Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

MARCELO DA SILVA SOUSA

PREVISÃO DE VALOR VENAL DE IMÓVEIS POR MEIO DE APRENDIZADO DE MÁQUINA:
O caso da Secretaria de Patrimônio da União

Brasília
2020

MARCELO DA SILVA SOUSA

PREVISÃO DE VALOR VENAL DE IMÓVEIS POR MEIO DE APRENDIZADO DE MÁQUINA:

O caso da Secretaria de Patrimônio da União

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.
Orientador: Prof. Msc. Saul Berardo

Brasília

2020

REFERÊNCIA BIBLIOGRÁFICA

Sousa, Marcelo. **Previsão de valor venal de imóveis por meio de aprendizado de máquina**: o caso da Secretaria de Patrimônio da União. 2020. Trabalho de Conclusão de Curso da Pós-Graduação *lato sensu* em Análise de Dados para o Controle – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF. 43 fl.

CESSÃO DE DIREITOS

NOME DO AUTOR: Marcelo da Silva Sousa
TÍTULO: Previsão de valor venal de imóveis por meio de aprendizado de máquina: o caso da Secretaria de Patrimônio da União
GRAU/ANO: Especialista/2020

Marcelo da Silva Sousa
marcelo.pacote@tcu.gov.br

Ficha catalográfica

Sousa, Marcelo da Silva Previsão de valor venal de imóveis por meio de aprendizado de máquina: O caso da Secretaria de Patrimônio da União / Marcelo da Silva Sousa; orientador, Saul Berardo, 2020. 43 p. Monografia (especialização) - Escola Superior do Tribunal de Contas da União, Curso de Pós-Graduação lato sensu em Análise de Dados para o Controle, Brasília, 2020. Inclui referências. 1.Controle Externo. 2. Análise de Dados. 3. Aprendizado de máquina. 4. *Machine Learning*. 5. Modelos preditivos. I. Berardo, Saul. II. Escola Superior do Tribunal de Contas da União. Pós-Graduação lato sensu em Análise de Dados para o Controle.

MARCELO DA SILVA SOUSA

PREVISÃO DE VALOR VENAL DE IMÓVEIS POR MEIO DE APRENDIZADO DE MÁQUINA:

O caso da Secretaria de Patrimônio da União

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle, realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.

Brasília, 31 de março de 2020.

Banca Examinadora:

_____

Prof.ª Saul Berardo, Mestre.

Orientador

Universidade do Pará

_____

Prof. Eduardo Chaves Ferreira, Doutor.

Laboratório Nacional de Computação Científica

RESUMO

A Secretaria de Patrimônio da União (SPU) detém cerca de meio milhão de imóveis disponíveis para alienação. Em seus editais de concorrência pública, imóveis anunciados com valores muito acima ou muito abaixo do preço de mercado podem representar indício de operação fraudulenta. Faz-se, pois, necessário fiscalizar o valor de venda do patrimônio da União. Este trabalho propõe um modelo preditivo que permite determinar o valor venal de imóveis da Secretaria de Patrimônio da União. Apresenta-se diferentes estratégias para criação de modelos, baseadas na setorização de regiões do país por meio do uso do Código de Endereçamento Postal (CEP). Utiliza-se a ferramenta *SAS Enterprise Miner* para avaliar diversos algoritmos de aprendizado de máquina como árvores de decisão, redes neuronais, regressão, *gradient boosting* e *emsemble*. Mostra-se a aplicação do modelo construído para previsão do valor de venda de apartamentos em edital de Concorrência Pública da Secretaria de Patrimônio da União.

Palavras-chave: Mineração de dados. Análise de dados. Aprendizado de máquina. Modelo preditivo. Árvore de decisão. *Gradient Boosting*. Regressão. *Emsemble*. Rede Neuronal. Previsão de preços de imóveis. *SAS Enterprise Miner*.

ABSTRACT

The Secretaria de Patrimônio da União (Federal Patrimony Secretariat - SPU) holds approximately half a million properties available for sale. In public bidding documents, properties far above or far below the market price may indicate fraudulent operation. Therefore, it is necessary to monitor the sale value of Union's assets. This work proposes a predictive model that allows determining the venal value of properties of SPU. We present different strategies for creating models based on sectorization of regions of the country through the use of the Postal Address Code (CEP). SAS Enterprise Miner tool is used to evaluate several machine learning algorithms such as decision trees, neural networks, regression, gradient boosting and emsemble. This work shows the usage of the model generated for forecasting the sale value in a public bidding notice of SPU.

