar este trabalho no congresso anual da AIChE nos Estados Unidos e, finalmente, por toda orientação e atenção que dedicou a mim e a este trabalho.

A toda a equipe da CETESB pelo fornecimento dos dados e por compartilharem sua experiência de modo a enriquecer esta pesquisa e ao IAG pelo fornecimento dos dados meteorológicos.

À CAPES pela bolsa de mestrado, ao Departamento de Engenharia Química da Escola Politécnica da USP e a todos que direta ou indiretamente contribuíram para a realização deste trabalho.

A alegria de ver e compreender é a mais bela dádiva da natureza

Albert Einstein

Resumo

O presente estudo visa à análise comparativa do desempenho dos modelos de estatística multivariada *Multi-layer Perceptron Neural Networks*, *Random Forests* e *Support Vector Machine* na previsão de máxima concentração diária de ozônio na baixa atmosfera na Região Metropolitana de São Paulo (RMSP), caracterizada pela alta concentração de habitantes e intensa atividade econômica, onde a qualidade do ar é afetada principalmente por episódios de altos níveis de ozônio. Foram aplicados tanto modelos de regressão quanto de classificação. Nos casos de classificação, estudou-se também o desempenho de dois modelos de análise de discriminantes: *Linear Discriminant Analysis* e *Fisher Discriminant Analysis*. Para a construção dos modelos utilizou-se uma base de dados com medições de variáveis meteorológicas, além da concentração de ozônio, fornecida pela Companhia Ambiental do Estado de São Paulo (CETESB). Dada a grande importância e a complexidade do processo de formação de ozônio na baixa atmosfera, a Universidade de São Paulo (USP) e a CETESB têm desenvolvido estudos no tema desde 1999, através dos quais produziram-se modelos de previsão baseados em redes neurais, implementados pela equipe da CETESB. O presente estudo é uma continuação do desenvolvimento anterior e contém as seguintes inovações quanto à metodologia e resultados esperados: (1) ajuste de novos modelos com novas estruturas, incluindo-se técnicas de *Support Vector Machine*, *Random Forests* e Discriminação; (2) uso de uma base de dados mais ampla e atualizada, de modo a melhorar a representatividade dos modelos; (3) ajuste dos modelos à nova legislação, Decreto Estadual 59.113 de 23/04/2013, que estabelece novos padrões de qualidade do ar para os poluentes atmosféricos, dentre os quais o ozônio. Embora nos casos de classificação nenhum dos modelos tenha apresentado bons resultados, nos casos de regressão foi possível obter resultados melhores do que os esperados. O modelo de *Multi-layer Perceptron* foi o que mostrou melhor desempenho para prever concentrações máximas de ozônio, tanto para a previsão de máximas concentrações baseadas em médias horárias quanto em médias móveis de 8 horas, que resultaram em coeficientes de correlação 0,867 e 0,891, respectivamente.

Abstract

The present study aims to compare the performance of the multivariate statistical models Multi-layer Perceptron Neural Networks, Random Forests and Support Vector Machine applied to the prediction of daily maximum concentrations of ground-level ozone in the Metropolitan Area of São Paulo (MASP), characterized by the high population density and the intense economic activity, where the air quality is mostly affected by high ozone levels. Both regression and classification models were applied. In the classification cases, two more models were applied: Linear Discriminant Analysis and Fisher Discriminant Analysis. The models were constructed using a database containing meteorological variables and daily maximum ozone concentration values, which were provided by the Environmental Agency of São Paulo State (CETESB). Given the great importance and complexity of the process of ozone formation in the troposphere, the University of São Paulo (USP) and CETESB have made studies in this area since 1999 and developed a prediction model based on neural networks, which was implemented by CETESB. The present study is a continuation of the previous one and contains the following innovations regarding the methodology and expected results: (1) comparison with other models such as support vector machines, random forests and discriminant analysis; (2) use of a wider and up-to-date database, which improves the representativeness of the models; (3) the models took into acount the new legislation, State decree 59113 of 04/23/2013, that establishes new air quality standards for ozone. Although none of the classification models had a good performance, the regression models yielded better than expected results. The multi-layer perceptron model was the one with higher performance in the prediction of daily maximum ozone concentrations based both on hourly averages and on eight-hour moving averages, which yielded correlation coefficients of 0.867 and 0.891 respectively.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BFGS - Método de Broyden-Fletcher-Goldfarb-Shanno

CART - *Classification and Regression Trees*

CETESB - Companhia Ambiental do Estado de São Paulo

COV - Compostos Orgânicos Voláteis

CPTEC - Centro de Previsão do Tempo e Estudos Climáticos

DVG - Direção do Vento Global

IAG - Instituto de Astronomia, Geofísica e Ciências Atmosféricas

ID - Índice de Desempenho

INPE - Instituto Nacional de Pesquisas Espaciais

KKT - Karush-Kuhn-Tucker

LDA - *Linear Discriminant Analysis*

LSCP - Laboratório de Simulação e Controle de Processos

MAE - *Mean Absolute Error*

MBE - *Mean Bias Error*

MLP - *Multi-layer Perceptron Neural Networks*

MSE - *Mean Squared Error*

QDA - *Quadratic Discriminant Analysis*

RF - *Random Forests*

RMSP - Região Metropolitana de São Paulo

SVM - *Support Vector Machine*

USP - Universidade de São Paulo

UV - Ultravioleta

VC - Validação Cruzada

VV - Velocidade do Vento

Lista de Símbolos

T - Temperatura

UR - Umidade Relativa

Pr - Pressão

U - Componente Leste-Oeste da Velocidade do Vento

V - Componente Norte-Sul da Velocidade do Vento

λ - Comprimento de Onda da Luz

K - Número de classes possíveis

N - Número de observações na base de dados

P - Número de variáveis

$w_{p,j}$ - Peso aplicado no neurônio para compor o neurônio j

S_j - Combinação linear dos neurônios da camada anterior com os pesos dessa camada para formar o neurônio j da próxima camada, dada pela seguinte equação: $S_j = w_{P+1,j} + \sum_{p=1}^{P} w_{p,j}.x_p$

L - Função de Verossimilhança (*Likelihood*)

E - Função Entalpia Cruzada

$p(k|X^{(n)})$ - Probabilidade condicional de uma observação n pertencer à classe k

N_f - Número de observações na folha f

$N_{k,f}$ - Número de observações da classe k na folha f

F - Número de folhas em uma árvore

B - Número de *subsets* gerados por *bagging*

m - Número de variáveis selecionadas para a decisão em um nó da árvore

d - Discriminante

a - Vetor de pesos das observações para compor o discriminante

B - Matriz da soma ponderada das distâncias quadráticas entre grupos

W - Matriz da soma da distância quadrática entre observações $\mathbf{X_k^{(n)}}$ de um grupo k e a média $\overline{\mathbf{X}}_k$ deste grupo

σ - Razão a ser maximizada para encontrar os pesos dos discriminantes

w - Vetor de pesos que determina o plano separador na técnica SVM

b - Escalar que compõe o plano separador na ténica de SVM

μ - Multiplicadores KKT

ξ e $\xi*$ - Variável de tolerânica da técnica de SVM

C - Custo das tolerâncias na técnica de SVM

$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)})$ - Kernel

ϵ - Distância das margens nos casos de regressão por SVM

R - Coeficiente de correlação

n_k - Número de observações pertencentes ao grupo, ou classe, k

μ_k - Média da população do grupo k

Σ_k - Matriz de covariância da amostra da população k no *training set*

π_k - Probabilidade *a priori* de uma observação pertencer ao grupo k

$discQ_k$ - Discriminante quadrático para o grupo k

Sumário

1 Introdução

Segundo a Organização Mundial da Saúde, em 2012, cerca de 7 milhões de pessoas no mundo morreram devido à exposição à poluição atmosférica. Ou seja, uma em cada oito mortes no mundo foi provocada pelo que hoje é considerado o maior risco ambiental à saúde [WHO, 2016]. Por esse motivo, a previsão de episódios críticos de poluição do ar é um assunto de crescente preocupação das autoridades, especialmente em grandes centros urbanos, onde há grande concentração de pessoas e ocorre emissão mais intensa de poluentes.

O uso de modelos determinísticos para a previsão dos níveis de ozônio (O_3) na baixa atmosfera apresenta limitações, uma vez que ainda não são entendidos em profundidade todos os fenômenos químicos e físicos que influenciam na produção e na dispersão dos poluentes na atmosfera urbana. A complexidade é ainda maior pelo fato de o O_3 ser um poluente secundário, ou seja, é gerado na atmosfera como consequência de interações entre poluentes precursores de sua formação, como óxidos de nitrogênio (NO_x) e compostos orgânicos voláteis (COVs), e condições meteorológicas locais.

Em substituição aos modelos determinísticos, vários modelos estatísticos têm sido propostos para a previsão de níveis de ozônio troposférico [Guardani et al., 2003, Burrows et al., 1994, Borges et al., 2012, Lu and Wang, 2008, Sucar et al., 1997, Abdul-Wahab and Al-Alawi, 2002]. Dentre os vários modelos propostos, destacam-se a análise de agrupamentos [Guardani et al., 2003], árvores de decisão [Burrows et al., 1994], redes neurais do tipo *multi-layer perceptron* [Borges et al., 2012] e *support vector machine* [Lu and Wang, 2008]. Baseados na análise de dados históricos obtidos a partir de estações monitoras da qualidade do ar, os modelos estatísticos são capazes de estabelecer relações entre variáveis de entrada e de saída (predições) sem estabelecer relações de causalidade.

Das várias metodologias estatísticas existentes, as redes neurais *multi-layer perceptron* têm se mostrado eficazes na captura das relações complexas e não lineares entre variáveis meteorológicas e concentrações de poluentes no ar [Guardani et al., 1999, Arhami et al., 2013], a exemplo do que se conhece quanto à aplicação desses modelos a vários sistemas complexos, em diferentes áreas de interesse.

Além da seleção do modelo apropriado, o sucesso na aplicação de modelos

estatísticos multivariados depende principalmente de dois fatores:

1. Qualidade dos dados disponíveis para ajuste de parâmetros dos modelos;
2. Seleção consistente das variáveis de entrada e de saída, de modo a representar as correlações a serem estabelecidas.

A qualidade do ar em grandes centros urbanos é afetada por uma grande quantidade de variáveis que representam as condições meteorológicas e as fontes emissoras, sendo, portanto, altamente dependente das condições locais em cada região específica. Tal fato tem sido evidenciado nos relatórios de qualidade do ar publicados anualmente pela CETESB [CETESB, 2016b], assim como em publicações anteriores da equipe de pesquisa da USP sobre o tema [Guardani and Nascimento, 2004, Guardani et al., 2003, Borges et al., 2012]. Desta forma, a construção de um modelo apropriado, capaz de fazer previsões com alto nível de confiança de forma rápida para cada região urbana específica tem sido o foco de muitas pesquisas recentes.

1.1 Motivação

A Região Metropolitana de São Paulo (RMSP) é uma das maiores regiões urbanas do mundo e é considerada o principal centro econômico do Brasil. Está localizada a cerca de 50 km do mar, cerca de 700 m de altitude, entre duas cadeias montanhosas e na área de transição do sub-trópico para o trópico, tem clima local irregular com grandes variações de condições em curtos períodos de tempo. Este comportamento contrasta com boa parte dos demais grandes centros urbanos do hemisfério norte, os quais apresentam clima relativamente regular, o que facilita as previsões climáticas e de qualidade do ar. A RMSP contém uma população de cerca de 21 milhões de habitantes, ou seja, 48% da população total do estado de São Paulo, concentrados em apenas 3,2% do território estadual [CETESB, 2014]. A deterioração atmosférica na região é causada, especialmente, por emissões veiculares, visto que a região contabilizou, em 2012, 49% da frota total de veículos do estado [CETESB, 2014].

Um dos poluentes mais problemáticos na RMSP é o O_3, que é um poluente secundário formado a partir de NO_x e COVs. De acordo com o relatório de qualidade do ar de 2015 da CETESB [CETESB, 2016b], a RMSP apresenta um longo histórico

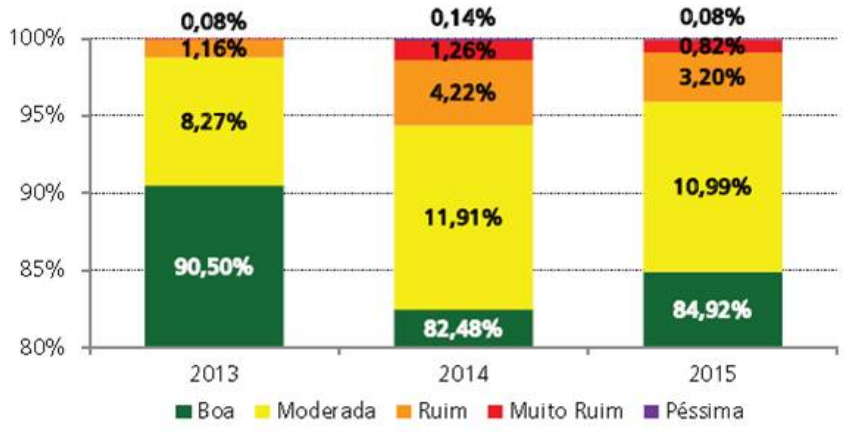

Figura 1.1: O3 - Distribuição percentual da qualidade do ar na RMSP [CETESB, 2016b]

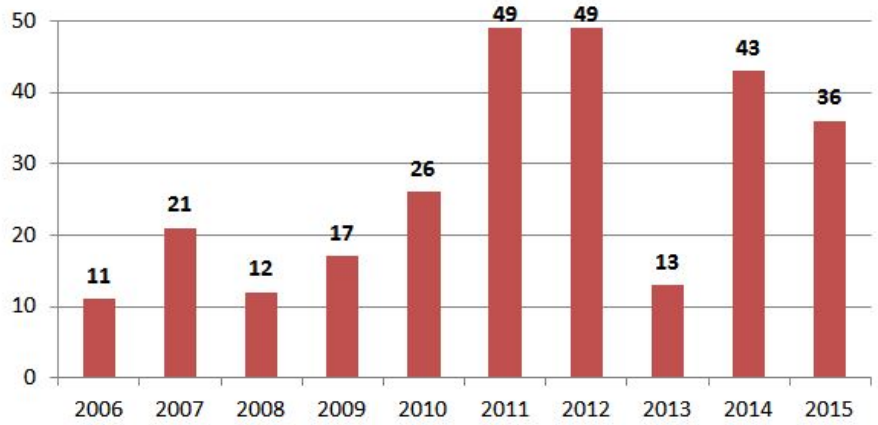

Figura 1.2: O3 - Número de dias de ultrapassagem do padrão de qualidade do ar estadual na RMSP [CETESB, 2016b]

de ultrapassagem dos padrões de qualidade do ar de O_3 (Figura 1.1). Em 2015, foram observados 36 dias em que houve violação do padrão estadual de 8 horas (140 μg/m3), considerando-se todas as estações que medem os níveis de O_3, conforme mostra a Figura 1.2. O padrão estadual foi recentemente atualizado pelo Decreto Estadual nº 59.113 em 2013 e estabelece metas mais rígidas para a redução gradual dos níveis de poluentes atmosféricos de modo a alcançar os padrões estabelecidos pela Organização Mundial da Saúde (WHO).

A formação de O_3 depende tanto de fenômenos químicos quanto físicos como, por exemplo, a presença de radiação solar, a forma como essa radiação se distribui no ambiente, a dispersão dos poluentes primários, a existência ou não de vento e sua direção, entre outros. Como as condições meteorológicas na RMSP variam intensamente

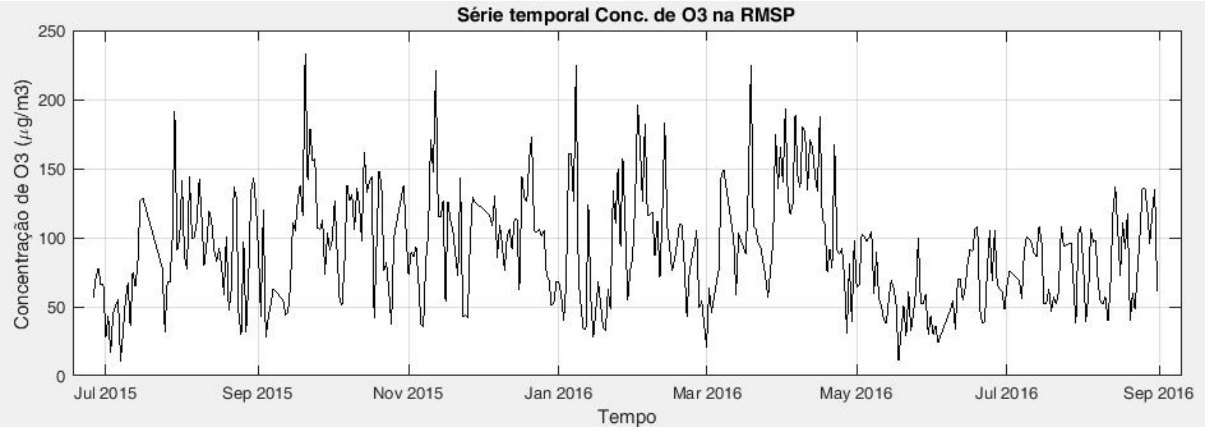

Figura 1.3: Série temporal das concentrações máximas diárias de O_3 baseadas nas médias horárias, medidas na estação Ibirapuera da CETESB

em intervalos curtos de tempo, o processo de formação de O_3 é afetado, resultando em concentrações máximas diárias que diferem muito de um dia para o dia seguinte, como mostra a Figura1.3, cujos dados provêm da estação Ibirapuera da CETESB.

A CETESB possui na RMSP um total de 26 estações monitoras da qualidade do ar, sendo que 19 delas são capazes de medir concentração de O_3, conforme mostra a Figura 1.4. A estação localizada no parque Ibirapuera, a 5,2 km do centro de São Paulo, é geralmente utilizada como referência nos estudos de concentração de O_3 da região por dois motivos principais:

1. ela apresenta os mais elevados níveis de O_3 de toda a região

2. o comportamento dos seus níveis de O_3 é representativo do de várias outras estações da RMSP, conforme foi constatado através de análises de agrupamentos em um estudo anterior do departamento de Engenharia Química da USP em conjunto com a CETESB [Guardani et al., 2003].

Portanto, a previsão de episódos de alta concentração de O_3 baseada em análises com medidas da estação Ibirapuera tem forte representatividade da RMSP como um todo. Prever esses episódios é de essencial importância para órgãos públicos, especialmente nos últimos tempos devido à atualização dos padrões de qualidade do ar pelo Decreto Estadual nº 59.113. A previsão possibilita a execução de medidas de controle de emissões com antecipação, além de permitir que a população seja alertada sobre os cuidados a serem observados.

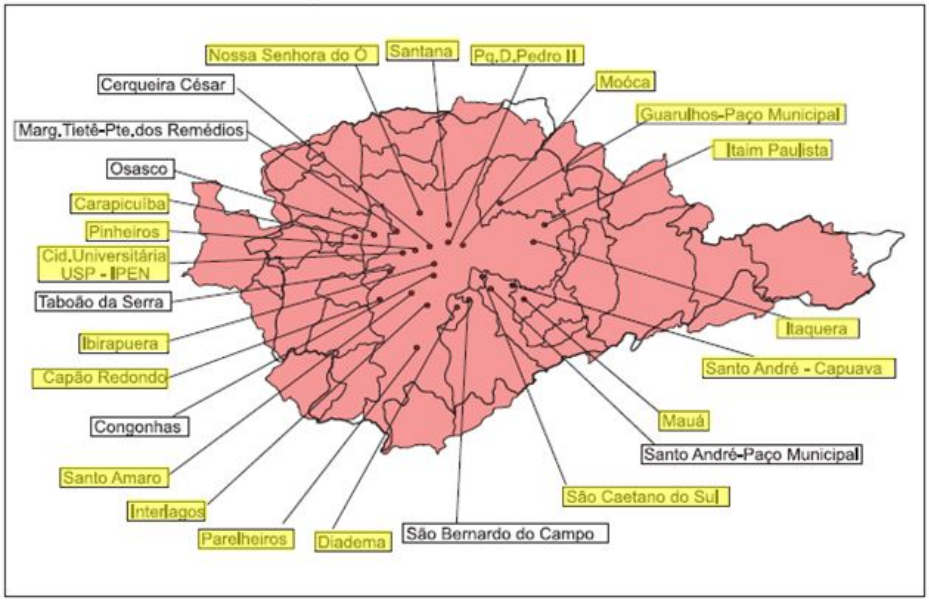

Figura 1.4: Distribuição das estações da Rede Automática da CETESB na RMSP [CETESB, 2014]

Os primeiros trabalhos publicados sobre aplicação de técnicas estatísticas multivariadas na previsão de níveis do poluente atmosférico O_3 surgiram no início da década de 1990 [Burrows et al., 1994, Comrie, 1997, Robeson and Steyn, 1990]. Desde então, muitos artigos dedicados ao tema têm sido publicados, sendo a grande maioria deles aplicada a cidades ou regiões específicas. Essas publicações têm contribuído para o desenvolvimento da área por meio de aplicações envolvendo essencialmente combinações de técnicas, que têm resultado em melhorias na qualidade das previsões, dependendo do local específico em que tais estudos são aplicados. Alguns exemplos são apresentados por Feng et al. (2011), que combinou técnicas de redes neurais e *Support Vector Machine* na predição de níveis de ozônio para a cidade de Pequim na China e, mais recentemente, por Sundaramoorthi (2014), que combinou a técnica de árvores de decisão com estimativa da densidade kernel.

No departamento de Engenharia Química da Universidade de São Paulo, estudos com redes neurais do tipo *multi-layer perceptron* têm sido desenvolvidos desde a década de 1990 [Guardani et al., 1999, Guardani and Nascimento, 2004, Borges et al., 2012], sendo que o último deles dedicou-se a criar um modelo para as concentrações média e máxima diárias de O_3 usando como variáveis de entrada apenas dados meteorológicos.

Diferentes modelos de estatística multivariada foram, portanto, desenvolvidos por autores ao redor de todo o mundo mas ainda não existe uma análise compara-

tiva da eficiência dos modelos mais comuns (*multi-layer perceptron*, *support vector machine* e análise de discriminantes) e mais recentes (métodos de conjunto de especialistas como *random forests*, criado a partir de árvores de decisão) aplicados à previsão de níveis de concentração de O_3 para dados de uma mesma região.

1.2 Objetivos

Este estudo tem por objetivo realizar uma análise comparativa da precisão dos métodos de *Multi-layer Perceptron* (MLP), *Random Forests* (RF), *Support Vector Machine* (SVM) e *Discriminant Analysis* (LDA e Fisher) na predição da concentração máxima diária de O_3 troposférico em três diferentes análises:

1. valores diários máximos de médias horárias;
2. valores diários máximos de médias móveis de 8 horas;
3. classificação das máximas de médias móveis de 8 horas em cinco diferentes grupos, de acordo com critérios estabelecidos pelas CETESB.

As aplicações dos métodos foram feitas utilizando uma base de dados históricos dos anos de 2000 a 2016 contendo médias horárias de concentração de O_3, temperatura (T), umidade relativa (UR), pressão (Pr) e componentes da velocidade do vento nos sentidos Norte-Sul (V) e Leste-Oeste (U). Os dados utilizados foram fornecidos pela CETESB, provenientes da estação de monitoração automática localizada no Parque Ibirapuera, e pelo Instituto de Astronomia, Geofísica e Ciências Atmosféricas da USP (IAG-USP), provenientes de sua estação meteorológica localizada no Parque de Ciência e Tecnologia da USP.

O estudo consiste de 3 etapas: seleção das variáveis mais representativas do fenômeno, a aplicação e comparação dos diferentes modelos para o caso de regressão (previsão de níveis de ozônio) e aplicação e comparação dos diferentes modelos para o caso de classificação das observações.

1.3 Estrutura da Dissertação

Esta dissertação é composta de cinco capítulos. No primeiro capítulo, Introdução, foram apresentadas as motivações do estudo e seus objetivos.

No segundo capítulo, Revisão Bibliográfica, são explicados os mecanismos de formação de ozônio troposférico, bem como a teoria de cada um dos métodos de Estatística Multivariada aplicados na predição do poluente.

No terceiro capítulo, Materiais e Métodos, são apresentadas a origem e a estrutura dos dados utilizados e nas técnicas de predição e a metodologia empregada nas simulações. Neste capítulo também são definidos os indicadores de desempenho para comparação dos modelos e a análise de sensibilidade aplicada para avaliar a importância das variáveis de entrada.