Keywords: Data Mining. Data Analysis. Machine Learning. Predictive Model. Decision Trees. *Gradient Boosting*. Regression. *Emsemble*. Neural Network. House Price Prediction. *SAS Enterprise Miner*.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS (opcional)

SUMÁRIO

1 INTRODUÇÃO

A Secretaria de Patrimônio da União (SPU) detém cerca de meio milhão de imóveis disponíveis para alienação. Políticas econômicas do governo podem determinar aumento expressivo do número de imóveis a serem vendidos. Imóveis anunciados com preço muito acima ou muito abaixo do preço de mercado podem ser indício de operação fraudulenta. Faz-se, pois, necessário fiscalizar o valor de venda do patrimônio da União.

Precificar corretamente o valor de venda de um imóvel é um desafio. Trata-se de tarefa que demanda esforço em termos de tempo e capacidade de análise. Profissionais especializados levam em consideração dezenas de variáveis como localização, estado de conservação, tamanho do imóvel, infraestrutura, acabamento e tempo de vida da construção para determinar o valor de um imóvel.

No âmbito do TCU, a fiscalização da atividade supradita é realizada por amostragem ou em virtude de denúncias de editais com suspeita de fraude. Caso o número de imóveis alienados aumente de forma substancial, a estrutura do Tribunal não poderia dedicar recursos humanos suficientes para acompanhar este crescimento. Em um cenário de restrições orçamentárias e, em especial, com a diminuição do número de auditores, a fiscalização realizada unicamente por seres humanos não é escalável.

Diante do exposto, o presente trabalho apresenta um modelo preditivo capaz de precificar imóveis da União em todo território nacional.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Desenvolver modelo preditivo que permita determinar valor venal de imóveis da Secretaria de Patrimônio da União.

1.1.2 Objetivos Específicos

a) identificar imóveis cuja alienação oferece maior risco haja vista valor diferente do praticado no mercado;
b) priorizar imóveis a serem fiscalizados, quando de sua alienação;

2 DESENVOLVIMENTO

A seção 2.1 dispõe sobre o referencial teórico associado ao aprendizado de máquina e aos modelos preditivos utilizados neste trabalho. A seção 2.2 explica todos os passos realizados no trabalho, iniciando pelo entendimento do negócio (seção 2.2.1) e dos dados (seção 2.2.2), a preparação dos dados (seção 2.2.3) e, finalmente, a modelagem (seção 2.2.4). Os resultados e a discussão estão na seção 2.3.

2.1 FUNDAMENTAÇÃO TEÓRICA

2.1.1 Aprendizado de máquina

Conforme descrito na seção 1.1.1, trata-se de trabalho cujo objetivo é desenvolver modelo a partir do qual é possível usar dados existentes para prever possíveis saídas para dados novos. A criação e uso de modelos que aprendem a partir de dados é o cerne do aprendizado de máquina. Segundo (FACELI, 2015):

> Em Aprendizado de Máquina, computadores são programados para aprender com a experiência passada. Para tal, empregam um princípio denominado indução, no qual se obtêm conclusões genéricas a partir de um conjunto particular de exemplos. Assim, algoritmos de Aprendizado de Máquina aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de dados que representam instâncias do problema a ser resolvido.

De acordo com (KONAR, 1999) o Aprendizado de Máquina pode ser classificado em três categorias: i) supervisionada, ii) não supervisionada e iii) por reforço.

Os métodos supervisionados buscam descobrir a relação entre a entrada de atributos (variáveis independentes) e uma classe alvo (variável dependente). A relação descoberta é representada por meio de um modelo. Os modelos descrevem e explicam os fenômenos que estão escondidos no conjunto de dados e podem ser usados para prever o valor da classe alvo a partir do conhecimento dos valores dos atributos de entrada. (MAIMON e ROKACH, 2005)

Consoante (HAN, KAMBER e PEI, 2011), há dois modelos de aprendizagem supervisionada, os modelos de classificação, no qual a predição é feita para um atributo classificador que assume valores discretos e os modelos de regressão, no qual a variável alvo é contínua. O modelo de regressão foi utilizado neste trabalho tendo em vista a natureza do problema a ser resolvido. O detalhamento será apresentado na seção 2.2

Há diversos mecanismos que implementam modelos de regressão. As seções 2.1.2, 2.1.3, 2.1.4 e 2.1.5 descrevem aqueles que serão empregados na modelagem dos dados (seção 2.2.4).