No quarto capítulo, Resultados e Discussão, são apresentados a comparação do desempenho dos modelos e os melhores ajustes obtidos, além da análise dos indicadores de desempenho utilizados através do teste t de amostras pareadas.

No quinto e último capítulo, Conclusões, apresenta-se a síntese dos resultados do estudo além de sugestões para próximos estudos.

2 Revisão Bibliográfica

2.1 Ozônio Troposférico

O acúmulo de ozônio na troposfera poderia ser explicado por dois eventos diferentes: transporte vertical da estratosfera, onde se localiza a camada de ozônio, ou produção *in situ*.

Estima-se que o fluxo global de O_3 da estratosfera para a troposfera seja $1-2\times 10^{13}$ mol O_3 por ano. Cada molécula de O_3 pode participar de uma reação de fotólise e produzir duas moléculas de OH, resultando em, no máximo, $2-4\times 10^{13}$ mols de $OH^{\bullet}$ por ano na troposfera.

A fotólise do O_3 (através de radiação com comprimento de onda $\lambda < 319nm$) produz O_2 e um radical de oxigênio, que pode tanto estar no seu estado fundamental (O) quanto no seu estado singlete ($O(^1D)$).

$$O_3 + hv \longrightarrow O_2 + O$$
$$O_3 + hv \longrightarrow O_2 + O(^1D)$$

O átomo no estado fundamental se combina rapidamente com O_2 para reformar o O_3, na presença de uma terceira molécula (M). Porém o oxigênio singlete pode reagir com com uma molécula de H_2O e produzir dois radicais OH, como descrito abaixo.

$$O + O_2 + M \longrightarrow O_3 + M$$
$$O(^1D) + H_2O \longrightarrow 2\,OH^{\bullet}$$

Sabe-se que a principal fonte de $OH^{\bullet}$ é a cadeia de reações acima, iniciadas pela fotólise do O_3, e que os principais sumidouros de $OH^{\bullet}$ da troposfera são o CH_4 e o CO.

Para se manter o balanço troposférico de CH_4 e o CO seriam necessários aproximadamente 10^{14} mol $OH^{\bullet}$ por ano. Se todo o O_3 fornecido pelo fluxo estratosfera-troposfera é capaz de originar apenas $2-4\times 10^{13}$ mols de $OH^{\bullet}$ por ano, pode-se inferir que a parcela restante de $OH^{\bullet}$ é produzida a partir de O_3 que é formado na própria troposfera. Portanto, a formação e consumo de O_3 *in situ* domina o balanço de O_3 troposférico.

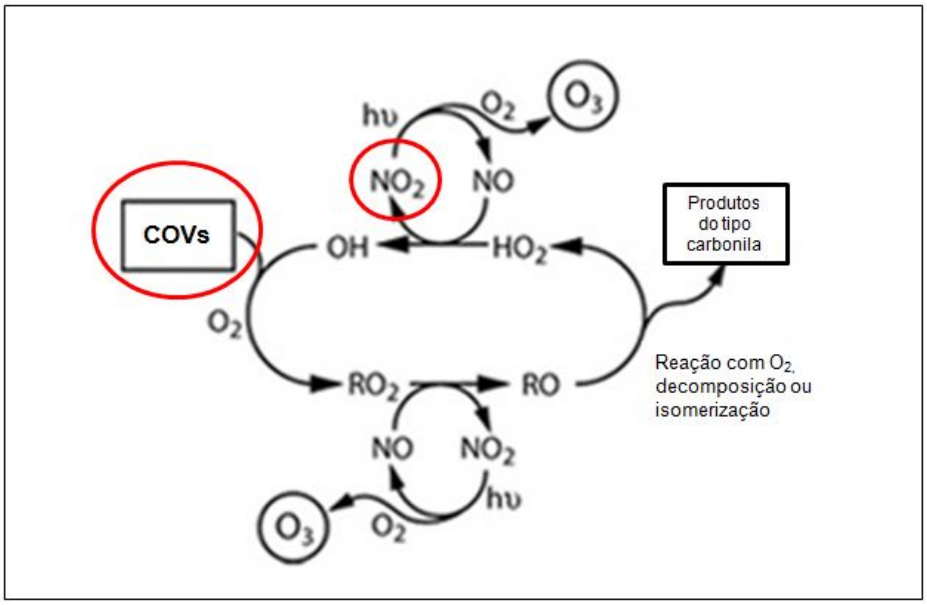

Figura 2.1: Um diagrama de oxidação de COVs na Troposfera adaptado de [Clapp and Jenkin, 2001] (traduzido para o português) [Man, 2017]

A formação de ozônio *in situ*, ou seja, na baixa atmosfera, pode ser descrita de uma forma simplificada a partir da decomposição de dióxido de nitrogênio (NO_2) e de compostos orgânicos voláteis (COVs), conforme mostra o diagrama da Figura 2.1.

Por uma reação fotoquímica com a luz solar, o NO_2 gera o átomo de oxigênio que, ao se combinar com as moléculas de oxigênio presentes no ar, dá origem ao ozônio de acordo com as reações seguintes [Finlayson-Pitts and Pitts Jr., 2012, CETESB, 2000, Seinfeld and Pandis, 2006]:

$$NO_2 + hv \xrightarrow{\lambda <= 424nm} NO + O^{\bullet}$$
$$O^{\bullet} + O_2 + M \longrightarrow O_3 + M$$

Em que M é uma terceira molécula responsável pela estabilização do intermediário formado pela adição de O com O_2.

Vale notar que a reação fotoquímica ocorre na faixa da radiação da cor violeta e logo abaixo da região visível (raios UV), como mostra a Figura 2.2.

O NO, gerado como produto da fotólise do NO_2 reage rapidamente com o ozônio formado para regenerar o NO_2 [CETESB, 2000], segundo a reação abaixo:

$$O_3 + NO \longrightarrow NO_2 + O_2$$

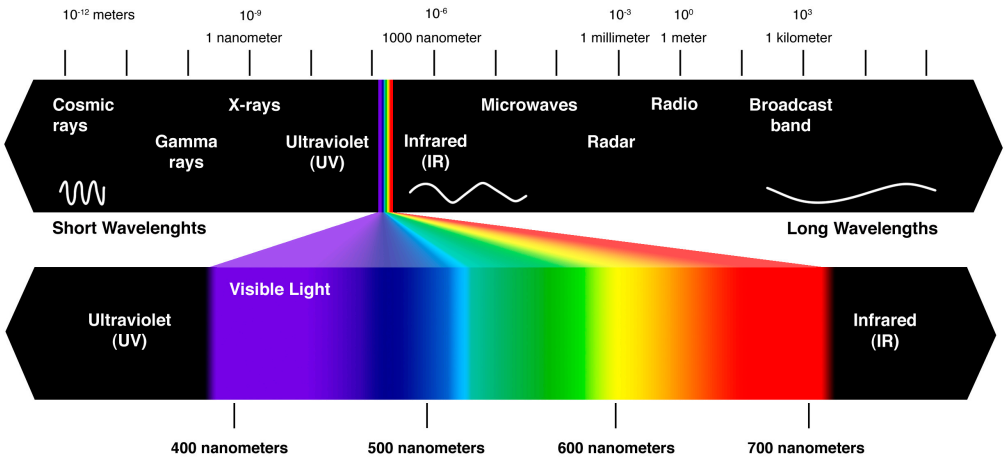

Figura 2.2: Região do espectro visível e de raios UV

Dessa maneira, o O_3 mantém-se num estado estacionário que depende da velocidade de fotólise do NO_2 e da razão $[NO_2]/[NO]$. Assim, se nenhum outro processo convertesse NO em NO_2 a concentração de ozônio não aumentaria significativamente [CETESB, 2016a]. A Figura 2.3 apresenta um diagrama da reação fotoquímica envolvendo a produção e das reações envolvendo o consumo de O_3. Já a Figura 2.4 apresenta o comportamento típico ao longo do dia desses poluentes na RMSP. As concentrações horárias são uma média de todos os dias de que se tem medidas na estação Ibirapuera em Janeiro de 2017.

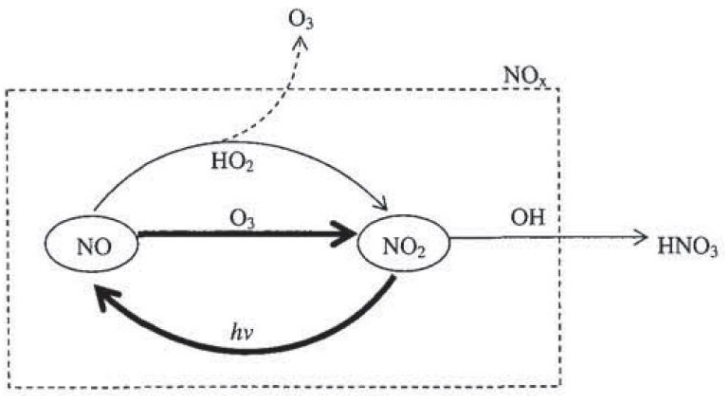

Figura 2.3: Reações envolvendo NO_x [Seinfeld and Pandis, 2006]

Percebe-se que a concentração de NO_x diminui à medida que a concentração de ozônio aumenta. Os picos de O_3 geralmente ocorrem no meio da tarde. Ao final da tarde, devido à falta de radiação, prevalece a reação de O_3 com NO, o que provoca um aumento da concentração de NO_2.

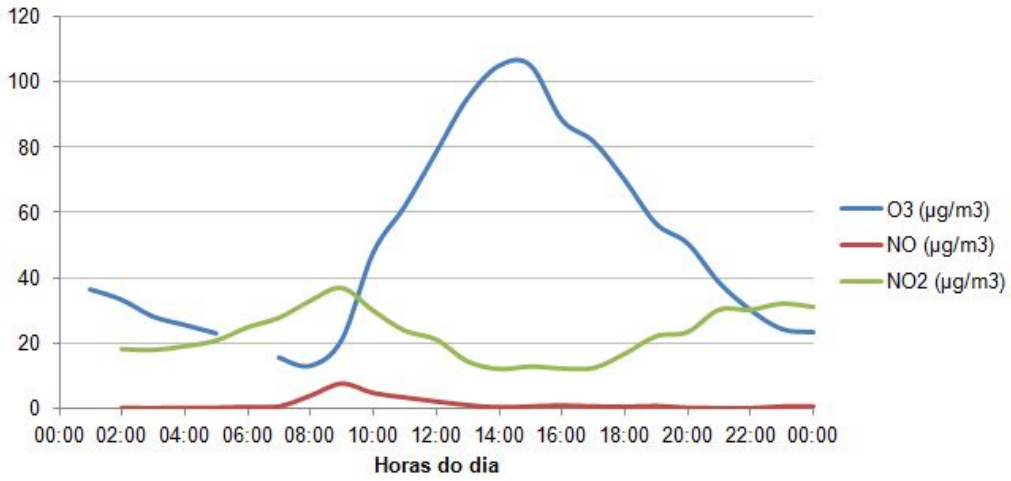

Figura 2.4: Perfil de concentrações de poluentes ao longo do dia; dados provenientes do período de Janeiro de 2017 da estação Ibirapuera da CETESB.

Conforme dito anteriormente, se nenhum processo convertesse NO em NO_2, a concentração de O_3 se manteria aproximadamente estável ao longo do dia. No entanto, quando os COVs estão presentes no ambiente, eles reagem para formar radicais que ou consomem NO, dificultando sua reação com O_3, ou convertem NO em NO_2 [Carter, 1994].

Os COVs são definidos pela CETESB como gases e vapores resultantes da queima incompleta e evaporação de combustíveis e de outros produtos orgânicos [CETESB, 2016a]. A União Européia define os COVs como qualquer composto orgânico com ponto de ebulição menor do que 250 graus Celsius na pressão de 101,3 kPa [EU, 2004]. Eles serão representados aqui pela molécula de metano (CH_4). A reação dos COVs com radicais hidroxil presentes na atmosfera gera intermediários que aceleram a prodrução de ozônio, conforme mostra a sequência de reações abaixo.

$$CH_4 + OH^{\bullet} \xrightarrow{O_2} CH_3O_2 + H_2O$$
$$CH_3O_2 + NO \xrightarrow{O_2} HCHO + HO_2 + NO_2$$
$$HCHO + hv \xrightarrow{O_2} CO + HO_2$$
$$3\,(HO_2 + NO \longrightarrow NO_2 + OH)$$
$$4\,(NO_2 + hv \xrightarrow{O_2} NO + O_3)$$

$$CH_4 + 8\,O_2 \longrightarrow CO + H_2O + 2\,OH + 4\,O_3$$

A oxidação da molécula de CO gera mais uma molécula de O_3, portanto a

máxima produção de ozônio a partir de CH_4 é dada pela reação abaixo.

$$CH_4 + 10\,O_2 \longrightarrow CO_2 + H_2O + 2\,OH^{\bullet} + 5\,O_3$$

Ou seja, cada molécula de CH_4 é capaz de gerar outras 5 moléculas de O_3. Essa produção máxima de ozônio ocorre somente quando os níveis de NO_x são suficientemente altos para que o CH_3O_2 reaja apenas com o NO.

Nas reações acima, o CH_4 foi utilizado como representante dos COVs. No entanto, na prática os COVs com maior potencial de formação de O_3 são aldeídos, formaldeídos e xilenos [Martins et al., 2008].

De forma simplificada, a chave para a formação de ozônio depende de dois fatores: a disponibilidade de radicais hidroxil na atmosfera ($OH^{\bullet}$) e a razão entre as concentrações de COV e NO_x.

Se a razão [COV]/[NO_x] é mais baixa que um determinado limite, haverá excesso de NO_2 na atmosfera, que reagirá com os radicais hidroxil produzindo HNO_3.

$$OH^{\bullet} + NO_2 + M \longrightarrow HNO_3 + M$$

Essa reação remove tanto o reagente NO_2, precursor direto do ozônio, quanto os radicais $OH^{\bullet}$, que dão início à cadeia de oxidação de COVs.

Se, por outro lado, a razão [COV]/[NO_x] for tal que haja excesso de COVs, eles reagirão com os radicais hidroxil, dando início à cadeia de oxidação e, portanto, ao ciclo de formação de ozônio. Ou seja, para um determinado valor de [COV], existe uma concentração [NO_x] tal que o máximo possível de ozônio é gerado.

A concentração dessas espécies químicas na troposfera depende de fatores como a taxa de emissão, que ocorre principalmente a partir de fontes móveis imensuráveis, e a dispersão provocada pelas diferenças de pressão locais e pela velocidade e direção do vento, além da intensidade da radiação no ambiente, que impacta diretamente a formação de O_3.

Dada a complexidade de modelar esse sistema, os modelos determinísticos, baseados no equacionamento das relações químicas e físicas, ainda não são entendidos em profundidade, o que limita sua aplicação para fins de previsão. Além disso, há dificuldades adicionais, pois a frota de veículos na RMSP é baseada em gasolina e etanol como combustíveis, o que torna mais complexo o ciclo fotoquímico.

2.2 Métodos de Estatística Multivariada/Aprendizado de Máquina

A definição exata das áreas de estudo de Estatística Multivariada e Aprendizado de Máquina (*Machine Learning*) ainda é nebulosa e, portanto, no presente estudo, estes termos são usados como sinônimos, dado que ambas as áreas de estudo abordam as técnicas aplicadas.

Os modelos de estatística multivariada são aplicados para definir uma função ou um modelo que representa um fenômeno a partir de uma base de dados. A planilha de dados é normalmente organizada como apresentado a seguir: cada linha representa uma observação diferente e cada coluna representa uma variável que afeta o fenômeno. A coluna final contém as respostas que se deseja obter. No exemplo abaixo a coluna à direita contém as variáveis de saída (respostas).

Observações	x_1	x_2	...	x_P	y
1	.	.	.	.	.
2	.	.	.	.	.
...	.	.	.	.	.
N	.	.	.	.	.

Comumente, os modelos de estatística multivariada são aplicados a dois casos:

1. Regressão: obtenção de um ou mais valores contínuos a partir dos dados. Neste caso y é uma variável contínua.

2. Classificação: obtenção de uma classe para cada observação. Neste caso y é um vetor de valores 0 e 1 e de tamanho 1 $\times$ K, onde K representa o número de classes possíveis.

Observações	x_1	x_2	...	x_P	y_1	y_2	...	y_K
1	.	.	.	.	1	0	0	0
2	.	.	.	.	0	0	0	1
...	.	.	.	.	0	1	0	0
N	.	.	.	.	1	0	0	0

As observações da base de dados utilizada são normalmente dividas em dois conjuntos: *Training-set* e *Test-set*. O primeiro é utlizado para ajuste do modelo e o segundo é utilizado para verificar a eficácia deste modelo na previsão do fenômeno.

Todos os modelos utilizados, apresentados pela Figura 2.5, são explicados nas subseções seguintes.

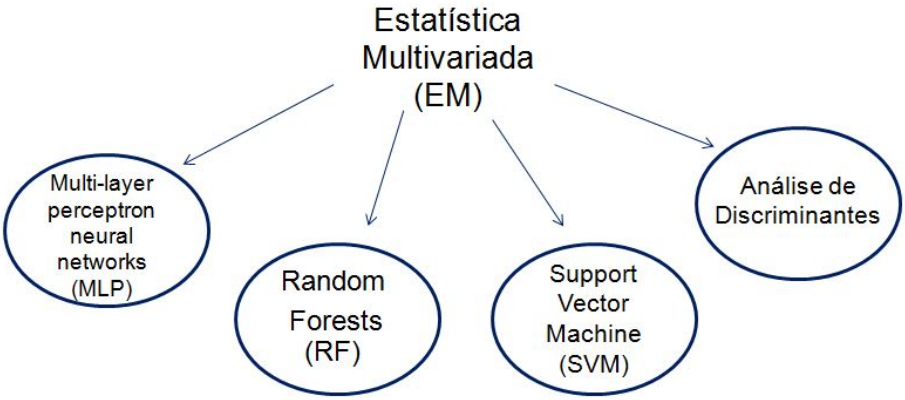

Figura 2.5: Modelos de estatística multivariada aplicados no estudo

2.2.1 *Multi-layer Perceptron Neural Networks* (MLP)

O método de MLP é uma ferramenta de modelagem poderosa da família das redes neurais com comprovada eficácia no tratamento de problemas complexos [Ordieres et al., 2005]. O método tem sido aplicado em diversos estudos de previsão de concentração de poluentes atmosféricos com resultados satisfatórios, sendo capaz de capturar as relações não lineares e complexas entre variáveis meteorológicas e os níveis de poluentes na atmosfera, sem detalhar as causas destas relações.

A técnica de MLP consiste na construção de camadas contendo neurônios, que são unidades de processamento matemático das informações providas, cujas saídas são obtidas segundo diferentes funções resposta, como mostra a Figura 2.6.

A primeira camada, ou camada de entrada, é composta pelas variáveis de entrada (*input*) extraídas da base de dados, portanto o número de neurônios desta camada será igual ao número de variáveis da base para cada observação.

A última camada, ou camada de saída, é composta por neurônios que representam as variáveis de resposta que se deseja obter. Para previsão de níveis de ozô-

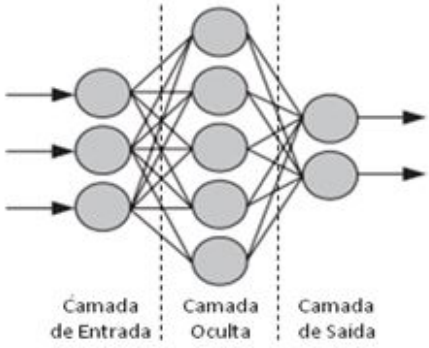

Figura 2.6: Representação de um MLP

nio, por exemplo, a camada de saída contém apenas um neurônio, que é o neurônio que expressa a concentração dessa espécie química.

A camada intermediária, chamada de camada oculta, é composta por tantos neurônios quantos se queiram. Cada um deles corresponde a um valor númerico gerado através da aplicação de uma função de ativação, ou função de transição, à combinação linear dos dados da camada de entrada.

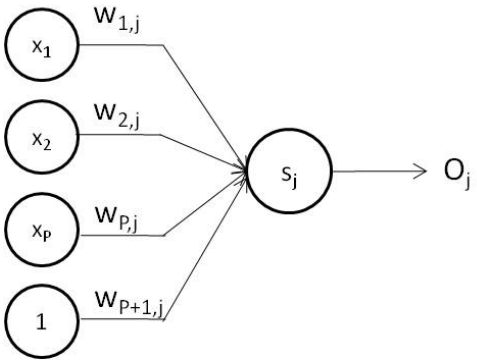

Figura 2.7: Representação de um neurônio j da estrutura da rede

A construção deste tipo de rede neural é feita aplicando-se um peso $w_{p,j}$ à saída de cada neurônio x_p da camada de entrada e somam-se todos os valores. Se o número de neurônios da camada de entrada for P, a soma conterá $P+1$ termos, onde o termo $P+1$ consiste do número 1 vezes um peso $w_{P+1,j}$. A este termo dá-se o nome de *bias*. A Equação 2.1 mostra o cálculo aplicado:

$$S_j = w_{P+1,j} + \sum_{p=1}^{P} w_{p,j}.x_p \ . \tag{2.1}$$

O valor S_j obtido é aplicado a uma função de ativação $f(S_j)$, conforme Equação 2.2, e o resultado O_j da função é utilizado para gerar os valores da camada seguinte num processo semelhante. A Figura 2.7 demonstra o processo de formação de um neurônio.

$$O_j = f(S_j) \ . \tag{2.2}$$

Funções de ativação comuns são a função sigmoidal, linear e radial, representadas respectivamente pelas Equações 2.3, 2.4 e 2.5. Costuma-se utilizar a mesma função de ativação para todos os neurônios de uma mesma camada. A escolha da função de ativação que melhor se ajusta ao modelo depende fortemente da base de dados. Neste estudo, as funções de ativação aplicadas foram a sigmoidal e linear por terem gerado resultados significativamente melhores do que a radial em testes preliminares:

$$f(z) = \frac{1}{1 + e^{-z}} \ , \tag{2.3}$$

$$f(z) = a + bz \ , \tag{2.4}$$

$$f(z) = e^{-z^2/2\sigma^2} \ , \tag{2.5}$$

onde a, b e σ são números escalares constantes.

As redes *feed-forward*, ou redes de sentido único, aquelas em que as entradas dos neurônios de uma camada são compostas apenas pelos valores das saídas dos neurônios da camada anterior, são as mais comuns. Normalmente existem apenas três camadas: camada de entrada, camada de saída e camada intermediária, como mostra a Figura 2.8, mas o número de camadas ocultas pode variar de zero a um número inteiro qualquer e o número de neurônios em cada camada oculta também pode ser definido conforme a conveniência.

A resposta da rede representada pela Figura 2.8 para uma observação n, considerando que a função de ativaçao $f(z)$ é a mesma para todos os neurônios da camada oculta e da camada de saída, é dada pela Equação 2.6:

$$y_{k,calc}^{(n)} = f\left(w_{J+1,k}^{(2)} + \sum_{j=1}^{J} w_{j,k}^{(2)}.f\left(w_{P+1,j}^{(1)} + \sum_{p=1}^{P} w_{p,k}^{(1)}.x_p\right)\right) \ . \tag{2.6}$$

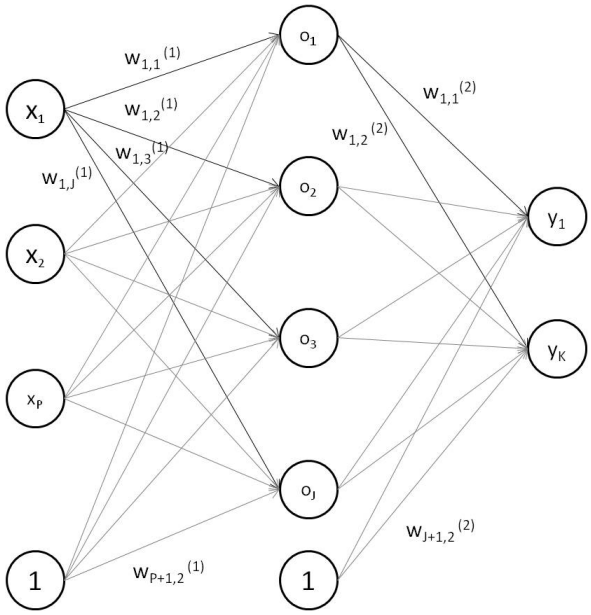

Figura 2.8: Representação de um MLP com pesos e bias

O treinamento (ou aprendizado) do MLP consiste de uma adaptação dos pesos $(w_{i,j})$ aplicados sobre as informações entre as camadas de neurônios, de forma a minimizar uma função objetivo, que pode ser a média dos erros quadráticos (Equação 2.7), geralmente utilizada nos casos de regressão, ou a entropia cruzada (*cross-entropy*, Equação 2.10), utilizadas nos casos de classificação.

A média dos erros quadráticos (MSE) representa o valor acumulado para todas as observações do *training-set* (de 1 a N), das somas dos erros quadráticos entre os valores experimentais e os calculados pelo MLP, para todas as saídas da rede (de 1 a K):

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(y_{k,calc}^{(n)} - y_{k,exp}^{(n)})^2 \ . \tag{2.7}$$

A função entropia cruzada (E) é obtida a partir do conceito de verossimilhança (L), que representa a probabilidade de se obter o conjunto de respostas da base de dados, dadas as variáveis de *input*, ou seja:

$$L = \prod_{n=1}^{N}\prod_{k=1}^{K} p(k|X^{(n)})^{y_{k,exp}^{(n)}} \ . \tag{2.8}$$

Se aplicarmos o log natural na função de verossimilhança multiplicada por -1, teremos a entropia cruzada, que é dada pela Equação 2.9 [Ripley, 1996]:

$$E = (-\ln L) = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{k,exp}^{(n)} \ln p(k|X^{(n)}) \ . \tag{2.9}$$

Na prática, porém, utiliza-se a definição de Entropia Cruzada dada pela Equação 2.10, conforme explica Bishop (1995):

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{k,exp}^{(n)} \ln \frac{p(k|X^{(n)})}{y_{k,exp}^{(n)}} \ , \tag{2.10}$$

onde a probabilidade condicional de uma observação n pertencer à classe k, $p(k|X^{(n)})$, é dada pela função *Softmax* [Bridle, 1989], como mostra a Equação 2.11:

$$p(k|X^{(n)}) = \frac{\exp y_{k,calc}^{(n)}}{\sum_{k=1}^{K} \exp y_{k,calc}^{(n)}} \ . \tag{2.11}$$

Para se obter o conjunto de pesos (w), é preciso, em primeiro lugar, que o conjunto de dados que constituem o *training-set* seja escolhido de tal forma que se contemplem todas as informações relevantes do processo. Tendo esse conjunto de dados, utiliza-se um algoritmo de otimização, a fim de se obter o conjunto de pesos que minimize a função objetivo dada pela Equação 2.7 ou pela Equação 2.10.

Levenberg-Marquardt e *Scaled Conjugated Gradient Backpropagation* são dois algorítmos de otimização mais comuns utilizados pelo software MATLAB para os casos de regressão e classificação, respectivamente.

Todos os algorítmos citados fazem uso das derivadas da função objetivo em relação a cada um dos pesos da rede $\left(\frac{\partial MSE}{\partial w_{p,j}^{(i)}} \quad \text{ou} \quad \frac{\partial E}{\partial w_{p,j}^{(i)}}\right)$. As derivadas parciais são obtidas através do algoritmo *back-propagation*, que foi popularizado no artigo de Rumelhart, Hinton e Williams (1986). O método detalhado para várias funções objetivo foi descrito por Ripley (1996).

2.2.2 Árvores de Decisão (CART) e *Random Forests* (RF)

As Árvores de Decisão (CART - *Classification and Regression Trees*) são ferramentas relativamente simples para classificação de dados. Existem dois tipos de árvores de decisão: as utilizadas em problemas de classificação e em problemas de regressão. A partir de um conjunto de dados, definem-se separadores (nós ou "*splitters*"), baseados em valores das variáveis de previsão envolvidas no sistema, de forma a separar, sequencialmente, grupos de indivíduos semelhantes. As Figuras 2.9 e 2.10 ilustram um exemplo de árvore de classificação.

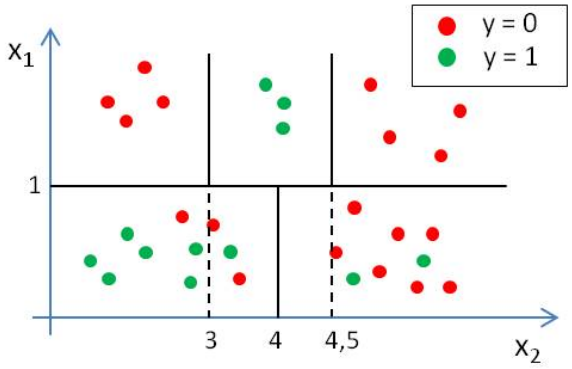

Figura 2.9: Distribuição gráfica de dados bidimensionais ($X \in \Re^2$) para um problema de classificação

O objetivo das técnicas baseadas em árvores de decisão é minimizar o erro em cada grupo final, também denominado folha. Para o caso de classificação, deseja-se minimizar a soma ponderada da entropia de cada folha, de modo que cada folha tenha o máximo possível de uniformidade. A entropia de uma folha f é dada pela Equação

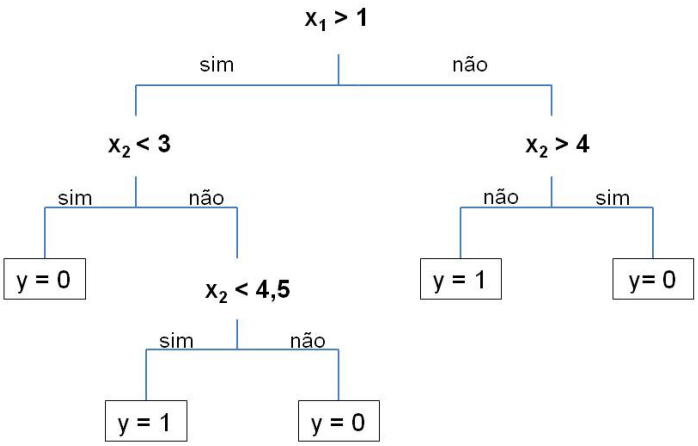

Figura 2.10: Estrutura da árvore de decisão

2.12:

$$E_f = -\sum_{k=1}^{K} \frac{N_{k,f}}{N_f} \log \frac{N_{k,f}}{N_f} \ , \tag{2.12}$$

onde N_f é o número de observações em uma folha e $N_{k,f}$ é o número de observações pertencentes à classe k na folha.

A soma ponderada das entropias de cada folha, ou a entropia total, para uma árvore com F folhas é dada pela Equação 2.13:

$$E = \sum_{f=1}^{F} \frac{N_f}{N} . E_f \ , \tag{2.13}$$

onde N é o número de observações no *training-set*.

Assim, para árvores de classificação, a menor entropia ocorre quando cada folha contém o máximo de observações pertencentes ao mesmo grupo. No processo de construção da árvore, a cada nó escolhe-se uma variável x_p e um valor para essa variável que dividirá o grupo original em dois outros grupos de forma que a entropia do estado final seja a menor possível. Essa divisão ocorre até que seja alcançado um critério de parada ou até que exista apenas uma observação em cada folha (crescimento máximo). O rótulo final de cada folha é igual à classe k com o maior número de representantes nessa folha.

É importante notar que quanto mais a árvore se aproxima do seu crescimento máximo, mais ela tende a ficar super ajustada aos dados do *training-set* (*overfitting*) e pior será a previsão para as observações do *test-set*. Portanto é preciso encontrar um equilíbrio entre o crescimento da árvore e a minimização da entropia.

No caso de árvores de regressão, deseja-se obter o menor desvio quadrático possível da média da folha para cada folha. Ou seja, o objetivo é minimizar a expressão dada pela Equação 2.14:

$$E = \sum_{f=1}^{F} \sum_{n=1}^{N_f} \left(y_f^{(n)} - \overline{y}_f \right)^2 \ , \tag{2.14}$$

onde cada $y_f^{(n)}$ é o próprio valor experimental de y para a observação n alocada na folha f. Na construção das árvores de regressão, não se fazem estimativas para o valor y, uma vez que o que se deseja otimizar são os nós, ou as divisões, que compõem a

árvore. O valor de y para uma nova observação será igual ao valor da média da folha ($\overline{y}_f$) em que essa observação for alocada.

Árvores de decisão, em geral, são consideradas técnicas fracas de aprendizagem, uma vez que são altamente dependentes do conjunto de dados iniciais e podem apresentar elevada variância com alterações nesses dados. Para solucionar este problema utiliza-se a técnica de RF, que consiste na combinação de múltiplas árvores de decisão. Esta técnica baseia-se no princípio de que várias técnicas de aprendizagem fracas combinadas podem gerar um modelo de aprendizagem forte [Breiman, 1999].

O algoritmo de *Random Forests* possui duas etapas bem definidas:

1. *Bootstrap Aggregation* ou *Bagging*: Criação de B *subsets* da base original (*training-set*) por amostragem com reposição

2. Construção das árvores de decisão para cada *subset* criado

Na primeira etapa do algoritmo, para um *training-set* com N observações, constroem-se B outros subconjuntos de dados com a mesma quantidade N de observações através de amostragem aleatória com reposição. A este procedimento dá-se o nome de *bagging*. O número B de *subsets* a serem gerados pode ser escolhido conforme a conveniência. Quanto maior o número de *subsets*, maior a acurácia do modelo e, da mesma forma, maior o custo computacional.

Na segunda etapa, constroi-se uma árvore de decisão para cada *subset* gerado. As árvores devem ser construídas até sua máxima extensão, ou seja, até que haja apenas uma observação em cada folha [Breiman, 2001]. Durante a construção da árvore, para cada nó seleciona-se aleatoriamente um subconjunto m de variáveis x_p dentre as P variáveis existentes. A decisão de que variável será utilizada como critério de separação do nó será feita com base no subconjunto de variáveis selecionadas. Assim, para uma base de dados cujas observações são descritas por $X = [x_1, x_2, ..., x_P]$, selecionam-se aleatoriamente, por exemplo, duas variáveis ($m = 2$): x_1 e x_P. A decisão de que separador utilizar neste nó específico deve levar em consideração apenas x_1 e x_P. Esse processo de seleção aleatória de variáveis é feito em todos os nós de todas as árvores.

A definição dos valores B (número de *subsets* gerados) e m (número de variáveis selecionadas aleatoriamente) é feita conforme a conveniência e o desempenho

das diferentes estruturas do modelo.

Obtém-se, assim, uma floresta de árvores construídas de forma aleatória, onde cada nova observação a ser classificada passa por todas as árvores de decisão. Em *Random Forests* não ocorre *overfitting* mesmo quando o número de árvores é muito elevado. Esta constatação foi feita por Breiman (2001), que demonstrou, usando a Lei dos Grandes Números, que a probabilidade de haver mais acertos do que erros de classificação converge para um número determinado quando o número de árvores cresce.

Para problemas de classificação, a observação será alocada no grupo com a maior quantidade de votos e para problemas de regressão, o valor y da nova observação será uma média dos valores previstos por árvore.

2.2.3 Análise de Discriminantes

Discriminante Linear de Fisher

A técnica de LDA conhecida como Discriminante Linear de Fisher é utilizada para problemas de classificação de observações. O objetivo é encontrar uma combinação linear de todas as variáveis $\mathbf{X}_{p\times 1} = [x_1, x_2, ..., x_P]$ de forma que o valor resultante dessa combinação, chamado discriminante, tenha valores muito diferentes para observações pertencentes a classes diferentes e valores próximos para observações da mesma classe [Hastie et al., 2008]. Ou seja, o discriminante d é um valor escalar obtido pelo produto escalar entre vetor das variáveis X e um vetor de pesos a para cada uma das observações, como mostra a Equação 2.15:

$$d = \mathbf{a}' \cdot \mathbf{X} \,. \tag{2.15}$$

O discriminante, portanto, pode ser visto como a projeção dos dados sobre um eixo de direção dada pelo vetor **a**, sendo este eixo escolhido de tal forma que as projeções dos dados pertencentes a grupos diferentes tenham a maior separação possível, como mostra a Figura 2.11.

Dada umas base de dados com as seguintes características:

K - número total de classes, ou grupos, existentes;

n_k - número total de observações no grupo k;

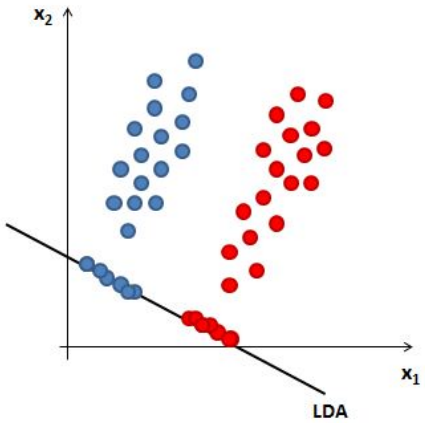

Figura 2.11: Representação gráfica dos discriminantes para dados bidimensionais

$\overline{\mathbf{X}}_k$ - vetor $p \times 1$ das médias das variáveis para as observações dentro do grupo k;

$\overline{\mathbf{X}}$ - vetor $p \times 1$ das médias das variáveis para todas as observações.

Define-se a matriz $\mathbf{B}_{p\times p}$ que representa a soma ponderada das distâncias quadráticas entre grupos, dada pela Equação 2.16:

$$\mathbf{B} = \sum_{k=1}^{K} n_k(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})' \ . \tag{2.16}$$

Adicionalmente, define-se a matriz $\mathbf{W}_{p\times p}$, que representa a soma da distância quadrática entre as observações $\mathbf{X}_k^{(n)}$ de um grupo k e a média $\overline{\mathbf{X}}_k$ deste grupo, somadas para todos os grupos, como mostra a Equação 2.17:

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{n=1}^{n_k} (\mathbf{X}_k^{(n)} - \overline{\mathbf{X}}_k)(\mathbf{X}_k^{(n)} - \overline{\mathbf{X}}_k)' \ . \tag{2.17}$$

Dado que o objetivo é separar o máximo possível os determinantes d de cada grupo, deve-se maximizar a distância das médias dos determinantes de cada grupo e minimizar a distância da dos determinantes das observações dentro de um mesmo grupo. Ou seja, deve-se maximizar a razão dada por σ na Equação 2.22:

$$\sigma = \frac{\sum_{k=1}^{K} n_k(\overline{d}_k - \overline{d})^2}{\sum_{k=1}^{K} \sum_{n=1}^{n_k} (d_k^{(n)} - \overline{d}_k)^2} \ , \tag{2.18}$$

$$\sigma = \frac{\sum_{k=1}^{K} n_k(\mathbf{a}' \cdot \overline{\mathbf{X}}_k - \mathbf{a}' \cdot \overline{\mathbf{X}})^2}{\sum_{k=1}^{K} \sum_{n=1}^{n_k} (\mathbf{a}' \cdot \mathbf{X}_k^{(n)} - \mathbf{a}' \cdot \overline{\mathbf{X}}_k)^2} \ , \tag{2.19}$$

$$\sigma = \frac{\sum_{k=1}^{K} n_k \mathbf{a}'(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})'\mathbf{a}}{\sum_{k=1}^{K}\sum_{n=1}^{n_k} \mathbf{a}'(\mathbf{X}_k^{(n)} - \overline{\mathbf{X}}_k)(\mathbf{X}_k^{(n)} - \overline{\mathbf{X}}_k)'\mathbf{a}} \ , \tag{2.20}$$

$$\sigma = \frac{\mathbf{a}'\left(\sum_{k=1}^{K} n_k(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})'\right)\mathbf{a}}{\mathbf{a}'\left(\sum_{k=1}^{K}\sum_{n=1}^{n_k} (\mathbf{X}_k^{(n)} - \overline{\mathbf{X}}_k)(\mathbf{X}_k^{(n)} - \overline{\mathbf{X}}_k)'\right)\mathbf{a}} \ , \tag{2.21}$$

$$\sigma = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} \ . \tag{2.22}$$

O vetor **a** procurado é o obtido a partir da Equação 2.23:

$$\mathbf{a} = arg\,\max \sigma = arg\,\max \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} \ . \tag{2.23}$$

Derivando-se σ tem-se:

$$\frac{\partial \sigma}{\partial \mathbf{a}} = \frac{2\mathbf{B}\mathbf{a}(\mathbf{a}'\mathbf{W}\mathbf{a}) - 2(\mathbf{a}'\mathbf{B}\mathbf{a})\mathbf{W}\mathbf{a}}{(\mathbf{a}'\mathbf{W}\mathbf{a})^2} = 0 \ . \tag{2.24}$$

Aplica-se então a relação dada pela Equação 2.22, resultando nas seguintes equações:

$$\frac{\mathbf{B}\mathbf{a}(\mathbf{a}'\mathbf{W}\mathbf{a}) - \sigma(\mathbf{a}'\mathbf{W}\mathbf{a})\mathbf{W}\mathbf{a}}{(\mathbf{a}'\mathbf{W}\mathbf{a})^2} = 0 \ , \tag{2.25}$$

$$\mathbf{B}\mathbf{a}(\mathbf{a}'\mathbf{W}\mathbf{a}) - \sigma(\mathbf{a}'\mathbf{W}\mathbf{a})\mathbf{W}\mathbf{a} = 0 \ , \tag{2.26}$$

$$(\mathbf{B} - \sigma\mathbf{W})\mathbf{a} = 0 \ , \tag{2.27}$$

$$(\mathbf{W}^{-1}\mathbf{B} - \sigma I)\mathbf{a} = 0 \ . \tag{2.28}$$

O problema inicial resume-se, desta forma, a um problema de autovalores e autovetores em relação à matriz $\mathbf{W}^{-1}\mathbf{B}$. As soluções não triviais deste problema correspondem aos vetores de pesos **a** procurados. Pode existir, para este problema, mais de uma solução, isso significa que existe mais de um eixo sobre o qual projetar os dados de forma que as condições propostas sejam satisfeitas.

Ao invés de se selecionar apenas uma das soluções para o cálculo do discrimi-

nante e alocação de novas observações a um dos grupos, utilizam-se todos os pesos $\mathbf{a}_i$ obtidos como resposta.

Desta forma, considerando um nova observação $\mathbf{X}^{(0)}$ não contida na base de dados e considerando-se que a resolução da Equação 2.28 tenha gerado D respostas possíveis ($\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_i$, ..., $\mathbf{a}_D$) a alocação em um dos grupos ocorre segundo os passos abaixo:

1. Calculam-se os valores de todos os discriminantes possíveis para $\mathbf{X}^{(0)}$, um para cada peso $\mathbf{a}_d$ encontrado ($d_1^{(0)}$, $d_2^{(0)}$,..., $d_D^{(0)}$)

2. Calculam-se os valores de todos os discriminantes possíveis para as médias $\overline{\mathbf{X}}_k$ de todos os grupos k ($\overline{d}_{1,k}$, $\overline{d}_{2,k}$, ..., $\overline{d}_{1,k}$)

3. Aloca-se a observação $\mathbf{X}^{(0)}$ no grupo para o qual a distância entre o discriminante da observação e o discriminante da média do grupo é a menor, segundo a Equação 2.29:

$$\text{min} \sum_{i=1}^{D} (d_i^{(0)} - \overline{d}_{i,k})^2 \ , \tag{2.29}$$

calculada para todos os grupos $k = 1, 2, ..., K$.

O número total (D) de discriminantes, ou seja, de pesos $\mathbf{a}_i$ que podem ser obtidos é igual a

$$D = min(K - 1; P) \ , \tag{2.30}$$

onde P é o número de variáveis que descrevem o problema, ou seja, o tamanho do vetor $\mathbf{X}$.

Discriminante Linear (LDA)

A técnica de LDA se baseia em duas hipóteses:

1. Todas as populações, que dão origem aos grupos de dados, possuem curva de distribuição $f_k(\mathbf{x})$ normal;

2. As matrizes de covariância de cada grupo de dados, pertecentes a diferentes populações, são iguais. Ou seja, $\Sigma_k = \Sigma$.

A técnica se baseia simplesmente na maximização das probabilidades *a posteriori* $Pr(K = k|X = \mathbf{x})$, ou seja, a probabilidade de uma observação pertencer a um grupo k, dado que ela é definida por x. Assim, pelo Teorema de Bayes, tem-se

$$Pr(K = k|X = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{x})\pi_l} \ , \tag{2.31}$$

onde:

K - número total de classes, ou grupos, existentes;

π_k - probabilidade *a priori* de uma observação pertencer ao grupo k.

Conforme a hipótese inicial, todas as populações possuem distribuição normal, que é definida pela Equação 2.32:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k)\right) \ , \tag{2.32}$$

onde:

μ_k - média da população k;

p - dimensão do vetor **x**;

Σ_k - matriz de covariância da população k.

Ao se comparar as probabilidades *a posteriori* de duas classes diferentes, tem-se uma equação linear em **x**, como mostram as Equações 2.33 e 2.34:

$$\log \frac{Pr(K = k|X = \mathbf{x})}{Pr(K = l|X = \mathbf{x})} = \log \frac{f_k(\mathbf{x})}{f_l(\mathbf{x})} + \log \frac{\pi_k}{\pi_l} \ , \tag{2.33}$$

$$\log \frac{Pr(K = k|X = \mathbf{x})}{Pr(K = l|X = \mathbf{x})} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)'\Sigma^{-1}(\mu_k - \mu_l) + \mathbf{x}'\Sigma^{-1}(\mu_k - \mu_l) \ . \tag{2.34}$$

Como na prática não se conhecem os parâmetros das populações, eles são estimados pelos dados do *training set*:

1. $\hat{\pi_k} = \frac{n_k}{N}$ onde n_k é o número de observações no grupo k e N é o número total de observações no *training set*;

2. $\hat{\mu_k} = \sum_{i=1}^{n_k} x_i/n_k$;

3. $\hat{\Sigma} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_i - \hat{\mu_k})(x_i - \hat{\mu_k})/(N - K)$.

A Equação 2.34 implica que o limite de cada grupo com relação a **x**, definido pela igualdade $Pr(K = k|X = \mathbf{x}) = Pr(K = l|X = \mathbf{x})$, é um hiperplano de dimensão p. Calculando-se o limite entre cada par de classes, pode-se dividir $\mathbb{R}^p$ em subregiões a serem classificadas conforme cada uma das classes.

Desta forma, aloca-se uma nova observação $\mathbf{x}_0$ na classe cuja probabilidade *a posteriori*, dada pela Equação 2.35, é a maior dentre todas as outras:

$$\log Pr(K = k|X = \mathbf{x}_0) = \log \pi_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \mu_k'\Sigma^{-1}\mathbf{x}_0 \ . \tag{2.35}$$

Discriminante Quadrático (QDA)

O QDA se diferencia do LDA pela segunda hipótese. Ou seja, no QDA não se assume que as matrizes de covariâncias de cada grupo são iguais. Desta forma, a Equação 2.34, que define os hiperplanos que separam as regiões de diferentes classes, não será mais linear e sim quadrática em relação a **x**.

As probabilidades a posteriori assumirão a forma abaixo:

$$\log Pr(K = k|X = \mathbf{x}) = \log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) - \frac{p}{2}\log(2\pi) \ . \tag{2.36}$$

Como o termo $-\frac{p}{2}\log(2\pi)$ é o mesmo para qualquer grupo, a comparação entre as probabilidades *a posteriori* dos diferentes grupos é feita baseada nos valores dos demais termos. Assim, define-se o discriminante quadrático como:

$$discQ_k = \log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k) \ . \tag{2.37}$$

Aloca-se uma observação $\mathbf{x}_0$ no grupo em que o valor de $discQ_k$ for o maior em comparação com os demais grupos.

2.2.4 *Support Vector Machine* (SVM)

A técnica de SVM surgiu nos anos 90 em meio à comunidade de Ciências da Computação e tem crescido em popularidade desde então [James et al., 2013]. Essa técnica pode ser usada tanto para aplicações de classificação quanto para regressão.

Sua ideia fundamental é a de criar ”margens”, ou hiperplanos, que separem os dados de classes diferentes e que estejam afastadas desses dados o máximo possível.

SVM para Classificação

Tomando como exemplo os dados da Figura 2.12, pode-se notar que as observações possuem uma separação bem definida, ou seja, não existem observações de uma classe misturadas às observações de outra. Desta forma, existem inúmeros hiperplanos que separariam essas observações de forma eficaz. O objetivo da técnica de SVM é definir o hiperplano que separa os dados de forma a maximizar a sua distância em relação a eles.