2.1.2 Árvores de decisão

As árvores de decisão fazem uma representação gráfica baseada em hierarquias para identificar grupos de indivíduos com características de interesse comuns. Utiliza mecanismo recursivo que divide a amostra inicial em subamostras, baseando-se em resultados observados das variáveis independentes e em suas interações. Formam-se, portanto, grupos para os quais a variável resposta apresenta comportamento homogêneo dentro dos grupos e heterogêneo entre eles (BREIMAN, FRIEDMAN, *et al.*, 1984).

Uma árvore de decisão é chamada de Árvore de Classificação se a variável dependente for categórica, ou Árvore de Regressão, caso seja numérica (TACONELI, 2008). Neste trabalho, foram utilizadas Árvores de Regressão, tendo em vista que a variável dependente é valor de venda de um imóvel da União.

Há diversos algoritmos para indução de árvores de decisão. São exemplos: CHAID (*Chi-square Automatic Interaction Detection*) (KASS, 1980), CART (*Classification and Regression Trees*) (BREIMAN, FRIEDMAN, *et al.*, 1984), ID3 (*Iteractive Dichotomizer 3*) (QUINLAN, 1986) e C4.5 (QUINLAN, 1993). De forma geral, o processo de indução de árvores inicia por meio de uma amostra, nomeada nó raiz. Ela é dividida em subamostras, denominadas nós filhos. Essas subamostras, ao serem subdivididas, são chamadas de nós pais haja vista gerarem nós filhos. Quando uma subamostra não puder mais ser subdivida, conforme algum critério de parada, é então denominada de nó final ou nó folha.

Dentre os algoritmos supramencionados, foi escolhido o CART para este trabalho. A escolha foi motivada pelo fato de o algoritmo não fazer restrição quanto à natureza das variáveis explicativas, por gerar resultados de fácil interpretação e por estar implementado na ferramenta SAS Enterprise Miner.

2.1.3 Regressão

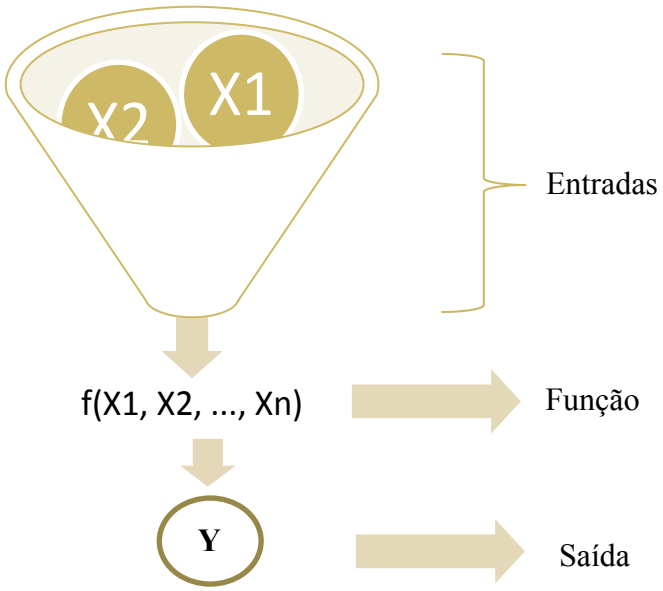
Trata-se de método de modelagem que avalia a relação entre uma variável dependente contínua Y e uma ou mais variáveis independentes X1, X2..., Xk. O objetivo da análise de regressão é identificar uma função que descreve, da melhor forma, a relação entre essas variáveis para que se possa prever qual valor a variável dependente assumirá quando forem atribuídos valores para as variáveis independentes (RAGSDALE, 2001).

A função supradita pode ser do tipo linear (equação da reta ou do plano) ou não linear (equação exponencial, geométrica *etecetera*). Ademais, pode ser simples (apenas uma variável independente) ou multivariada, quando há múltiplas variáveis independentes para análise. Na seção 2.2.4 será utilizada a regressão linear multivariada para previsão do valor de venda de imóveis. A seção 2.1.3.1 descreve brevemente a regressão linear múltipla.

2.1.3.1 Regressão linear múltipla

Introduzida por (STERNBERG, STILLO e SCHWENDEMAN, 1960), a Regressão Linear Múltipla (RLM) é adequada para casos nos quais o problema de pesquisa envolve uma variável dependente (alvo) relacionada a duas ou mais variáveis independentes. A Figura 1 ilustra o relacionamento entre a variável dependente, as diversas variáveis independentes e a função que descreve a relação entre as variáveis.

Figura 1 - Relação entre variáveis independentes, função preditiva e saída



Fonte: Elaborada pelo autor (2020).

Consoante (HAIR, ANDERSON, *et al.*, 2009), o modelo teórico de regressão linear múltipla é descrito pela seguinte equação linear:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Onde:

- Y é o valor da variável dependente.
- Xi são os valores das variáveis independentes (constantes conhecidas).
- β_i são parâmetros ou coeficientes de regressão.
- ε é o erro aleatório do modelo.

2.1.4 Redes Neuronais

De acordo com Muller e Fill (2003), as Redes Neuronais Artificiais (RNA) são agrupamentos de unidades de processamento (neurônios), interconectadas e estruturadas, cujo funcionamento é similar ao de uma estrutura neural de organismo inteligente.

As RNA extraem seu poder computacional da capacidade de distribuição de sua estrutura de alto grau de paralelismo e de sua capacidade de aprender/generalizar, tornando

possível a resolução de problemas complexos em diversas áreas do conhecimento humano. (HAIKIN, 2001)

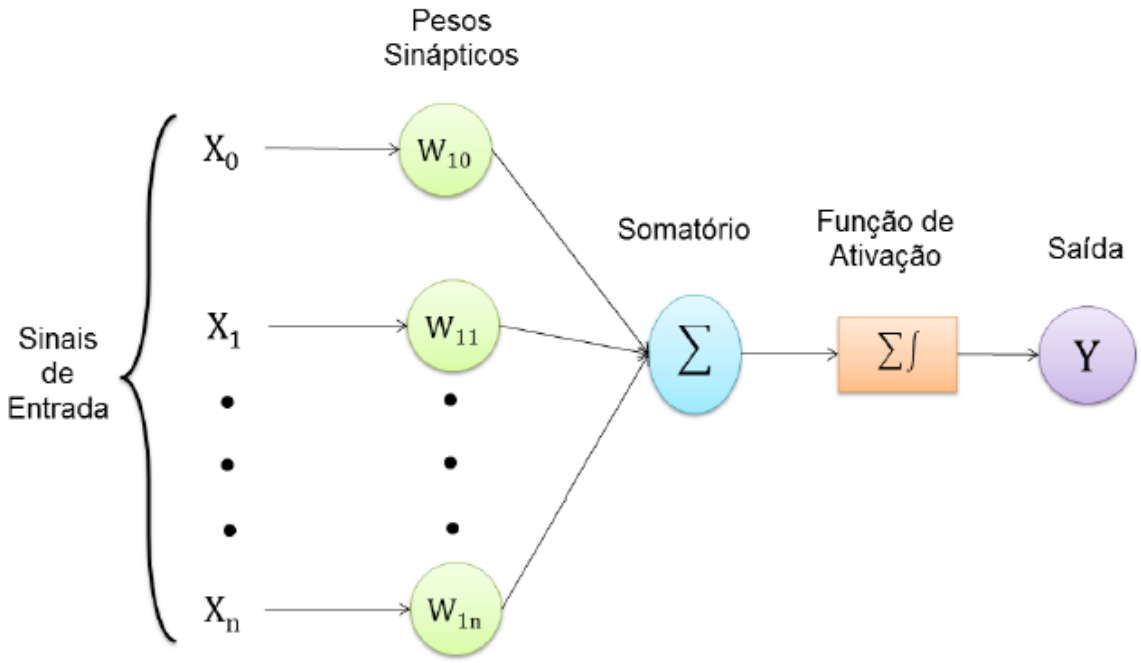
O mesmo autor acrescenta que as redes RNA são sistemas paralelos distribuídos, compostas por unidades básicas de processamento responsáveis por calcular funções matemáticas. Estas unidades são dispostas em uma ou mais camadas e interligadas por um número expressivo de conexões. Na maioria dos modelos, essas conexões têm pesos associados que, após o processo de aprendizagem, armazenam o conhecimento adquiridos pela rede. O funcionamento dessas redes é inspirado em estrutura conhecida do ponto de vista biológico: o cérebro humano. (HAIKIN, 2001)