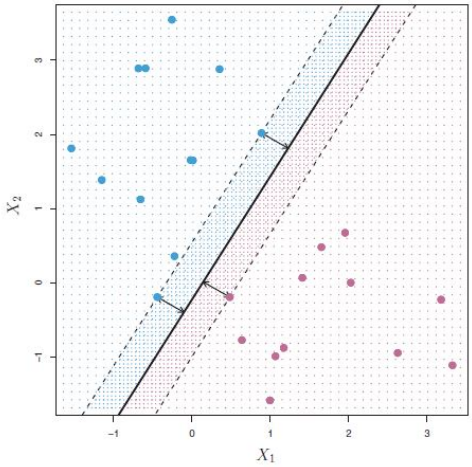

Figura 2.12: Dispersão de dados bidimensionais e separador SVM [James et al., 2013]

Neste processo de maximização, somente os dados que estão mais próximos dos limites entre as classes serão utilizados. Esses dados, ou observações, são chamados ”vetores de suporte”, e é por essa razão que o modelo recebe o nome de *Support Vector Machine*.

Para introduzir o modelo, considera-se primeiramente o caso em que as observações estão divididas em apenas duas classes diferentes. As variáveis que descrevem o problema são $\mathbf{X}^{(n)} = [x_1^{(n)}, x(n)_2, ..., x_P^{(n)}]$, com $\mathbf{X} \in \Re^P$ e $y^{(n)} \in \{-1, 1\}$. A base de dados teria então o aspecto da tabela abaixo.

Observações	x_1	x_2	...	x_P	y
1	.	.	.	.	1
2	.	.	.	.	-1
...	.	.	.	.	-1
N	.	.	.	.	1

A equação de um hiperplano para a separação dos dados é apresentada pela Equação 2.38, onde **w** é um vetor de tamanho $P \times 1$ e b é um escalar:

$$\mathbf{w}' \cdot \mathbf{X} + b = 0 \ . \tag{2.38}$$

Assim, qualquer dado que se localiza acima deste hiperplano possui $y = 1$, e qualquer dado que se localiza abaixo possui $y = -1$. Então se definirmos uma função $g(\mathbf{X}^{(n)}) = \mathbf{w}' \cdot \mathbf{X}^{(n)} + b$, teremos que $g(\mathbf{X}^{(n)}) > 0$ quando $y^{(n)} = 1$ e, da mesma forma, $g(\mathbf{X}^{(n)}) < 0$ quando $y^{(n)} = -1$. Ou seja, a Equação 2.39 a seguir é válida para todas as observações da base de dados:

$$y^{(n)}(\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) > 0 \quad \text{para} \quad n = 1, 2, ..., N \ . \tag{2.39}$$

Podem-se definir duas outras margens, uma acima e outra abaixo do hiperplano separador, Equações 2.40 e 2.41, respectivamente:

$$\mathbf{w}' \cdot \mathbf{X} + b - 1 = 0 \ , \tag{2.40}$$

$$\mathbf{w}' \cdot \mathbf{X} + b + 1 = 0 \ . \tag{2.41}$$

Assim, como as observações estarão sempre nas margens, acima da margem superior ou abaixo da margem inferior, a Equação 2.39 pode ser adaptada para a seguinte forma:

$$y^{(n)}(\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) \geq 1 \quad \text{para} \quad n = 1, 2, ..., N \ . \tag{2.42}$$

A distância do hiperplano separador em relação à margem superior é $\frac{1}{||\mathbf{w}||}$, e em relação à margem inferior é $-\frac{1}{||\mathbf{w}||}$.

O problema se resume então a uma otimização da forma:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \frac{1}{||\mathbf{w}||} \\ \text{sujeito a} \quad & y^{(n)}(\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) \geq 1 \quad n = 1, \ldots, N. \end{aligned}$$

Mas maximizar a quantidade $\frac{1}{||\mathbf{w}||}$ é o mesmo que minimizar $||\mathbf{w}||^2$. Fazendo essa alteração, termos um problema de otimização com função objetivo convexa e restrições lineares, o que garante que é possível obter o ótimo global.

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{sujeito a} \quad & y^{(n)}(\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) \geq 1 \quad n = 1, \ldots, N. \end{aligned}$$

Por se tratar de uma otimização limitada por restrições que são inequações, aplicam-se as condições de Karush-Kuhn-Tucker (KKT). Os multiplicadores KKT ($\mu^{(1)}$, $\mu^{(2)}$, ..., $\mu^{(N)}$) apontarão quais restrições estão ativas. Como cada restrição é proveniente de uma observação da base de dados, as restrições ativas correspondem às observações que estão sendo utilizadas na otimização e, portanto, elas indicam quais são os vetores de suporte.

Finalmente, pode-se demonstrar que este problema de maximização pode ser transformado em um problema de otimização dual da forma abaixo, que fornece o vetor **w** e o escalar b procurados [Ng, 2016].

$$\begin{aligned} \max_{\mu} \quad & W(\mu) = \sum_{n=1}^{N} \mu^{(n)} - \frac{1}{2} \sum_{n,i=1}^{N} y^{(n)} y^{(i)} \mu^{(n)} \mu^{(i)} (\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)}) \\ \text{sujeito a} \quad & \mu^{(n)} \geq 0 \quad n = 1, \ldots, N \\ & \sum_{n=1}^{N} \mu^{(n)} y^{(n)} = 0 \ . \end{aligned}$$

Porém esta otmização é aplicável apenas quando as observações de classes diferentes estão bem separadas. Para o caso em que há sobreposição de observações de diferentes classes, como mostra a Figura 2.13, é preciso introduzir uma variável de tolerância ($\xi^{(n)}$) para a distância de algumas observações em relação ao hiperplano separador.

A maximização da distância das margens adicionando-se as variáveis de tolerância fica, então, da forma abaixo. O parâmetro C regula o custo de se introduzir a

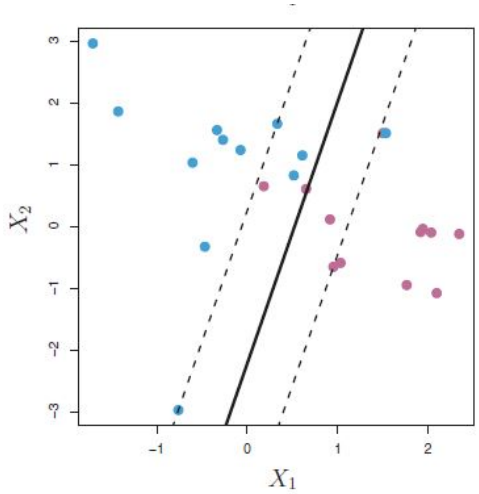

Figura 2.13: Exemplo de dados com sobreposição de observações com de classes diferentes [James et al., 2013]

tolerância para as observações e seu valor pode ser escolhido pelo usuário. Quanto maior o valor de C, menores serão as tolerâncias e, portanto, as margens superior e inferior se aproximarão do hiperplano separador e vice-versa.

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi^{(n)} \\ \text{sujeito a} \quad & y^{(n)}(\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) \geq 1 - \xi^{(n)} \quad n = 1, \ldots, N\ . \\ & \xi^{(n)} \geq 0 \quad n = 1, \ldots, N\ . \end{aligned}$$

Transformando-se para o problema de otimização dual, tem-se:

$$\begin{aligned} \max_{\mu} \quad & W(\mu) = \sum_{n=1}^{N}\mu^{(n)} - \frac{1}{2}\sum_{n,i=1}^{N} y^{(n)}y^{(i)}\mu^{(n)}\mu^{(i)}(\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)}) \\ \text{sujeito a} \quad & 0 \leq \mu^{(n)} \leq C \quad n = 1, \ldots, N \\ & \sum_{n=1}^{N}\mu^{(n)}y^{(n)} = 0\ . \end{aligned}$$

Até então, só foi abordado o caso em que os dados poderiam ser separados por fronteiras lineares (hiperplano). Para generalizar o modelo para fronteiras não lineares, substitui-se o produto escalar $(\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)})$ da função $W(\mu)$ por Kernels $K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)})$.

Os Kernels mais comuns são o Linear, o Polinomial, o Radial e o Sigmoidal, apresentados pelas Equações 2.43, 2.44, 2.45 e 2.46, respectivamente [Lorena and Carvalho, 2007, Meyer et al., 2015]:

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = (\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)}) \ , \tag{2.43}$$

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = (\delta(\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)}) + \kappa)^d \ , \tag{2.44}$$

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = \text{exp}\left(-\delta||\mathbf{X}^{(n)} - \mathbf{X}^{(i)}||^2\right) \ , \tag{2.45}$$

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = \text{tanh}\left(\delta(\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)}) + \kappa\right) \ , \tag{2.46}$$

onde δ, κ e d são parâmetros a serem definidos conforme a conveniência.

A utilização dos Kernels torna o problema de classificação aplicável a inúmeras estruturas de dados. A Figura 2.14 apresenta dois exemplos de classificação com Kernels. No quadro da esquerda, foi utilizado um kernel polinomial de ordem 3 e no quadro da direita um kernel radial [James et al., 2013].

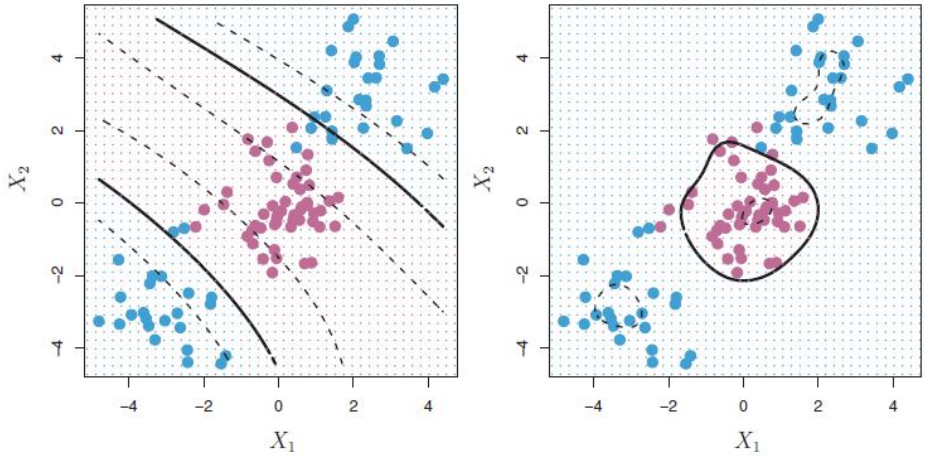

Figura 2.14: Esquerda: SVM com Kernel polinomial de ordem 3. Direita: SVM com Kernel radial [James et al., 2013].

SVM para Regressão

A diferença dos casos de regressão para os de classificação se encontra principalmente nas restrições da otimização. Na regressão deseja-se que as observações estejam distantes do hiperplano separador de um valor máximo ε [Smola and Scholkopf, 1998].

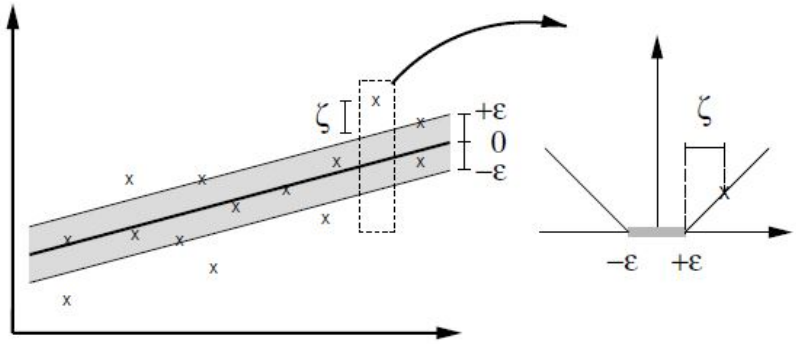

Figura 2.15: Representação gráfica de um caso de regressão com SVM [Smola and Scholkopf, 1998].

$$
\begin{aligned}
\min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 \\
\text{sujeito a} \quad & y^{(n)} - (\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) \leq \varepsilon \quad n = 1, \ldots, N \\
& \mathbf{w}' \cdot \mathbf{X}^{(n)} + b - y^{(n)} \leq \varepsilon \quad n = 1, \ldots, N \ .
\end{aligned}
$$

Aplicando-se variáveis de tolerância ($\xi^{(n)}$, para quando o valor real supera o valor da margem superior, e $\xi*^{(n)}$, para quando o valor real é menor do que o da margem inferior) ao problema, tem-se a formulação proposta por Vapnik (1995).

$$
\begin{aligned}
\min_{\mathbf{w}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}(\xi^{(n)} + \xi*^{(n)}) \\
\text{sujeito a} \quad & y^{(n)} - (\mathbf{w}' \cdot \mathbf{X}^{(n)} + b) \leq \varepsilon + \xi^{(n)} \quad n = 1, \ldots, N \\
& \mathbf{w}' \cdot \mathbf{X}^{(n)} + b - y^{(n)} \leq \varepsilon + \xi*^{(n)} \quad n = 1, \ldots, N \\
& \xi^{(n)}, \xi*^{(n)} \geq 0 \quad n = 1, \ldots, N \ .
\end{aligned}
$$

Assim, apenas as observações que distam do hiperplano separador de um valor maior que ε contribuem com o custo de minimização da função objetivo. A Figura 2.15 traz uma representação do problema.

Da mesma forma que nos casos de Classificação, este problema pode ser colocado no formato de otimização dual e podem se utilizar os Kernels para regressão de funções não lineares.

3 Materiais e Métodos

3.1 Bases de Dados

Os dados de concentração de poluentes foram coletados em tempo real pela CETESB [CETESB, 2016a], que possui uma rede de monitoração automática distribuída por todo o Estado de São Paulo. A RMSP conta com 26 estações fixas, sendo que 19 delas são capazes de medir a concentração de ozônio. As estações estão distribuídas conforme o mapa já mostrado na Figura 1.4, o qual é reproduzido a seguir pela Figura 3.1, com destaque para as estações que realizam medições de ozônio.

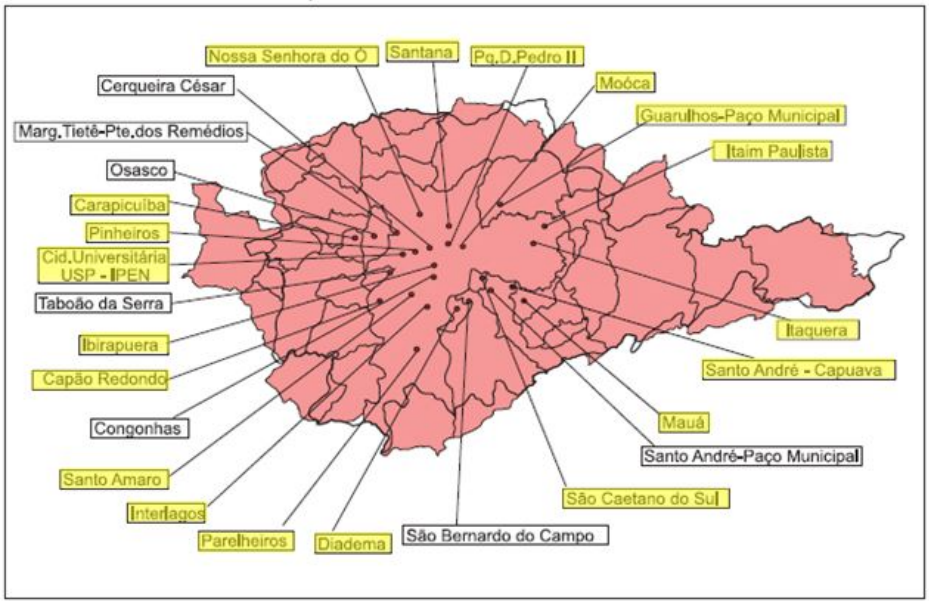

Figura 3.1: Distribuição das estações da Rede Automática da CETESB na RMSP [CETESB, 2014]

Os dados são disponibilizados a qualquer usuário via o sistema QUALAR, acessível em http://qualar.cetesb.sp.gov.br/qualar/home.do. Para o presente estudo, foram utilizados os dados medidos na Estação Ibirapuera, com séries históricas de 01 de Janeiro de 2000 a 31 de Agosto de 2016, que foram fornecidos diretamente, com base no convênio existente entre a USP e a CETESB.

Os dados meteorológicos, usados como variáveis entrada para o modelo, foram obtidos da Estação Meteorológica do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da USP (IAG-USP), localizada no Parque de Ciência e Tecnologia da USP, a aproximadamente 7,9 km da estação Ibirapuera da CETESB, conforme Figura 3.2.

A base de dados utilizada contém médias horárias das medidas das seguintes

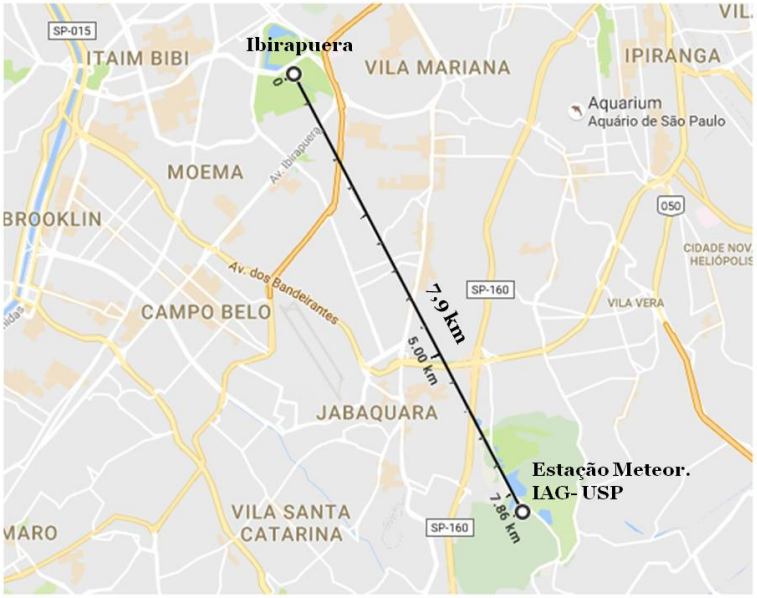

Figura 3.2: Distância entre o Parque Ibirapuera e o Parque de Ciência e Tecnologia da USP

variáveis medidas na superfície (pouco acima do nível do solo): Temperatura (em graus Celsius), Pressão (em hPa), Umidade Relativa (em porcentagem), Direção do Vento (de acordo com os pontos cardeais, colaterais e intermédios), Velocidade do Vento (em km/h) e Concentração Máxima diária de Ozônio (em $\mu g/m^3$). Para tornar o cálculo mais coerente em termos numéricos, as variáveis Velocidade do Vento (VV) e Direção do Vento (DV), esta última convertida para graus, foram combinadas de forma a se obter as componentes da velocidade do vento no sentido Norte-Sul (V) e Leste-Oeste (U):

$$U = VV \sin\left(DV \frac{\pi}{180}\right) , \tag{3.1}$$

$$V = VV \cos\left(DV \frac{\pi}{180}\right) . \tag{3.2}$$

A seleção destas variáveis meteorológicas como representativas do fenômeno de formação de O_3 foi baseada em resultados obtidos de estudos anteriores elaborados pelo Laboratório de Simulação e Controle de Processos (LSCP) do Departamento de Engenharia Química da Universidade de São Paulo em conjunto com a CETESB [Guardani et al., 1999, Borges et al., 2012]. Como neste estudo o objetivo é construir e analisar o desempenho de modelos para a previsão de concentração de O_3, é necessário que as variáveis a serem consideradas sejam passíveis de ser previstas. Por

este motivo, utilizaram-se variáveis meteorológicas e não medidas da concentração de poluentes primários como NO_x e COVs.

No entanto, como esses poluentes precursores do O_3 são emitidos, em geral, por fontes móveis, cujo padrão de circulação na RMSP difere em dias de semana e em finais de semana, adicionou-se uma variável binária para diferenciar esses dias, estabelecendo assim, indiretamente, um padrão de emissão de precursores.

Adicionou-se também como variável de entrada a concentração máxima de O_3 do dia anterior a cada observação, devido à pressuposição de que concentrações do dia anterior podem afetar a concentração do dia seguinte. Foram excluídas da base de dados quaisquer observações com medidas faltantes.

A partir desses dados, foram construídas três diferentes bases para análise das máximas diárias considerando as médias horárias de O_3 (Figura 3.3) :

MtMM	MMM	MMT	
T máxima	T máxima	T méd. manhã	T méd. tarde
P mínima	P mínima	P méd. manhã	P méd. tarde
UR mínima	UR mínima	UR méd. manhã	UR méd. tarde
U médio tarde	U médio dia inteiro	U méd. manhã	U méd. tarde
V médio tarde	V médio dia inteiro	V méd. manhã	V méd. tarde
Dias da Semana (binária)	Dias da Semana (binária)	Dias da Semana (binária)	
Conc. O_3 máx. do dia anterior	Conc. O_3 máx. do dia anterior	Conc. O_3 máx. do dia anterior	

Figura 3.3: Bases de dados utilizadas e suas variáveis de entrada

1. MtMM: Temperatura diária máxima, pressão e umidade relativa mínimas, média da tarde (das 13:00 às 17:00) das componentes da velocidade do vento nas direções Norte-Sul e Leste-Oeste, variável binária indicando dias da semana ou final de semana e concentração máxima de O_3 do dia anterior. Total de 7 variáveis mais a concentração máxima diária de O_3 e 5603 observações.

2. MMM: Temperatura diária máxima, pressão e umidade relativa mínimas, média diária das componentes da velocidade do vento nas direções Norte-Sul e Leste-Oeste, variável binária indicando dias da semana ou final de semana e concen-

tração máxima de O_3 do dia anterior. Total de 7 variáveis mais a concentração máxima diária de O_3 e 5594 observações.

3. MMT: Médias da manhã (das 8:00 às 12:00) e tarde (das 13:00 às 17:00) da temperatura, pressão, umidade relativa e componentes da velocidade do vento nas direções Norte-Sul e Leste-Oeste, variável binária indicando dias da semana ou final de semana e concentração máxima de O_3 do dia anterior. Total de 13 variáveis mais a concentração máxima diária de O_3 e 5598 observações.

As três bases citadas foram testadas para os mesmos métodos de estatística multivariada duas vezes: uma com a variável binária que indica dias da semana e outra sem, uma vez que essa variável parece aprimorar os resultados em alguns testes e prejudicar em outros.

A base de dados que gerou os melhores resultados para a previsão da máxima diária das médias horárias de O_3 foi, em seguida, aplicada para a previsão da máxima diária considerando as médias móveis de 8 horas.

Esta última foi, então, aplicada para os casos de classificação, aqueles em que a resposta obtida é o rótulo de uma classe, e as concentrações máximas de ozônio foram classificadas em 5 classes diferentes segundo os índices de qualidade do ar definidos pela CETESB, considerando as médias móveis de 8 horas para a concentração de O_3:

BOA: Concentração máxima de O_3 de 0 a 100 $\mu g/m^3$
MODERADA: Concentração máxima de O_3 de 100 a 130 $\mu g/m^3$
RUIM: Concentração máxima de O_3 de 130 a 160 $\mu g/m^3$
MUITO RUIM: Concentração máxima de O_3 de 160 a 200 $\mu g/m^3$
PÉSSIMA: Concentração máxima de O_3 maior que 200 $\mu g/m^3$

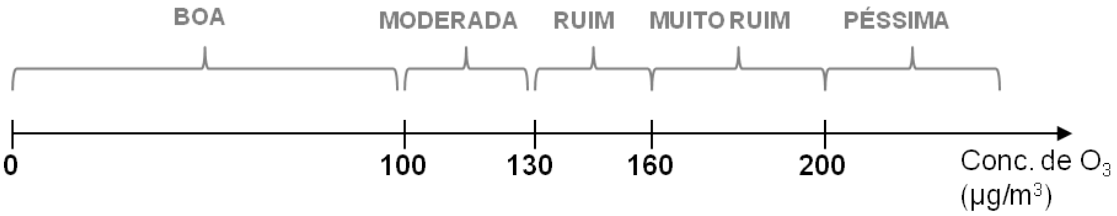

Figura 3.4: Classes segundo a concentração máxima de O_3, considerando as médias móveis de 8 horas, conforme definido pela CETESB

3.2 Metodologia de simulação

Regressão

Cada base de dados foi testada duas vezes para todas as configurações dos modelos de regressão (MLP, RF e SVM): na primeira vez incluindo a variável binária de dias da semana e na segunda excluindo-a. Foi então selecionada a configuração ótima para cada modelo, conforme descrito nas seções 3.2.1, 3.2.3, 3.2.4, 3.2.5 e 3.2.6, e a ela foram aplicados indicadores de desempenho (IDs), descritos na seção 3.2.2. A base de dados que gerou os melhores valores de IDs foi mantida para análises seguintes e as demais foram descartadas.

Os IDs obtidos para esta base foram, então, comparados através do teste t para amostras dependentes, ou seja, o teste t aplicado para populações cujos elementos estão relacionados, uma vez que as previsões são feitas exatamente para os mesmos dias. O test t, descrito na seção 3.2.7, permitiu analisar se os resultados obtidos para um modelo são estatisticamente diferentes dos obtidos para outro. Finalmente, pode-se concluir qual o melhor modelo de regressão para prever concentrações máximas de O_3 segundo as médias horárias. A Figura 3.5 apresenta um fluxograma deste processo.

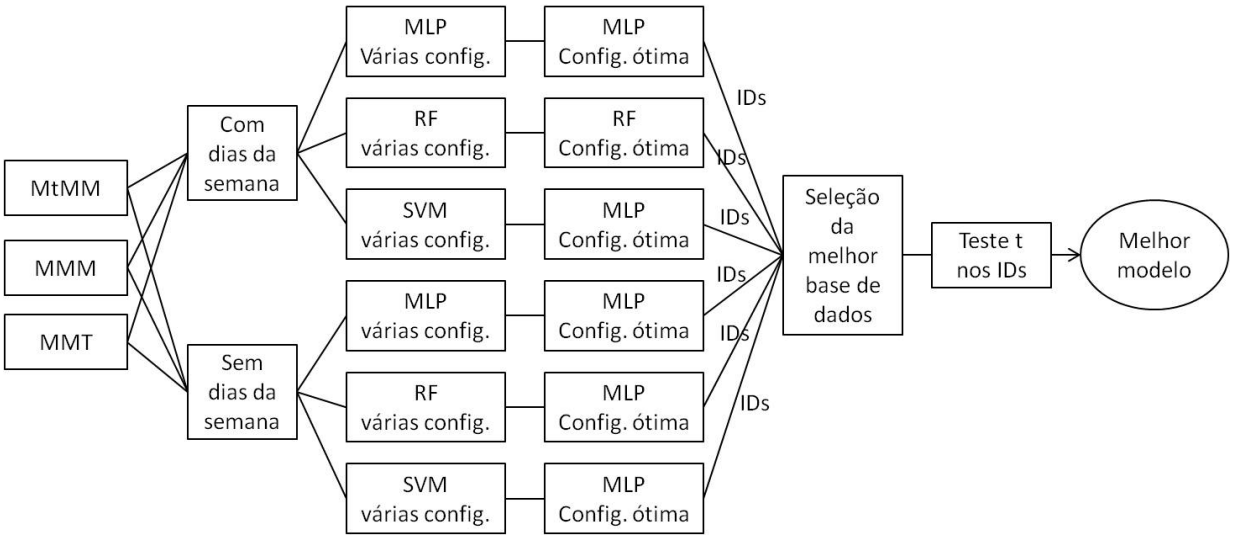

Figura 3.5: Fluxograma do processo de decisão do melhor modelo para os casos de regressão

O melhor modelo, com as melhores configurações, foi então aplicado para a previsão da máxima diária de O_3 considerando as médias móveis de 8 horas, utilizando

como variáveis de entrada as provenientes da base de dados com melhores resultados na análise anterior.

Classificação

A base de dados usada na previsão da máxima diária de O_3 considerando as médias móveis de 8 horas foi adaptada conforme descrito na seção 3.1 para ser aplicada em modelos de classificação. Ela foi então testada duas vezes para todos os modelos da classificação (MLP, RF, SVM, LDA e Fisher), de forma semelhante ao descrito nos casos de regressão: na primeira vez incluindo a variável binária de dias da semana e na segunda excluindo-a. A configuração ótima para cada modelo foi selecionada, conforme descrito nas seções 3.2.1, 3.2.3, 3.2.4, 3.2.5 e 3.2.6. Os erros percentuais de classificação (descritos nas seções 3.2.1 e 3.2.2) dos modelos ótimos foram comparados através do teste t para amostras dependentes e, finalmente, concluiu-se qual o melhor modelo para esta aplicação. A Figura 3.6 apresenta o fluxograma do processo de análise.

3.2.1 Validação Cruzada (VC)

Quando se comparam diferentes modelos e diferentes configurações de modelos, o método mais indicado para se comparar o desempenho de cada modelo é o de validação cruzada (VC) do tipo *k-fold*, que é um método que permite a utilização de toda a base de dados para avaliação do desempenho de um modelo. Neste método se particiona a base de dados inicial em k partes com o mesmo, ou aproximadamente o mesmo, número de observações, treina-se um modelo em $k-1$ partes, sendo elas consideradas os dados de treino (*training-set*), e testa-se o desempenho do modelo na parte não usada no treinamento, considerada o *test-set*. Repete-se este processo k vezes, até se obter k diferentes valores de desempenho do modelo. A média desses valores é então considerada uma estimação do erro de previsão do modelo com determinados parâmetros [Hastie et al., 2008].

Neste estudo aplicou-se o método de VC do tipo *10-fold*, assim cada base de dados foi incialmente ordenada de forma aleatoria e em seguida divida em 10 grupos com o mesmo, ou aproximadamente o mesmo, número de observações, conforme

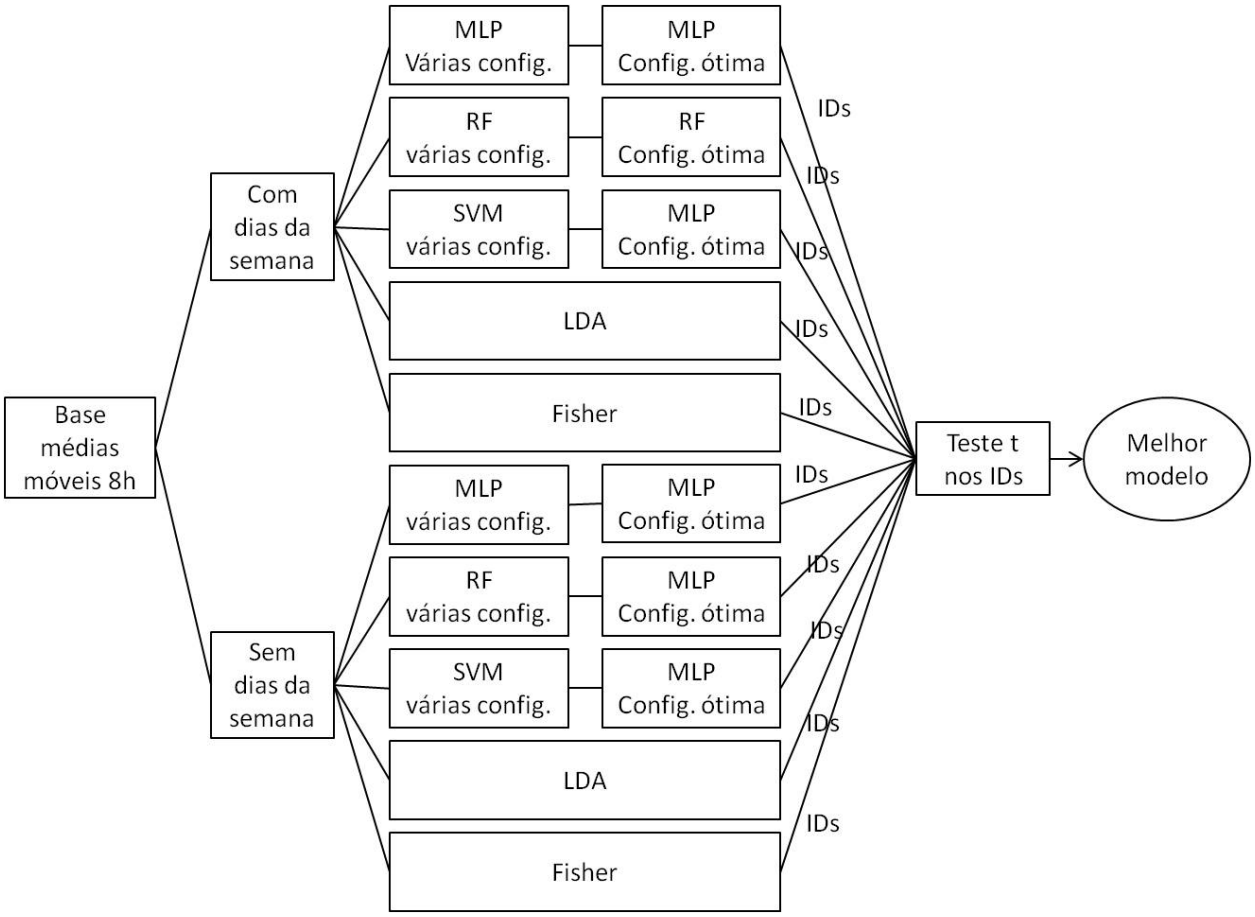

Figura 3.6: Fluxograma do processo de decisão do melhor modelo para os casos de classificação

Figura 3.7. Então, para cada configuração de um modelo treinou-se o modelo em 9 das 10 partes e testou-se seu desempenho, através da aplicação dos indicadores de desempenho (IDs) descritos na seção 3.2.2 desta dissertação, na parte restante não utilizada para o treinamento, que está representada por cinza na Figura 3.7. Esse processo foi, então, repetido 10 vezes de forma que cada parte da base de dados tenha sido utilizada como *test-set* uma vez. A estimativa do erro total de predição do modelo é considerada a média dos 10 diferentes valores obtidos em cada uma das iterações na VC, conforme Equação 3.3:

$$\overline{ID} = \sum_{k=1}^{10} \frac{1}{10} ID_k \ . \tag{3.3}$$

O desvio padrão dessa estimativa é dado pela Equação 3.4:

$$dp_{ID} = \sqrt{\sum_{k=1}^{10} \frac{(ID_k - \overline{ID})^2}{(10-1)}} \ . \tag{3.4}$$

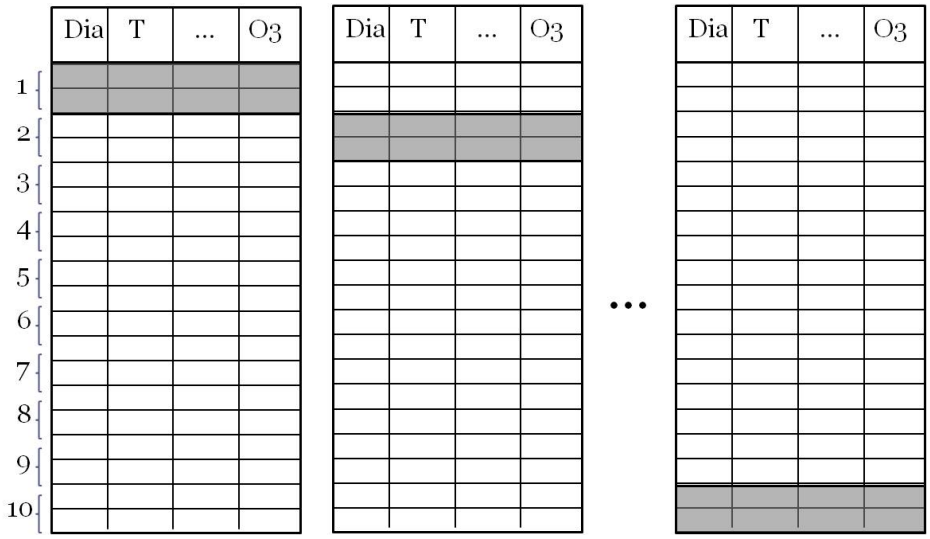

Figura 3.7: Representação da divisão da base de dados e aplicação do método de VC do tipo *10-fold*; as seções em cinza representam o *test-set* em cada uma das iterações do método

Para exemplicar o processo, toma-se como exemplo um caso de regressão através de um MLP com um número h de neurônios na camada oculta. Para casos de regressão, um dos IDs descrito na seção 3.2.2 é o MSE (Equação 2.7). Treina-se então o MLP 10 vezes neste processo, obtendo-se 10 valores de MSE diferentes. O erro de predição do modelo MLP para h neurônios na camada oculta é, então, dado pela Equação 3.5, e o desvio padrão pela Equação 3.6:

$$\overline{MSE} = \sum_{k=1}^{10} \frac{1}{10} MSE_k \ , \tag{3.5}$$

$$dp_{MSE} = \sqrt{\sum_{k=1}^{10} \frac{(MSE_k - \overline{MSE})^2}{(10-1)}} \ . \tag{3.6}$$

3.2.2 Indicadores de Desempenho (IDs)

Como IDs de todos os modelos de regressão foram usadas 4 medidas: Média dos Erros Quadráticos (MSE, Equação 3.7), Média dos Erros Absolutos (MAE, Equação 3.8), Média dos Desvios (MBE, Equação 3.9) e Coeficiente de Correlação (R, Equação 3.10).

$$MSE = \frac{1}{N} \sum_{n=1}^{N} (y_{calc}^{(n)} - y_{exp}^{(n)})^2 \ . \tag{3.7}$$

$$MAE = \frac{1}{N}\sum_{n=1}^{N} |y_{calc}^{(n)} - y_{exp}^{(n)}| \ . \tag{3.8}$$

$$MBE = \frac{1}{N}\sum_{n=1}^{N} (y_{calc}^{(n)} - y_{exp}^{(n)}) \ . \tag{3.9}$$

$$R = \sqrt{\frac{\sum_{n=1}^{N} (y_{exp}^{(n)} - \overline{y}_{exp}^{(n)})^2 - \sum_{n=1}^{N} (y_{exp}^{(n)} - y_{calc}^{(n)})^2}{\sum_{n=1}^{N} (y_{exp}^{(n)} - \overline{y}_{exp}^{(n)})^2}} \ . \tag{3.10}$$

As duas primeiras indicam a distância dos resultados calculados por meio dos modelos e os valores reais, ou experimentais. A MBE indica a tendência do modelo de superestimar ou subestimar os resultados em relação aos valores reais. Para esses três indicadores o melhor valor é o mais próximo de zero.

O R indica a força da linearidade do modelo. Ou seja, num gráfico onde a abscissa corresponde aos dados experimentais e a ordenada aos dados calculados, o coeficiente R indica a relação de linearidade entre essas duas medidas. Quanto mais próximo de 1, mais forte é a relação.

Para os casos de classificação, o erro é calculado pela porcentagem observações classificadas erroneamente, conforme Equação 3.11:

$$Erro[\%] = \frac{\sum_{n=1}^{N} I(y_{exp}^{(n)} \neq y_{calc}^{(n)})}{N} \ . \tag{3.11}$$

3.2.3 *Multi-layer Perceptron Neural Networks* (MLP)

Preprocessamento de dados

As redes neurais MLP para regressão e classificação foram ambas construídas no software MATLAB com as funções *feedforwardnet()* e *patternnet()*, respectivamente. Essas funções realizam um preprocessamento dos dados automaticamente em que todos os valores da base de entrada são convertidos em valores no intervalo [-1,1]. O valor máximo cada uma das variáveis considerando todas as observações da base é convertido em 1, e o mínimo em -1.

A mudança da escala dos dados aumenta o desempenho do algoritmo de otimização, uma vez que coloca todas as variáveis na mesma ordem de grandeza.

Após a otimização do MLP, os dados são automaticamente convertidos para a escala original.

Regressão

Para o caso de regressão, foram testadas 26 arquiteturas diferentes para cada base de dados, conforme apresentado pela Figura 3.8. Utilizou-se apenas uma camada oculta, cujo número de neurônios variou de 3 a 15. A razão para a escolha dessa configuração é que um MLP com uma única camada oculta com função de ativação sigmoidal, sem restrições no número de neurônios ou tamanho dos pesos, pode aproximar qualquer função continua com uma precisão arbitrária, conforme demonstrou Cybenko (1989).

Foram testadas também duas configurações diferentes quanto às funções de ativação. Uma com a função sigmoidal (Equação 2.3) na camada oculta e a função linear (Equação 2.4) na camada de saída e outra com a função sigmoidal nas duas camadas.

Com a aplicação de VC do tipo *10-fold*, cada uma das arquiteturas para uma mesma base de dados foi simulada 10 vezes, uma para cada porção diferente de cada base como *test-set*.

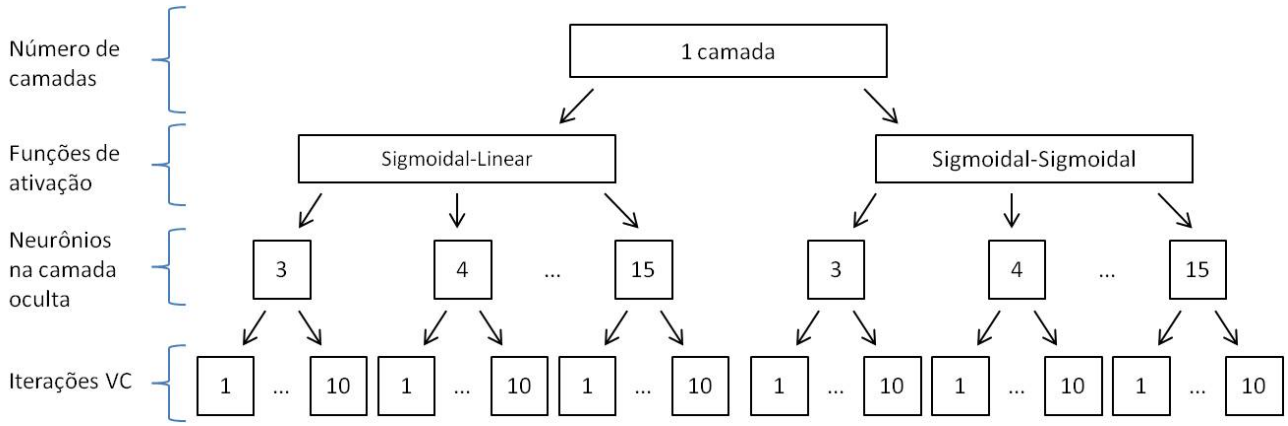

Figura 3.8: Arquiteturas de MLPs testadas para os casos de regressão e iterações de VC

O treinamento de MLPs apresenta um fator de aleatoriedade no momento em que os pesos w são inicializados na primeira etapa do processo de otimização. Portanto, cada treinamento resulta em valores diferentes dos pesos e, consequentemente,

em um ajuste diferente da MLP aos dados. Para minimizar o efeito da inicialização aleatória, cada configuração das MLPs foi treinada 10 vezes. Destas 10, o treinamento que produziu o melhor resultado foi o escolhido.

A função objetivo utilizada em todos os casos de regressão foi a Média dos Erros Quadráticos (MSE, Equação 2.7) e o algoritmo de otimização foi o Levenberg-Marquardt, que é uma boa alternativa ao método de Gradiente Descendente devido à sua rapidez de resolução. O treinamento de um MLP pelo MATLAB é interrompido quando um dos dois critérios abaixo é satisfeito:

1. O módulo do gradiente da função objetivo no ponto estudado é menor do que um valor pré-determinado (1e-7);

2. O valor da função objetivo calculado para um conjunto de observações de validação não diminui ao longo de 25 iterações, onde cada iteração corresponde à apresentação de todas as observações para o algoritmo de otimização uma vez;

3. Um número máximo de 100 iterações é atingido;

4. O valor do parâmetro de Levenberg-Marquardt alcança um valor máximo (1e+10) (para melhor entedimento do algoritmo Levenberg-Marquardt, consultar a referência [Marquardt, 1963]).

Classificação

Nos casos de classificação, foram testadas 13 arquiteturas diferentes, em que a função objetivo foi a Entropia Cruzada (Equação 2.10), conforme apresentado pela Figura 3.9. Para estes últimos casos, as funções de ativação foram sigmoidal, para a camada oculta, e *softmax* (Equação 2.11), para a camada de saída.

Assim como nos casos de regressão, os modelos foram construídos software MATLAB. O treinamento de cada MLP foi interrompido quando um dos três critérios abaixo foi satisfeito:

1. O módulo do gradiente da função objetivo no ponto estudado é menor do que um valor pré-determinado (1e-6);

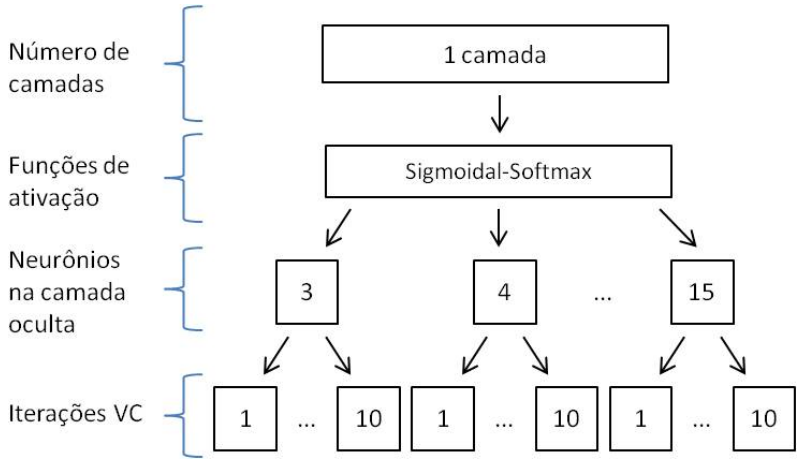

Figura 3.9: Arquiteturas de MLPs testadas para os casos de classificação e iterações de VC

2. O valor da função objetivo calculado para um conjunto de observações de validação não diminui ao longo de 25 iterações, onde cada iteração corresponde à apresentação de todas as observações para o algoritmo de otimização uma vez;

3. Um número máximo de 1000 iterações é atingido;

Seleção da Melhor Arquitetura

Comparou-se então o erro de predição das diferentes configurações, dado pela média do MSE (Equação 3.7) entre a iterações da VC (Equação 3.3), nos casos de regressão, e pela média do erro percentual (Equação 3.11) entre as iterações da VC nos casos de classificação. A arquitetura com o menor erro foi eleita como a arquitetura ótima dentre todas as outras.

3.2.4 *Random Forests* (RF)

Diferentemente do modelo MLP, o RF não exige normalização dos dados, uma vez que a diferença de escala das variáveis de entrada não prejudica o algoritmo de otimização.

Os modelos de RF foram construídos segundo a metodologia proposta por Breiman (2001), ou seja, fazendo-se sequencialmente os processos de *bagging* e construção de árvores até sua máxima extensão.

O número B de *subsets* gerados foi 150, pois para números maiores o custo computacional foi muito alto e a melhora no desempenho não foi significativa.

O número de váriáveis m a serem consideradas na definição de cada nó variou de 2 ao número total p de variáveis de entrada de cada base. Devido aos fatores de aleatoriedade do modelo, cada configuração foi testada 10 vezes e o melhor teste foi escolhido.

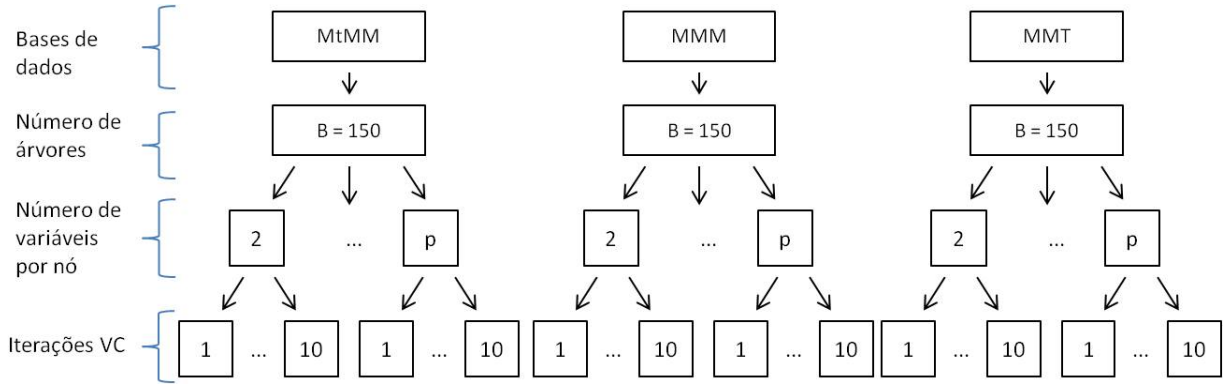

Figura 3.10: Configurações de RFs testadas

Todas as observações das bases de dados são usadas na construção dos *subsets* e, portanto no treinamento dos modelos. Em cada *subset* gerado por amostragem aleatória, aproximadamente um terço das observações não são utilizadas [Breiman, 2001]. Essas observações são chamadas de *out-of-bag*. A previsão de concentração de ozônio para uma dada observação é feita utilizando-se somente as árvores que foram construídas a partir de *subsets* que não continham essa observação. Esse método é chamado de estimação *out-of-bag*.

Com a aplicação de VC do tipo *10-fold*, cada um dos 10 *test-sets* foi usado para o cálculo do desempenho do modelo através da previsão *out-of-bag*.

A função objetivo é a MSE (Equação 2.7), para os casos de regressão, e a Entropia Cruzada (Equação 2.10), para os casos de classificação.

Seleção da Melhor Configuração

De forma similar ao MLP, comparou-se o erro de predição das diferentes configurações, dado pela média do MSE (Equação 3.7) entre a iterações da VC (Equação 3.3), nos casos de regressão, e pela média do erro percentual (Equação 3.11) entre as iterações da VC nos casos de classificação. A configuração com o menor erro foi eleita como a configuração ótima dentre as demais.

3.2.5 Análise de Discriminantes

Os métodos de LDA e Fisher, aplicados apenas nos casos de classificação, consistem na resolução de equações sem nenhum parâmetro ajustável. Não existe também nenhum fator de aleatoriedade no processo de resolução. Portanto, LDA e Fisher foram aplicados apenas 10 vezes para cada base de dados devido às iterações de VC do tipo *10-fold*. Todos as aplicações de LDA foram feitas através do software MATLAB e as de Fisher através do software R.

Neste estudo não foi usado o modelo de discriminante quadrático porque ele falhou em testes preliminares devido ao fato de o grupo de concentração mais alta de O_3 (grupo 'PÉSSIMA') conter apenas 7 observações, portanto a matriz de covariância obtida para este grupo não é considerada representativa de uma população inteira.

3.2.6 *Support Vector Machine* (SVM)

Os problemas de SVM consistem na resolução de um problema de otimização com função objetivo convexa e restrições lineares, dessa forma é possível atingir o ótimo global. Não existe nenhum fator de aleatoriedade no processo de resolução.

Como parâmetros ajustáveis tem-se a escolha dos kernels. As opções disponíveis no pacote *e1071* do software R, a ferramenta utilizada na aplicação desse método, são os kernels linear, polinomial, radial e sigmoidal (Equações 2.43, 2.44, 2.45 e 2.46, respectivamente). Os kernels polinomial, radial e sigmoidal possuem o seguintes parâmetros a serem definidos: δ, κ e d. Os valores utililizados foram os valores definidos como padrão no pacote *e1071*, de forma que os kernels aplicados são apresentados pelas equações abaixo:

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = (\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)}) \ , \tag{3.12}$$

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = \left(\frac{1}{P}(\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)})\right)^3 \ , \tag{3.13}$$

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = \exp\left(-\frac{1}{P}||\mathbf{X}^{(n)} - \mathbf{X}^{(i)}||^2\right) \ , \tag{3.14}$$

$$K(\mathbf{X}^{(n)}, \mathbf{X}^{(i)}) = \tanh\left(\frac{1}{P}(\mathbf{X}^{(n)} \cdot \mathbf{X}^{(i)})\right), \tag{3.15}$$

onde P é a dimensão dos dados, ou o número de colunas de **X**.

Para valor do custo C, outro parâmetro a ser definido, foram testados valores de 1 a 101, variando 10 a 10. O valor que gerou os melhores resultados foi o escolhido. Nos casos em que este valor foi 101, o limite do intervalo, testaram-se valores progressivamente maiores até se encontrar um valor ótimo. As possíveis configurações de SVM aplicadas para cada uma das bases de dados são representadas pela Figura 3.11.

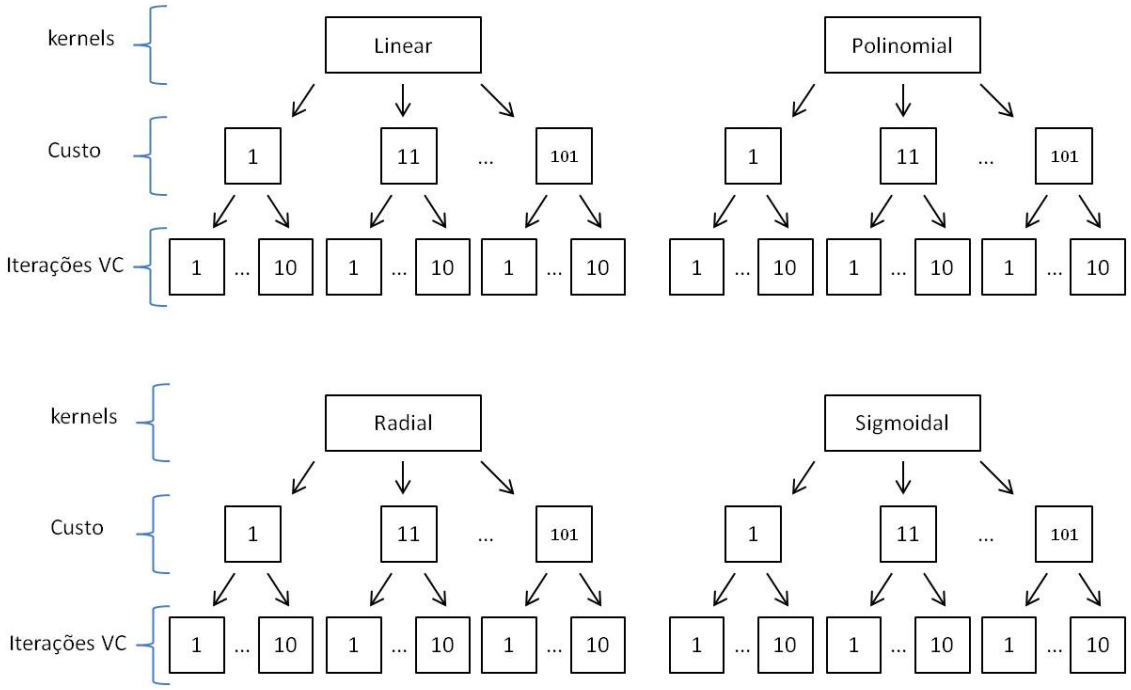

Figura 3.11: Configurações de SVMs testadas

Aqui, mais uma vez, cada uma das configurações foi simulada 10 vezes, uma para cada porção diferente de cada base devido à VC do tipo *10-fold*.

Seleção da Melhor Configuração

Assim como para MLP e RF, comparou-se o erro de predição das diferentes configurações, dado pela média do MSE (Equação 3.7) entre a iterações da VC (Equação 3.3), nos casos de regressão, e pela média do erro percentual (Equação 3.11) entre as iterações da VC nos casos de classificação. A configuração com o menor erro foi eleita como a configuração ótima dentre as demais.

3.2.7 Teste t para Amostras Dependentes

A aplicação de VC do tipo *10-fold* faz com que, para cada configuração ótima obtida para MLP, RF, SVM, LDA e Fisher, tenham-se 10 valores de cada ID. Assim, por exemplo, para um MLP de h neurônios na camada oculta, num caso de regressão, tem-se como resultado 10 valores de MSE, 10 de MBE, 10 de MAE e 10 valores de R. No caso de classificação, tem-se 10 valores de erro percentual de classificação. As médias de cada um destes indicadores representam o desempenho dessa configuração do modelo de MLP.

Muitas vezes, as médias destes IDs (Equação 3.3) para modelos diferentes possuem valores numericamente próximos, de forma que é impreciso considerar um modelo melhor do que outro apenas pela inspeção visual. Portanto, para se testar se as médias dos resultados de dois modelos são estatísticamente diferentes entre si, ou seja, pertencem a populações de dados diferentes, aplica-se o teste t para amostras dependentes para cada um dos IDs. As amostras, ou seja, os conjuntos dos 10 valores de cada ID são consideradas dependentes porque estes valores foram obtidos a partir de uma mesma base de dados.

Assim, por exemplo, num caso de regressão, onde se deseja decidir se a $\overline{MSE}^{MLP}$ obtida na configuração ótima de MLP é estatísticamente diferente da $\overline{MSE}^{RF}$ obtida na configuração ótima de RF, aplica-se o seguinte procedimento:

1. Define-se a variável $\bar{D}$, conforme Equações 3.16 e 3.17:

$$\bar{D} = \frac{1}{10}\sum_{k=1}^{10} D_k = \frac{1}{10}\sum_{k=1}^{10}\left(MSE_k^{MLP} - MSE_k^{RF}\right) , \tag{3.16}$$

$$\bar{D} = \overline{MSE}^{MLP} - \overline{MSE}^{RF} ; \tag{3.17}$$

2. Calcula-se o seu desvio padrão conforme Equação 3.18:

$$dp_D = \sqrt{\frac{1}{10-1}\sum_{k=1}^{10}(D_k - \bar{D})^2} ; \tag{3.18}$$

3. Calcula-se a estatística T, conforme Equação 3.19, cuja distribuição é do tipo t de Student com (10 -1) graus de liberdade, uma vez que as populações comparadas

possuem 10 representantes

$$T = \frac{\sqrt{10}(\bar{D} - \mu_D)}{dp_D} \ , \tag{3.19}$$

onde μ_D é o valor médio real da população de indivíduos D_k;

4. Aplica-se um teste de hipóteses do tipo:

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

para um nível de significância α, geralmente igual a 5 ou 10% ;

5. Finalmente, testa-se a estatística T, obtendo-se o valor p (*p-value*) relativo a ela. Se o valor p for maior do que α, isso significa que existe uma probabilidade significativa de se obter valores tão extremos quanto T numa nova amostragem, então não se pode rejeitar a hipótese H_0, e assim os valores $\overline{MSE}^{MLP}$ e $\overline{MSE}^{RF}$ não podem ser considerados estatisticamente diferentes. Ou seja:

Se $p < \alpha$ rejeita-se H_0 e os resultados dos modelos são diferentes;
se $p \geqslant \alpha$ não se rejeita H_0.

Este procedimento foi aplicado para todos os IDs (MSE, MBE, MAE e erro percentual) para comparar o resultado de todos os pares de modelos com configuração ótima obtidos a partir de uma mesma base de dados. Para um entendimento mais aprofundado do teste t para amostras dependentes, consultar a referência [Bussab and Moretin, 2013].

3.2.8 Análise de Sensibilidade

Uma vez obtido o modelo de regressão que melhor correlaciona as variáveis de entrada à concentração de ozônio, pode-se verificar a importância de cada uma

das variáveis de entrada através da análise de sensibilidade. Para cada modelo (MLP, SVM, RF, LDA e Fisher) deve-se aplicar um método diferente para a análise de sensibilidade. Já antecipando que os melhores resultados foram obtidos com o modelo MLP, neste estudo foi utilizado o método HIPR (*Holdback Input Randomization*) para análise de sensibilidade, uma vez que esse método foi desenvolvido para aplicação em redes neurais [Kemp et al., 2007].

O método HIPR consiste em comparar o desempenho das redes neurais, ou seja, o MSE, quando os valores de uma das variáveis de entrada são subtituídos por valores aleatórios. Ele funciona de acordo com as seguintes etapas:

1. Treina-se um modelo MLP com os dados originais
2. Utiliza-se o *test-set* para determinar a importância relativa das variáveis de entrada:
   (a) Subtituem-se os valores de uma das variáveis de entrada por valores aleatórios dentro do mesmo intervalo dos valores reais;
   (b) Calcula-se o MSE do *test-set* alterado;
   (c) Repete-se o procedimento para todas as variáveis de entrada.

Todo este processo foi repetido 30 vezes, visando minimizar qualquer possível viés obtido no processo de randomização dos valores. Obteve-se assim um valor de MSE médio para cada variável randomizada. A variável que, quando teve seus valores alterados, resultou num MSE médio mais elevado é a variável mais importante do modelo, a segunda variável com maior MSE médio é a segunda mais importante e assim por diante. Estabelece-se assim uma ordem de importância das variáveis de entrada.

4 Resultados e Discussão

4.1 Regressão

4.1.1 Médias horárias

Para os casos de regressão, foram testados os modelos MLP, RF e SVM, como mencionado anteriormente, uma vez que os modelos de discriminantes não são aplicáveis para regressão. Os resultados dos indicadores obtidos para todos os testes de regressão e as configurações ótimas dos modelos para cada caso são apresentados no Anexo A.

Os melhores IDs foram obtidos para a base MMT, que considerou as médias da manhã e tarde de cada variável, seguida pela MMM e MtMM.

Os melhores resultados para os casos de regressão e com a base de dados MMT juntamente com os parâmetros ótimos encontrados para cada modelo são apresentados na Tabela 4.1. Os resultados estão ordenados do menor valor de $\overline{MSE}$ para o maior.

Tabela 4.1: Melhores resultados e parâmetros para os casos de regressão

	Nº do Modelo	Base de dados	Dias da Semana	Modelo	Parâmetros Ótimos	$\overline{MSE}$	$\overline{MAE}$	$\overline{MBE}$	$\bar{R}$
1º	R19	MMT	sim	MLP	11 neurônios; sig. - sig.	560	17,2	-0,35	0,8674
2º	R20	MMT	não	MLP	14 neurônios; sig. - sig.	564	17,4	-0,18	0,8662
3º	R17	MMT	sim	MLP	14 neurônios; sigmoidal - linear	566	17,4	0,03	0,8662
4º	R18	MMT	não	MLP	15 neurônios; sigmoidal - linear	566	17,5	-0,01	0,8658
5º	R23	MMT	sim	SVM	Kernel: radial; Custo = 1	590	17,1	2,28	0,8619
6º	R24	MMT	não	SVM	Kernel: radial; Custo = 1	593	17,3	2,06	0,8607
7º	R21	MMT	sim	RF	150 árvores; m = 8	655	18,6	0,56	0,8434
8º	R22	MMT	não	RF	150 árvores; m = 6	657	18,6	0,60	0,8430

Observa-se que os modelos com a variável de dias da semana geraram resultados ligeiramente melhores do que seus pares sem dias da semana. Também é possível notar que os modelos de MLP apresentaram melhores valores dos IDs com relação aos demais modelos, com exceção do indicador $\overline{MAE}$, cujos valores foram semelhantes para os modelos de MLP e SVM.

Para se constatar se os resultados obtidos para cada modelo são estatísticamente diferentes, ou seja, pertencem a populações diferentes, aplicou-se então o teste t para amostras dependentes com confiança de 90% ($\alpha = 10\%$), cujos resultados são apresentados pelas Tabelas 4.2, 4.3, 4.4 e 4.5.

Tabela 4.2: Teste t para o índice $\overline{MSE}$ com 90% de confiança

$\overline{\text{MSE}}$		R17 MLP	R18 MLP	R19 MLP	R20 MLP	R21 RF	R22 RF	R23 SVM	R24 SVM
R17	MLP	-	0,93	0,47	0,86	0,06	0,06	0,58	0,52
R18	MLP	0,93	-	0,46	0,82	0,06	0,05	0,58	0,51
R19	MLP	0,47	0,46	-	0,67	0,05	0,04	0,49	0,44
R20	MLP	0,86	0,82	0,67	-	0,04	0,03	0,52	0,45
R21	RF	0,06	0,06	0,05	0,04	-	0,72	0,00	0,00
R22	RF	0,06	0,05	0,04	0,03	0,72	-	0,00	0,00
R23	SVM	0,58	0,58	0,49	0,52	0,00	0,00	-	0,32
R24	SVM	0,52	0,51	0,44	0,45	0,00	0,00	0,32	-

Tabela 4.3: Teste t para o índice $\overline{MAE}$ com 90% de confiança

$\overline{\text{MAE}}$		R17 MLP	R18 MLP	R19 MLP	R20 MLP	R21 RF	R22 RF	R23 SVM	R24 SVM
R17	MLP	-	0,20	0,37	0,95	0,02	0,02	0,57	0,91
R18	MLP	0,20	-	0,04	0,30	0,03	0,02	0,36	0,65
R19	MLP	0,37	0,04	-	0,25	0,01	0,01	0,79	0,84
R20	MLP	0,95	0,30	0,25	-	0,01	0,01	0,50	0,87
R21	RF	0,02	0,03	0,01	0,01	-	0,55	0,00	0,00
R22	RF	0,02	0,02	0,01	0,01	0,55	-	0,00	0,00
R23	SVM	0,57	0,36	0,79	0,50	0,00	0,00	-	0,01
R24	SVM	0,91	0,65	0,84	0,87	0,00	0,00	0,01	-

Tabela 4.4: Teste t para o índice $\overline{MBE}$ com 90% de confiança

$\overline{\text{MBE}}$		R17 MLP	R18 MLP	R19 MLP	R20 MLP	R21 RF	R22 RF	R23 SVM	R24 SVM
R17	MLP	-	0,85	0,17	0,64	0,41	0,37	0,00	0,00
R18	MLP	0,85	-	0,05	0,58	0,22	0,19	0,00	0,00
R19	MLP	0,17	0,05	-	0,57	0,09	0,09	0,00	0,00
R20	MLP	0,64	0,58	0,57	-	0,16	0,14	0,00	0,00
R21	RF	0,41	0,22	0,09	0,16	-	0,39	0,02	0,04
R22	RF	0,37	0,19	0,09	0,14	0,39	-	0,03	0,05
R23	SVM	0,00	0,00	0,00	0,00	0,02	0,03	-	0,03
R24	SVM	0,00	0,00	0,00	0,00	0,04	0,05	0,03	-

Tabela 4.5: Teste t para o índice $\overline{R}$ com 90% de confiança

$\overline{R}$		R17 MLP	R18 MLP	R19 MLP	R20 MLP	R21 RF	R22 RF	R23 SVM	R24 SVM
R17	MLP	-	0,73	0,55	0,99	0,00	0,00	0,30	0,17
R18	MLP	0,73	-	0,42	0,83	0,00	0,00	0,32	0,17
R19	MLP	0,55	0,42	-	0,64	0,00	0,00	0,19	0,09
R20	MLP	0,99	0,83	0,64	-	0,00	0,00	0,19	0,07
R21	RF	0,00	0,00	0,00	0,00	-	0,80	0,00	0,00
R22	RF	0,00	0,00	0,00	0,00	0,80	-	0,00	0,00
R23	SVM	0,30	0,32	0,19	0,19	0,00	0,00	-	0,17
R24	SVM	0,17	0,17	0,09	0,07	0,00	0,00	0,17	-

Analisando-se em especial a Tabela 4.2, podem-se fazer duas constatações:

1. Não existe uma real diferença estatística para os pares de modelos com e sem dias da semana, ainda que os primeiros tenham gerado IDs ligeiramente melhores;

2. Os modelos MLP e SVM, que apresentaram os melhores desempenhos, são estatísticamente diferentes dos de RF quanto ao ID $\overline{MSE}$, mas não podem ser diferenciados entre si.

No entanto, quando se analisam os resultados do teste t para outros IDs, especialmente o $\overline{MBE}$ e o $\overline{R}$, percebe-se que a diferença entre SVM e os melhores modelos de MLP é clara. Os modelos de SVM apresentam um desvio absoluto $\overline{MBE}$ muito maior do que os modelos de MLP. Além disso, os coeficentes $\overline{R}$ para os modelos R19 e R20 de MLP são ou considerados estísticamente diferentes daqueles dos SVM ou apresentam um valor p não maior do que 0,20 no teste t.

Portanto, pode-se concluir que de forma geral o modelo de MLP (R19) é o que apresenta melhor desempenho na previsão de níveis de O_3 para os casos de regressão seguido por SVM (R23) e RF (R21).

As Figuras 4.1, 4.2 e 4.3 apresentam a comparação dos valores calculados (*output*) pelos os valores medidos, ou experimentais (*target*), coletados pela estação Ibirapuera da CETESB, para o modelo R19 de MLP, o melhor obtido, R23 de SVM e R21 de RF, respectivamente.

Percebe-se que todos os modelos tendem a subestimar dados de alta concentração de O_3, ou seja, concentrações acima de 200 $\mu g/m^3$ no entanto o modelo MLP é o que apresenta maior coeficiente angular (0,76). Acredita-se que o pior desempenho

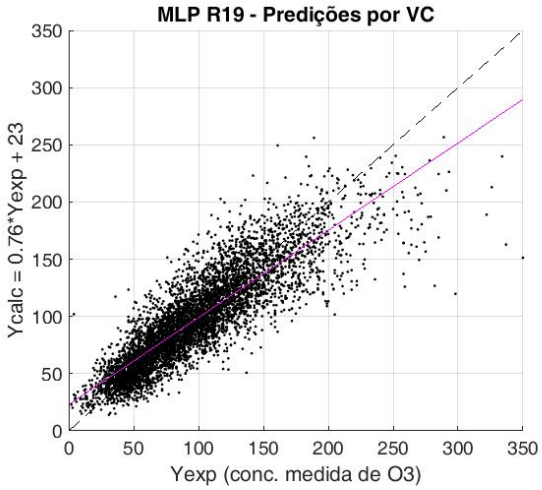

Figura 4.1: Comparação entre valores calculados pelo modelo R19 de MLP e os valores medidos

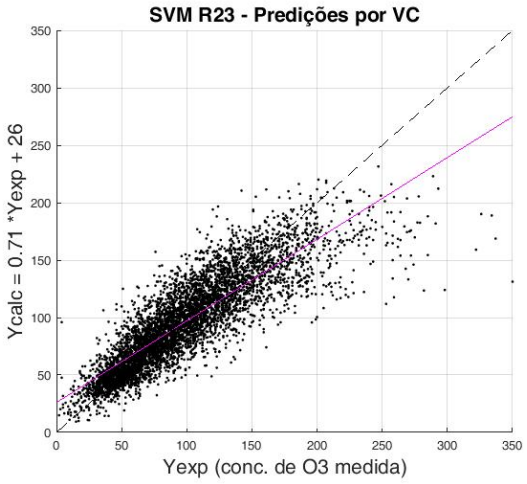

Figura 4.2: Coparação entre os valores calculados pelo modelo R23 de SVM e os valores medidos

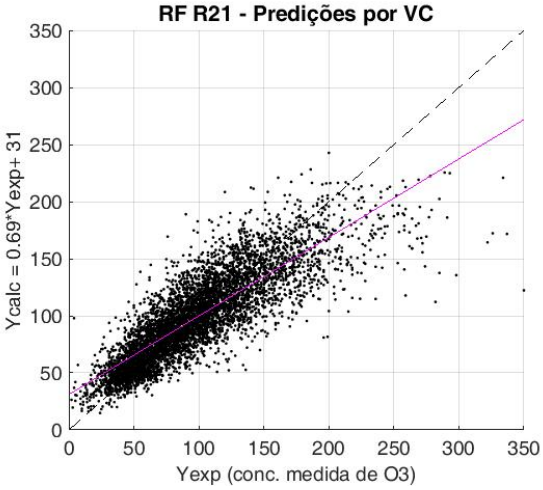

Figura 4.3: Comparação entre os valores calculados pelo modelo R21 de RF e os valores medidos

na previsão dos valores mais altos deve-se à pouca quantidade de observações desse tipo nas bases de dados, uma vez que os dias de ozônio muito alto são mais raros.

A Figura 4.4 traz a série temporal do valores previstos pelo modelo R19 de MLP, o melhor modelo, e dos valores experimentais para dados de Julho de 2015 até Agosto de 2016. Escolheu-se esse período de tempo para construção da série temporal por ele abranger um espaço de um ano e por se tratar dos dados mais recentes usados no modelo. Nota-se que, de forma geral, as curvas de dados calculados (Y_{calc}) e experimentais (Y_{exp}) apresentam o mesmo comportamento.

Análise de Sensibilidade

Uma vez obtido o modelo que melhor correlaciona as variáveis de entrada à concentração máxima diária de ozônio (médias horárias), é possível analisar a importância de cada variável de entrada na concentração final aplicando-se o método HIPR para o modelo R19 de MLP. Conforme explicado no capítulo Materiais e Métodos, este método verifica a ordem de importância das variáveis de entrada ao aplicar valores aleatórios para cada uma individualmente, verificando como estes valores impactam a medida de erro do modelo, ou seja, a MSE.

A ordem de importância das variáveis é apresentada pela Figura 4.5. Pode-se observar que as temperaturas e pressões impactam com maior intensidade os níveis de ozônio. É interessante notar também que a variável Vtarde, a componente da média da velocidade do vento a tarde na direção Norte-Sul, é mais importante que a componente Utarde (direção Leste-Oeste). Outro aspecto importante é o fato de a variável Dias da Semana afetar pouco os resultados, de acordo com o que foi verificado na Tabela 4.1.

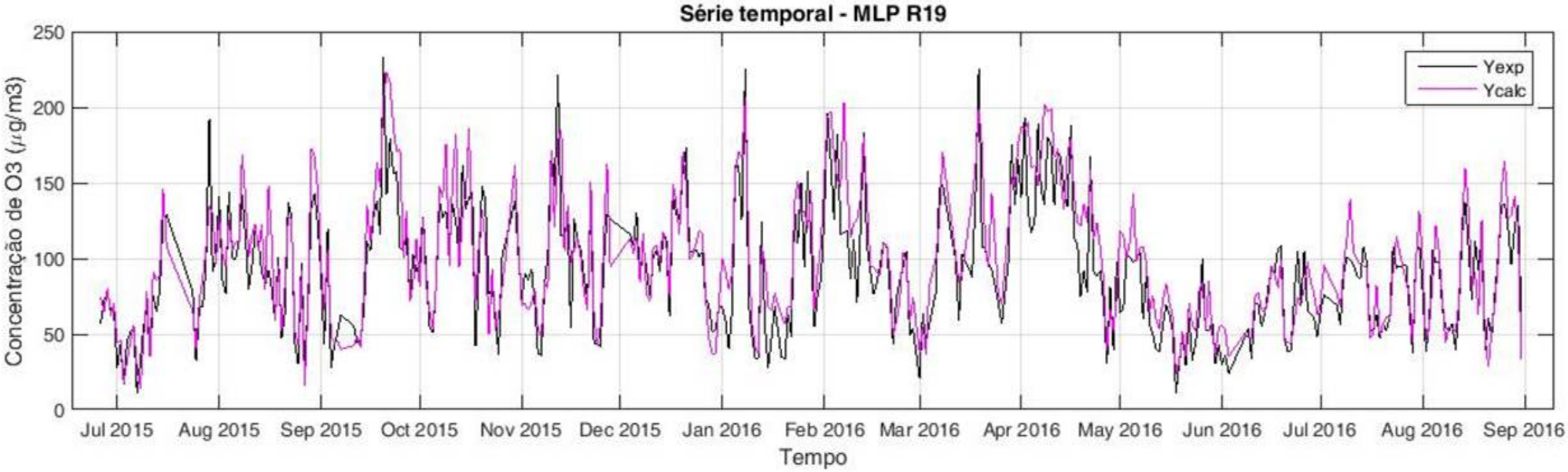

Figura 4.4: Série temporal dos valores calculados pelo modelo R19 de MLP e dos valores medidos

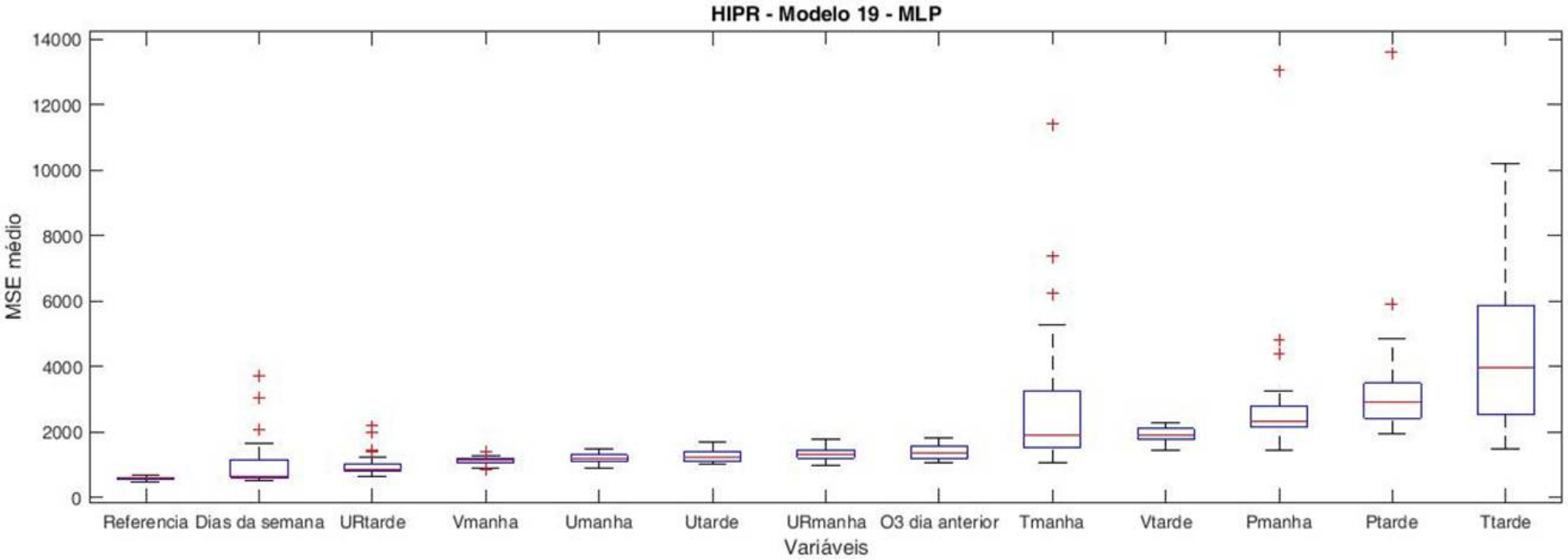

Figura 4.5: Aplicação do modelo HIPR para verificar a importância relativa das variáveis de entrada no modelo R19

4.1.2 Médias móveis de 8 horas

O melhor modelo, MLP, foi então aplicado na previsão de máximas diárias de O_3 considerando as médias móveis de 8 horas. As variáveis de entrada foram as mesmas da base de dados MMT. Neste caso, de forma semelhante aos anteriores, também foi aplicada a VC do tipo *10-fold* na determinação dos melhores parâmetros do modelo. Os resultados são apresentados pela Tabela 4.6 e pelas Figuras 4.7 e 4.6.

Tabela 4.6: Parâmetros ótimos para o modelo MLP considerando médias móveis de 8 horas para a concentração de O_3

Base de dados	Dias da Semana	Modelo	Parâmetros Ótimos	$\overline{MSE}$	$\overline{MAE}$	$\overline{MBE}$	$\bar{R}$
MMT médias móveis 8h	sim	MLP	14 neurônios; sig. - sig.	239	11,9	-0,11	0,8913

Percebe-se que os resultados obtidos para as máximas diárias considerando as médias móveis de 8 horas de O_3 foram melhores do que os resultados considerando médias horárias. A razão para isso é que os valores de médias móveis de 8 horas possuem variância menor quando comparados aos valores de médias horárias, como mostra a Figura 4.8.

4.1.3 Previsão para dias futuros

A última aplicação do modelo R19, o melhor modelo para o caso de regressão, foi feita para a previsão da concentração de ozônio para o dia posterior. Para tal,

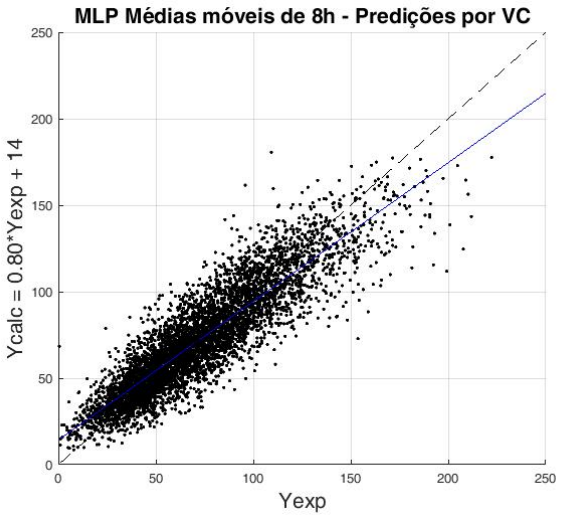

Figura 4.6: Comparação entre os valores calculados pelo modelo MLP e os valores medidos considerando médias móveis de 8 horas para a concentração de O_3

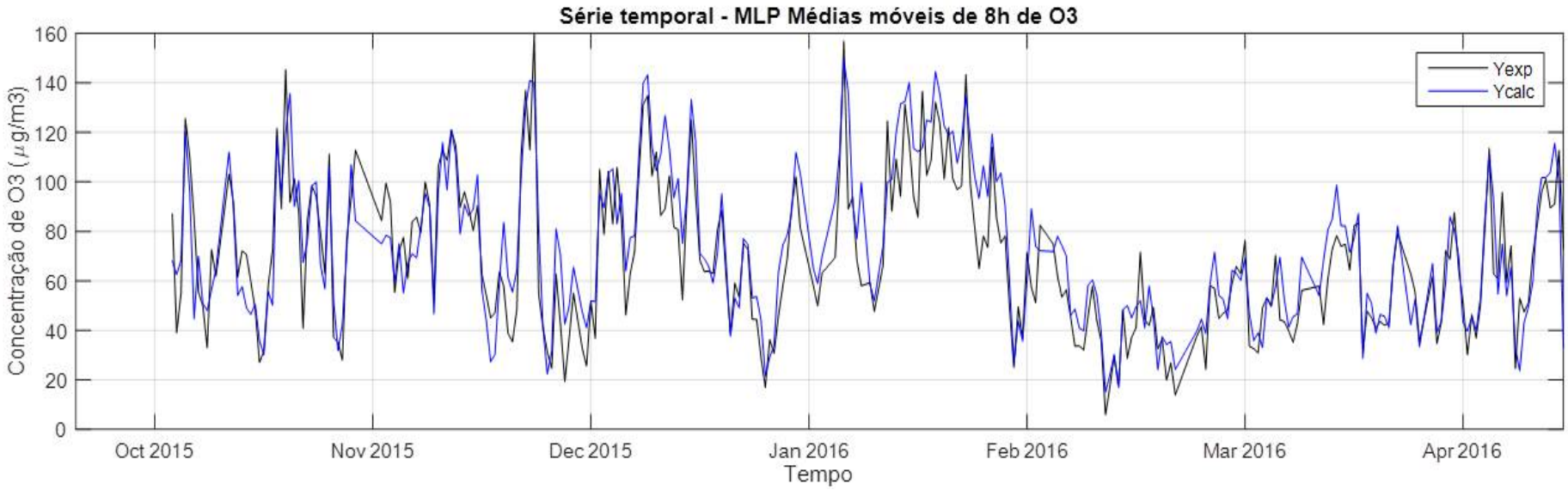

Figura 4.7: Série temporal dos valores calculados pelo modelo de MLP e dos valores experimentais considerando as médias móveis de 8 horas para a concentração de O_3

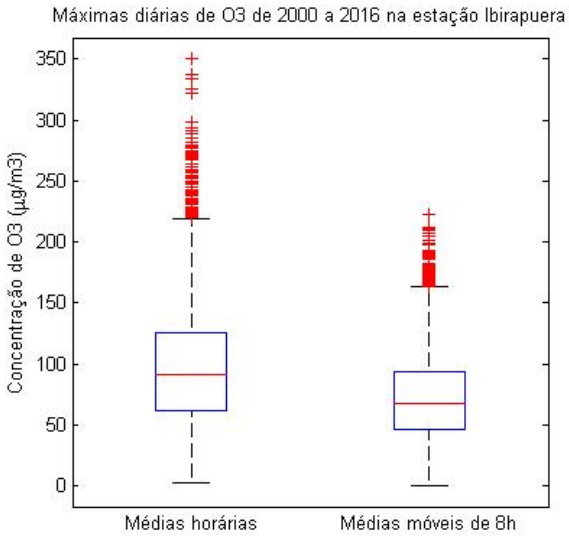

Figura 4.8: Concentrações máximas diárias de O_3 em médias horárias e médias móveis de 8 horas

utilizaram-se as previsões metereológicas do modelo ETA 40 km, modelo utilizado pelo Centro de Previsão do Tempo e Estudos Climáticos do Instituto Nacional de Pesquisas Espaciais (CPTEC-INPE). Os dados foram fornecidos pela CETESB.

A Figura 4.9 compara os valores previstos com os valores reais de concentração máxima diária de O_3 do dia 01 de Setembro de 2016 até o dia 30 de Novembro de 2016. Pode-se notar que as previsões não apresentam boa correlação com os valores reais de concentração de O_3. Atribui-se esse resultado ao fato de as previsões metereológicas não terem apresentado boa correlação com as condições metereológicas reais, como mostra o Anexo C.

A Figura 4.10 apresenta os resultados para o mesmo período de tempo utilizando as condições metereológicas reais como variáveis de entrada. A correlação neste caso foi alta, com $R^2 = 0,846$, tornando claro que o modelo desempenha bem o seu papel de correlacionar as variáveis de entrada com a concentração máxima de O_3.

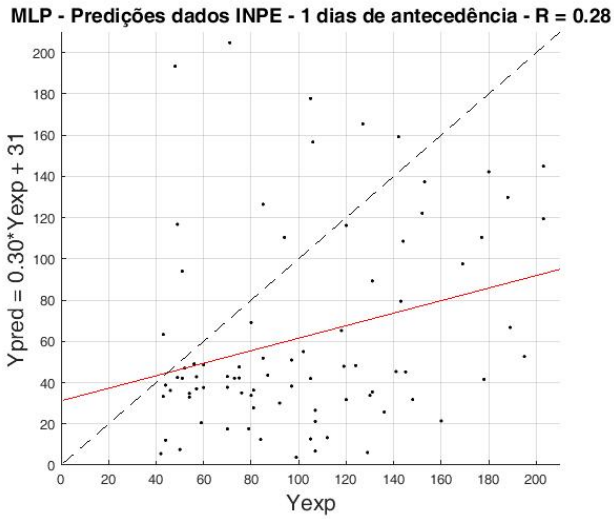

Figura 4.9: Previsão da concentração máxima de O_3 utilizando previsões metereológicas com 1 dia de antecedência como variáveis de entrada

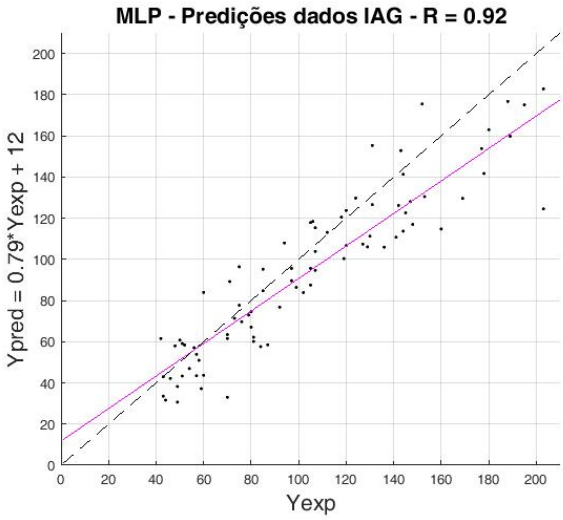

Figura 4.10: Previsão da concentração máxima de O_3 utilizando condições metereológicas reais medidas pela estação do IAG

4.2 Classificação

Nos casos de classificação foram testados os cinco modelos: RF, MLP, SVM, LDA e Fisher. Os cinco foram aplicados à base com as mesmas variáveis de entrada da base MMT, que foi a que gerou melhores resultados para os casos de regressão. A classificação das observações levou em conta os valores máximos diários das médias móveis de 8 horas.

Neste estudo não foi usado o modelo de discriminante quadrático devido ao pequeno número de observações no grupo de concentração mais alta de O_3 (grupo 'PÉSSIMA'). Como este grupo continha apenas 7 observações, a matriz de covariância

obtida para ele não é representativa de uma população inteira.

A Tabela 4.7 mostra os resultados obtidos para todos os modelos, em erro percentual de classificação, e as configurações ótimas.

Numa primeira análise, pode-se notar que o modelo C2 de MLP gerou o melhor resultado, e que todos os modelos de MLP, RF e SVM geraram resultados muito parecidos. As técnicas de LDA e Fisher geraram resultados iguais. Ao se analisar os resultados do teste t, apresentados pela Tabela 4.8, percebe-se que os resultados de MLP, RF e SVM são, efetivamente, iguais do ponto de vista estatístico. Eles diferem apenas dos resultados de LDA e Fisher, com exceção do modelo C1 de MLP.

Portanto não foi possível se verificar qual o melhor modelo para a análise de classificação dentre MLP, SVM e RF.

As Figuras 4.11 e 4.12 apresentam os resultados para o modelo C2 de MLP. A Figura 4.11 apresenta, considerando as classes dadas como resposta pelo modelo, as distribuição dos observações segundo suas classifcações reais. Nota-se que o modelo não classificou nenhuma das observações no grupo 'PÉSSIMA'.

Tabela 4.7: Resultados e configurações ótimas para os casos de classificação

	Nº do Modelo	Base de dados	Dias da Semana	Modelo	Parâmetros Ótimos	$\overline{\text{Erro}}$ %
1º	C2	MMTmm	não	MLP	6 neurônios; sigmoidal - softmax	14,95%
2º	C3	MMTmm	sim	RF	150 árvores; m = 8	15,13%
3º	C6	MMTmm	não	SVM	Kernel: radial; Custo = 1	15,26%
4º	C4	MMTmm	não	RF	150 árvores; m = 9	15,28%
5º	C1	MMTmm	sim	MLP	12 neurônios; sigmoidal - softmax	15,32%
6º	C5	MMTmm	sim	SVM	Kernel: radial; Custo = 1	15,33%
7º	C7	MMTmm	sim	LDA	-	16,30%
8º	C9	MMTmm	sim	Fisher	-	16,30%
9º	C10	MMTmm	não	Fisher	-	16,40%
10º	C8	MMTmm	não	LDA	-	16,40%

Tabela 4.8: Teste t para o índice de erro percentual com 90% de confiança

$\overline{\text{Erro}}$ %		C1 MLP	C2 MLP	C3 RF	C4 RF	C5 SVM	C6 SVM	C7 LDA	C8 LDA	C9 Fisher	C10 Fisher
C1	MLP	-	0,26	0,79	0,96	0,98	0,95	0,22	0,15	0,22	0,15
C2	MLP	0,26	-	0,73	0,53	0,64	0,68	0,10	0,07	0,10	0,07
C3	RF	0,79	0,73	-	0,44	0,67	0,74	0,03	0,03	0,03	0,03
C4	RF	0,96	0,53	0,44	-	0,91	0,96	0,06	0,06	0,06	0,06
C5	SVM	0,98	0,64	0,67	0,91	-	0,52	0,01	0,01	0,01	0,01
C6	SVM	0,95	0,68	0,74	0,96	0,52	-	0,01	0,01	0,01	0,01
C7	LDA	0,22	0,10	0,03	0,06	0,01	0,01	-	0,56	0,56	0,56
C8	LDA	0,15	0,07	0,03	0,06	0,01	0,01	0,56	-	0,56	0,26
C9	Fisher	0,22	0,10	0,03	0,06	0,01	0,01	0,56	0,56	-	0,56
C10	Fisher	0,15	0,07	0,03	0,06	0,01	0,01	0,56	0,26	0,56	-

A Figura 4.12 apresenta, considerando as classes reais dos dados, a porcentagem das classificações do modelo para cada grupo. O único grupo que apresentou um bom desempenho foi o grupo 'BOA', os demais apresentaram porcentagens maiores de dados classificados como pertencentes a outros grupos. Portanto os modelos de classificação estudados não geraram bons resultados.

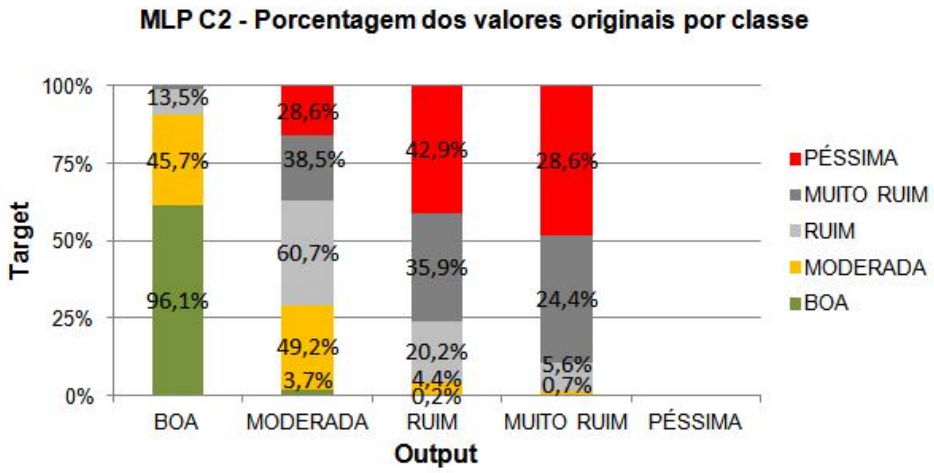

Figura 4.11: C2 MLP - Composição dos grupos definidos pelo modelo de acordo com as classificações originais (*target*)

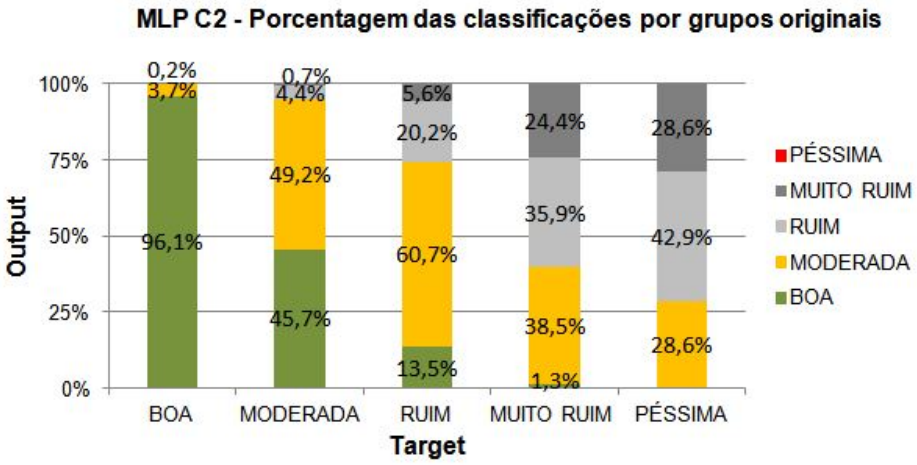

Figura 4.12: C2 MLP - Composição dos grupos originais de acordo com as classificações feitas pelo modelo (*output*)

.

5 Conclusões

A partir dos resultados de ajuste dos modelos multivariados apresentados, foram obtidas as conclusões listadas a seguir. A escolha de variáveis para previsão tem grande importância sobre o desempenho dos modelos de previsão de máxima diária de O_3. As variáveis de entrada que produziram os melhores ajustes foram as médias de manhã e tarde de temperatura, pressão, umidade relativa e as componentes da velocidade do vento nas direções Norte-Sul e Leste-Oeste, além da variável binária para classificação dos dias em dias da semana ou finais de semana (base MMT).

O estudo comparou diferentes modelos multivariados para a previsão da máxima concentração diária de ozônio usando apenas variáveis de entrada meteorológicas. Dentre todos os modelos, o *multilayer perceptron neural networks* (MLP) apresentou o melhor desempenho frente aos modelos RF e SVM. Ele apresentou melhores resultados para os indicadores média dos desvios quadráticos (MSE) e coeficiente de correlação (R). Para este último foi obtido o valor de 0,867, que é um valor alto para modelos aplicados à previsão de fenômenos ambientais. O modelo prevê bem dados de concentrações baixas e médias, mas tende a subestimar dados de concentrações altas, ou seja, concentrações maiores do que 200 $\mu g/m^3$.

Através do teste t para amostras pareadas, verificou-se que não houve impacto significativo da variável de dias da semana nos modelos. O mesmo fato foi constatado na análise de sensibilidade usando o método HIPR. A análise de sensibilidade apontou como principais variáveis as temperaturas e pressões, além da componente da velocidade do vento na direção Norte-Sul, a Vtarde.

A aplicação do MLP para a previsão de máximas de O_3 segundo médias móveis de 8 horas gerou um bom ajuste com relação aos dados medidos e um alto valor de R (0,8913).

A aplicação do modelo para previsão de concentrações máximas dos dias seguintes, utilizando como dados de entrada as previsões das variáveis metereológicas, não gerou bons resultados devido à baixa correlação das variáveis metereológicas previstas com as reais. No entanto, ao fazer a mesma previsão com as variáveis de entrada reais, o modelo teve excelente desempenho, podendo-se concluir então que o erro nas previsões para o dia seguinte é devido principalmente às limitações das previsões meteorológicas.

Nos casos de classificação, não houve diferença significativa entre os desempenhos dos modelos e nenhum dos modelos conseguiu classificar observações no grupo 'PÉSSIMA'. Este resultado deve-se ao fato de existirem muito raras observações de concentração máxima, segundo a média móvel de 8 horas, pertencentes a esse grupo. Por este mesmo motivo, não foi possível a aplicação de discriminantes quadráticos neste estudo, uma vez o que número de observações da classe 'PÉSSIMA' não era representativo da população. No entanto, em um estudo anterior do Departamento em conjunto com a CETESB em que se consideraram as médias horárias de concentração em vez das médias móveis de 8 horas, o modelo de discriminantes quadráticos foi aplicado e foi eficiente na classificação de dados do grupo de mais alta concentração de ozônio.

Assim, uma sugestão para os próximos passos seria aplicar o discriminante quadrático para classificar as máximas médias horárias em grupos e então aplicar um modelo de regressão por MLP para cada grupo individualmente. Essa abordagem poderia solucionar o problema da subestimação das concentrações mais altas de ozônio.

Uma outra sugestão é aplicar o modelo MLP de regressão e, a partir dos resultados, classificar as observações em grupos. Como o modelo de regressão fornece uma boa previsão, desde que os dados de meteorologia sejam também bem previstos, a classificação em grupos a partir dos seus resultados seria confiável.

Referências Bibliográficas

[Adv, 1989] (1989). *1989 Neural Information Processing Systems Conference*. Neural Information Processing Systems Foundation, Inc.

[EU, 2004] (2004). Directive 2004/42/ce of the european parliament and of the council. http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32004L0042.

[Man, 2017] (2017). The university of manchester website. http://www.cas.manchester.ac.uk/resactivities/atmosphericchemistry/topics/peroxyradical/.

[Abdul-Wahab and Al-Alawi, 2002] Abdul-Wahab, S. A. and Al-Alawi, S. M. (2002). Assessment and prediction of trophosperic ozone concentration levels using artificial neural networks. *Elsevier*, 17(3):219–228.

[Arhami et al., 2013] Arhami, M., Kamali, N., and Rajabi, M. M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by monte carlo simulation. *Environmental Science and Pollution Research*, 20:4777–4789.

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press.

[Borges et al., 2012] Borges, A. S., Andrade, M. F., and Guardani, R. (2012). Ground-level ozone prediction using a neural network model based on meteorological variables and applied to the metropolitan area of são paulo. *Int. J. of Environment and Pollution*, 49(1/2):1–15.

[Breiman, 1999] Breiman, L. (1999). *Bagging Predictors*. Kluwer Academic Publishers.

[Breiman, 2001] Breiman, L. (2001). Random forests. Technical report, UC Berkeley.

[Bridle, 1989] Bridle, J. S. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In [Adv, 1989], pages 211–217.

[Burrows et al., 1994] Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., and Thomson, B. (1994). Cart decision-tree statistical analysis and prediction of summer season maximum surface ozone for the vancouver, montreal and atlantic regions of canada. *Journal of Applied Meteorology*, 34:1848–1862.

[Bussab and Moretin, 2013] Bussab, W. O. and Moretin, P. A. (2013). *Estatística Básica*. Editora Saraiva, 8 edition.

[Carter, 1994] Carter, W. P. L. (1994). Development of ozone reactivity scales for volatile organic compounds. *J. of Air & Waste Management*, 44:881–899.

[CETESB, 2000] CETESB (2000). Estudo do comportamento do ozônio na região metropolitana de são paulo.

[CETESB, 2014] CETESB (2014). 2013 - qualidade do ar no estado de são paulo.

[CETESB, 2016b] CETESB (2016b). 2015 - qualidade do ar no estado de são paulo.

[CETESB, 2016a] CETESB (acesso em 01/06/2016a). http://ar.cetesb.sp.gov.br/poluentes/.

[Clapp and Jenkin, 2001] Clapp, L. J. and Jenkin, M. E. (2001). Analysis of the relationship between ambient levels of o3, no2 and no as function of nox in the uk. *Atmospheric Environment*, 35:6391–6405.

[Comrie, 1997] Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Wast Mangement*, 47(6):653–663.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Controls, Signals and Systems*, 2:303–314.

[Feng et al., 2011] Feng, Y., Zhang, W., and Sun, D. Z. (2011). Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. *Atmospheric Environment*, 45:441 – 459.

[Finlayson-Pitts and Pitts Jr., 2012] Finlayson-Pitts, B. J. and Pitts Jr., J. N. (2012). Atmospheric chemistry of tropospheric ozone formation: Scientific and regulatory implications. *Air & Waste*, 43(8):1091–1100.

[Guardani et al., 2003] Guardani, R., Aguiar, J. L., Nascimento, C. A. O., Lacava, C. I. V., and Yanagi, Y. (2003). Ground-level ozone mapping in large urban areas using multivariate statistical analysis: Application to the são paulo metropolitan area. *J. of Air & Waste Management*, 53:553–559.

[Guardani and Nascimento, 2004] Guardani, R. and Nascimento, C. A. O. (2004). Neural network-based study for predicting ground-level ozone concentration in large urban areasm applied to the são paulo metropolitan area. *Int. J. Environment and Pollution*, 22(4):1–19.

[Guardani and Nascimento, 2013] Guardani, R. and Nascimento, C. A. O. (2013). Analise estatistica multivariada aplicada a processos quimicos. Technical report, Escola Politecnica da Universidade de Sao Paulo.

[Guardani et al., 1999] Guardani, R., Nascimento, C. A. O., Guardani, M. L. G., Martins, M. H. R. B., and Romano, J. (1999). Study of atmospheric ozone formation by means of a neural network-based model. *J. of Air & Waste Management*, 49(3):316–323.

[Hastie et al., 2008] Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning*. Springer.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

[Johnson and Wichern, 2007] Johnson, R. A. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 6 edition.

[Kemp et al., 2007] Kemp, S. J., Zaradic, P., and Hansen, F. (2007). An approach for determining the relative input parameter importance and significance in artificial neural networks. *Elsevier*, 204:326–334.

[Lorena and Carvalho, 2007] Lorena, A. C. and Carvalho, A. C. P. L. F. (2007). Uma introdução às support vector machines. *RITA*, XIV(2):43–67.

[Lu and Wang, 2008] Lu, W. and Wang, D. (2008). Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. *Elsevier*, 395:109–116.

[Marquardt, 1963] Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441.

[Martins et al., 2008] Martins, L. D., Andrade, M. F., and Ynoue, R. Y. (2008). Ambiental volatile organic compounds in the megacity of são paulo. *Quim. Nova*, 31(8):2009–2013.

[Meyer et al., 2015] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., and Lin, C. (2015). Package 'e1071'. Technical report, Department of Statistics, Probability, Theory Group, TU Wien.

[Ng, 2016] Ng, A. (2016). Support vector machine. Technical report, Stanford University.

[Ordieres et al., 2005] Ordieres, J. B., Vergara, V. P., Capuz, R. S., and Salazar, R. (2005). Neural network prediction model for fine particulate matter (pm 2,5) on the us-mexico borderin el paso (texas) and ciudad juárez (chihuahua). *Elsevier*, 20:547 – 559.

[Ripley, 1996] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Press Syndicate of The University of Cambridge, 1 edition.

[Robeson and Steyn, 1990] Robeson, S. M. and Steyn, D. G. (1990). Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment*, 24B(2):303–312.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(9):533 – 536.

[Seinfeld and Pandis, 2006] Seinfeld, J. H. and Pandis, S. N. (2006). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, volume 1993. Wiley.

[Smola and Scholkopf, 1998] Smola, A. J. and Scholkopf, B. (1998). A tutorial on support vector regression. Technical report, NeuroCOLT.

[Sucar et al., 1997] Sucar, L. E., Pérez-Brito, J., Ruiz-Suárez, J. C., and Moralez, E. (1997). Learning structure from data and its application to ozone prediction. *Applied Intelligence*, 7(4):327–338.

[Sundaramoorthi, 2014] Sundaramoorthi, D. (2014). A data-integrated simulation model to forecast ground-level ozone concentration. *Annal of Operations Research*, 216:53 – 59.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

[WHO, 2016] WHO (acesso em 2016). http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/.

A Resultados e Configurações dos Modelos de Regressão

Nº do Modelo	Base de dados	N	Dias da Semana	Modelo	Parâmetros Ótimos	$\overline{MSE}$	dp $\overline{MSE}$	$\overline{MAE}$	dp $\overline{MAE}$	$\overline{MBE}$	dp $\overline{MBE}$	$\overline{R}$	dp $\overline{R}$
R1	MMM	5594	sim	MLP	12 neurônios; sigmoidal - linear	621	65,9	18,1	0,58	0,04	0,798	0,8517	0,008
R2	MMM	5594	não	MLP	13 neurônios; sigmoidal - linear	622	74,9	18,1	0,75	0,02	0,774	0,8515	0,011
R3	MMM	5594	sim	MLP	14 neurônios; sig. - sig.	619	66,4	18,0	0,59	-0,11	0,890	0,8524	0,008
R4	MMM	5594	não	MLP	14 neurônios; sig. - sig.	622	71,8	18,1	0,63	-0,19	0,938	0,8514	0,010
R5	MMM	5594	sim	RF	150 árvores; m = 2	657	76,0	18,5	0,80	0,44	0,700	0,8424	0,012
R6	MMM	5594	não	RF	150 árvores; m = 2	664	80,7	18,6	0,80	0,47	0,783	0,8407	0,012
R7	MMM	5594	sim	SVM	Kernel: radial; Custo = 1	642	87,9	18,0	0,79	2,21	0,758	0,8480	0,013
R8	MMM	5594	não	SVM	Kernel: radial; Custo = 1	643	86,4	18,1	0,79	2,22	0,759	0,8478	0,013
R9	MtMM	5603	sim	MLP	9 neurônios; sigmoidal - linear	652	74,0	18,4	0,78	-0,20	1,281	0,8444	0,010
R10	MtMM	5603	não	MLP	13 neurônios; sigmoidal - linear	657	78,0	18,5	0,71	-0,06	1,302	0,8430	0,011
R11	MtMM	5603	sim	MLP	11 neurônios; sig. - sig.	646	71,2	18,3	0,69	-0,03	1,129	0,8457	0,010
R12	MtMM	5603	não	MLP	11 neurônios; sig. - sig.	656	72,0	18,4	0,72	-0,22	0,972	0,8429	0,009
R13	MtMM	5603	sim	RF	150 árvores; m = 2	683	71,7	18,8	0,76	0,55	0,917	0,8360	0,009
R14	MtMM	5603	não	RF	150 árvores; m = 2	690	80,6	18,9	0,81	0,54	0,952	0,8340	0,011
R15	MtMM	5603	sim	SVM	Kernel: radial; Custo = 1	666	79,7	18,3	0,72	2,29	1,128	0,8423	0,011
R16	MtMM	5603	não	SVM	Kernel: radial; Custo = 1	668	79,0	18,3	0,72	2,12	1,157	0,8414	0,011
R17	MMT	5598	sim	MLP	14 neurônios; sigmoidal - linear	566	65,3	17,4	0,68	0,03	1,498	0,8662	0,009
R18	MMT	5598	não	MLP	15 neurônios; sigmoidal - linear	566	69,4	17,5	0,68	-0,01	1,010	0,8658	0,012
R19	MMT	5598	sim	MLP	11 neurônios; sig. - sig.	560	70,8	17,2	0,78	-0,35	1,020	0,8674	0,012
R20	MMT	5598	não	MLP	14 neurônios; sig. - sig.	564	57,2	17,4	0,56	-0,18	1,173	0,8662	0,011
R21	MMT	5598	sim	RF	150 árvores; m = 8	655	79,0	18,6	0,81	0,56	0,959	0,8434	0,011
R22	MMT	5598	não	RF	150 árvores; m = 6	657	81,1	18,6	0,80	0,60	1,053	0,8430	0,012
R23	MMT	5598	sim	SVM	Kernel: radial; Custo = 1	590	83,3	17,1	0,86	2,28	1,079	0,8619	0,013
R24	MMT	5598	não	SVM	Kernel: radial; Custo = 1	593	81,2	17,3	0,84	2,06	1,031	0,8607	0,012

B Resultados e Configurações dos Modelos de Classificação

Abaixo são apresentados todos os testes efetuados para os casos de classificação.

Nº do Modelo	Base de dados	N	Dias da Semana	Modelo	Parâmetros Ótimos	$\overline{\text{Erro}}$ %	dp $\overline{\text{Erro}}$ %
C1	MMTmm	5492	sim	MLP	12 neurônios; sigmoidal - softmax	15,32%	1,44%
C2	MMTmm	5492	não	MLP	6 neurônios; sigmoidal - softmax	14,95%	1,07%
C3	MMTmm	5492	sim	RF	150 árvores; m = 8	15,13%	0,92%
C4	MMTmm	5492	não	RF	150 árvores; m = 9	15,28%	0,95%
C5	MMTmm	5492	sim	SVM	Kernel: radial; Custo = 1	15,33%	1,75%
C6	MMTmm	5492	não	SVM	Kernel: radial; Custo = 1	15,26%	1,60%
C7	MMTmm	5492	sim	LDA	-	16,30%	1,49%
C8	MMTmm	5492	não	LDA	-	16,40%	1,55%
C9	MMTmm	5492	sim	Fisher	-	16,30%	1,49%
C10	MMTmm	5492	não	Fisher	-	16,40%	1,55%

C Comparação das Variáveis Metereológicas Previstas com as Medidas

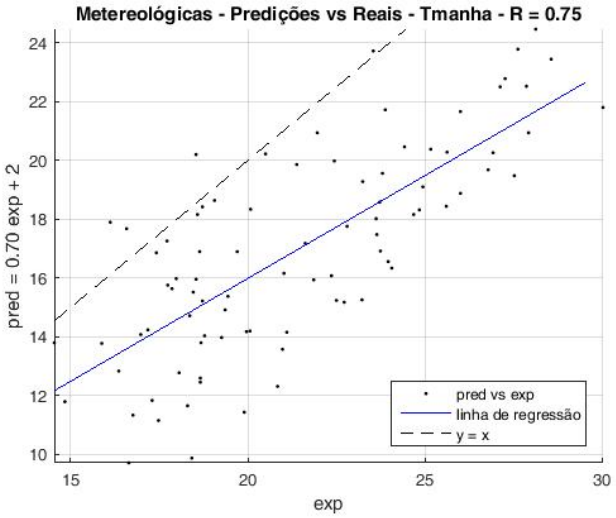

Figura C.1: Comparação dos valores previstos da variável metereológica Tmanha (modelo ETA 40 km) com os valores reais

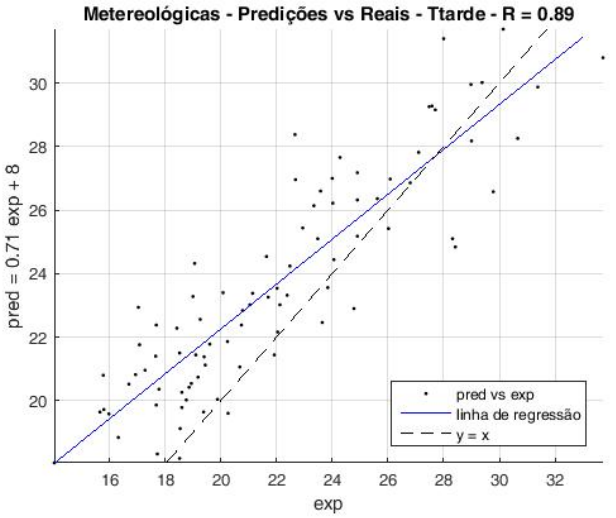

Figura C.2: Comparação dos valores previstos da variável metereológica Ttarde (modelo ETA 40 km) com os valores reais

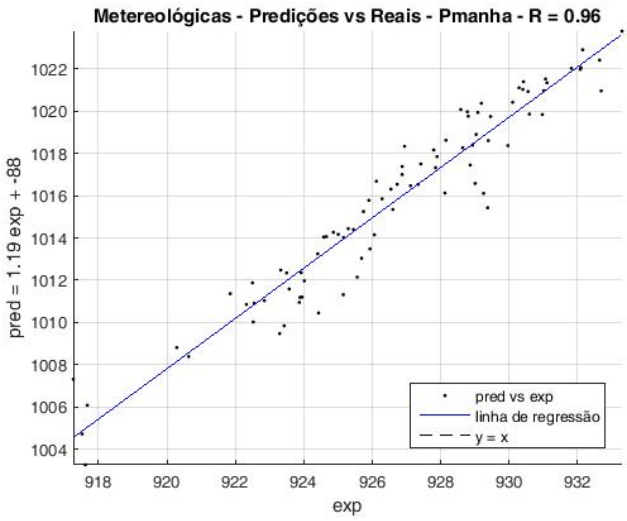

Figura C.3: Comparação dos valores previstos da variável metereológica Pmanha (modelo ETA 40 km) com os valores reais

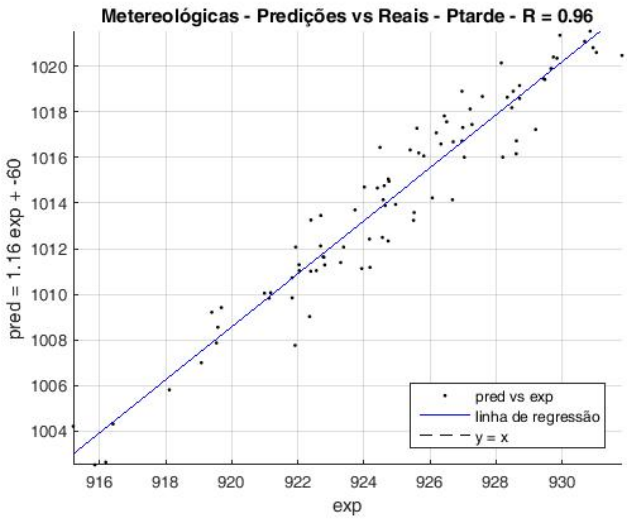

Figura C.4: Comparação dos valores previstos da variável metereológica Ptarde (modelo ETA 40 km) com os valores reais

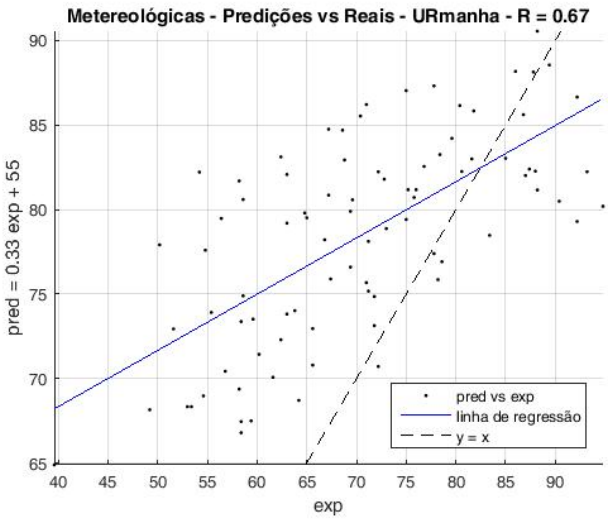

Figura C.5: Comparação dos valores previstos da variável metereológica URmanha (modelo ETA 40 km) com os valores reais

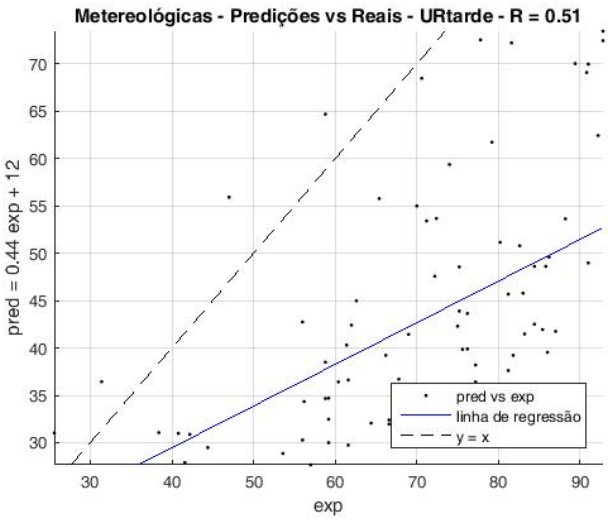

Figura C.6: Comparação dos valores previstos da variável metereológica URtarde (modelo ETA 40 km) com os valores reais

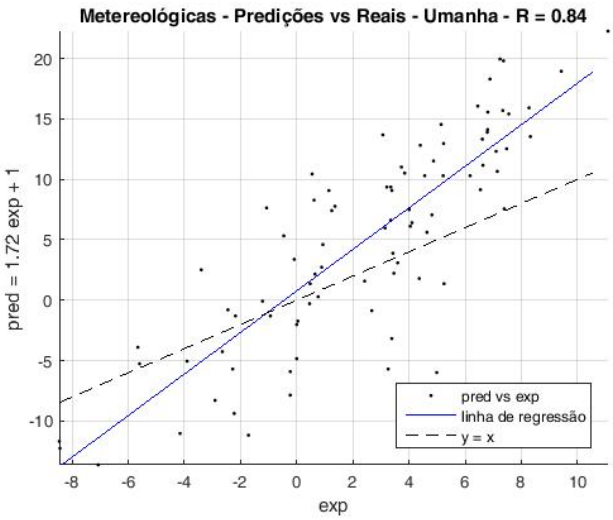

Figura C.7: Comparação dos valores previstos da variável metereológica Umanha (modelo ETA 40 km) com os valores reais

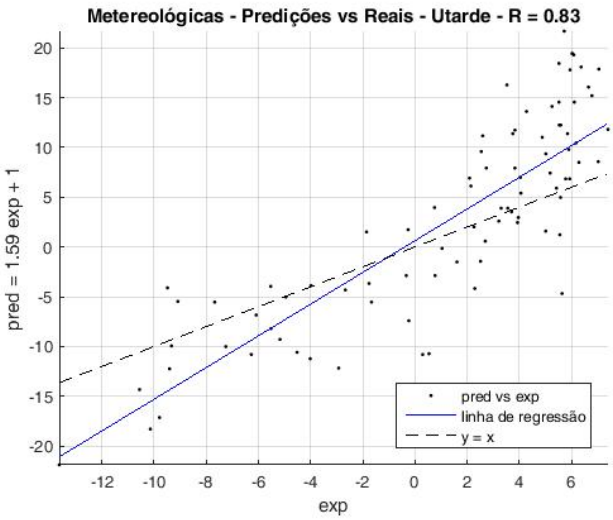

Figura C.8: Comparação dos valores previstos da variável metereológica Utarde (modelo ETA 40 km) com os valores reais

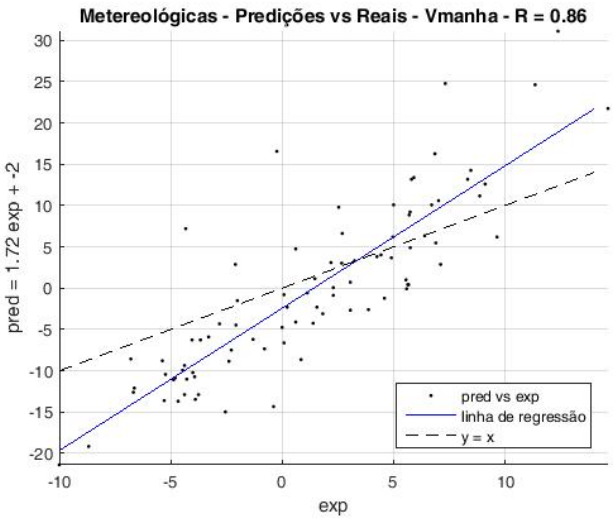

Figura C.9: Comparação dos valores previstos da variável metereológica Vmanha (modelo ETA 40 km) com os valores reais

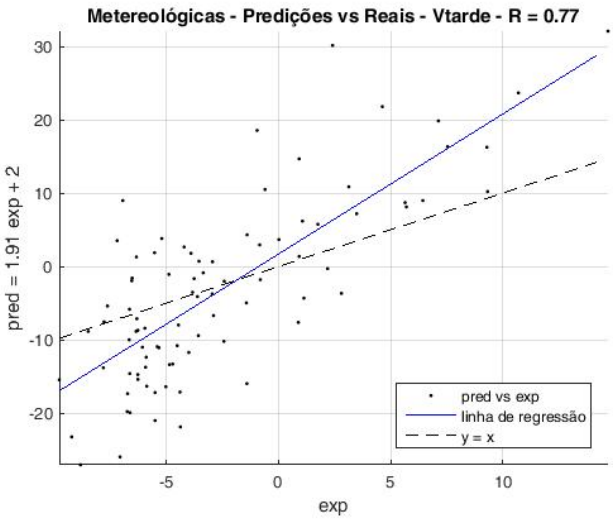

Figura C.10: Comparação dos valores previstos da variável metereológica Vtarde (modelo ETA 40 km) com os valores reais