Uma RNA se parece com o cérebro tendo em vista que:

i. o conhecimento é adquirido por meio de um processo de aprendizagem e
ii. são utilizadas forças de conexão entre os neurônios (pesos sinápticos), para armazenar o conhecimento adquirido. (HAIKIN, 2001)

Em 1943, os pesquisadores W.S. McCulloch e W.H. Pitts idealizaram o modelo matemático de um neurônio artificial. Assim, segundo Mcculloch e Pitts (1943), o neurônio é composto de conexões que emulam os dendritos, pesos que emulam as sinapses, uma função de mapeamento emulando o corpo celular e uma saída emulando um axônio. A Figura 2 ilustra o referido modelo matemático. Na figura, a função de ativação se refere a uma função interna, que determina o que fazer com o valor resultante do somatório das entradas ponderadas.

Figura 2 - Modelo matemático de um neurônio artificial



Fonte: (PEREIRA, 2017)

2.1.5 Gradient Boosting

Boosting é uma estratégia genérica para aprimorar o desempenho de qualquer algoritmo de aprendizado (FREUND, SCHAPIRE e ABE, 1999). Originalmente, o método foi proposto para tratar problemas de classificação de padrões, com a introdução do algoritmo *AdaBoost* por Freund e Schapire (1997). Posteriormente, surgiram diversas generalizações partir da estratégia original, entre elas, o algoritmo *Gradient Boosting* (FRIEDMAN, 2001). Conforme descreve (HASTIE, TIBSHIRANI e FRIEDMAN, 2009), o algoritmo pode ser aplicado tanto em problemas de classificação como em problemas de regressão. Como funciona para este, foi um dos escolhidos para os experimentos realizados na seção 2.2.4.

De forma sucinta, o que Friedman (2001) descreve é que o algoritmo *Gradient Boosting* consiste em um processo iterativo aditivo. O método inicia com uma previsão constante, cujo valor corresponde à média da variável de resposta na amostra de treinamento ($f_0(x) = \bar{y}$). A cada iteração, um novo termo é adicionado ao modelo corrente com o objetivo de reduzir gradualmente o erro de previsão. Assim, as atualizações são calculadas seguindo o sentido inverso do gradiente da função objetivo $\Psi(y_i, f(x_i))$, em relação às aproximações

correntes, f (x_i). O processo repete-se até que uma determinada condição de parada seja satisfeita, por exemplo, um número máximo de iterações, M.

A equação a seguir descreve a estratégia aditiva do algoritmo *Gradient Boosting*:

$$\hat{y}_i^{(0)} = f_0(\mathbf{x}_i) = \bar{y}$$
$$\hat{y}_i^{(1)} = f_0(\mathbf{x}_i) + \eta f_1(\mathbf{x}_i) = \hat{y}_i^{(0)} + \eta f_1(\mathbf{x}_i)$$
$$\hat{y}_i^{(2)} = f_0(\mathbf{x}_i) + \eta f_1(\mathbf{x}_i) + \eta f_2(\mathbf{x}_i) = \hat{y}_i^{(1)} + \eta f_2(\mathbf{x}_i)$$
$$\dots$$
$$\hat{y}_i^{(M)} = \sum_{m=0}^{M} f_m(\mathbf{x}) = \hat{y}_i^{(M-1)} + \eta f_M(\mathbf{x})$$

Onde $\hat{y}_i^{(m)}$ representa o valor estimado da variável de resposta para a i-ésima observação da amostra de treinamento, após a m-ésima iteração do algoritmo.

2.2 METODOLOGIA

2.2.1 Entendimento do negócio

Conforme descreve o art. 99 do Código Civil de 2002, são três os tipos de bens imóveis públicos: bens de uso comum do povo ou de domínio público (rios, ruas, praças, estradas etc.), bens de uso especial ou do patrimônio administrativo indisponível (destinados a serviços públicos, como repartições em geral, hospitais e escolas públicas) e bens dominiais ou do patrimônio disponível, que não possuem destinação pública determinada (por exemplo: prédios públicos desativados e terrenos de marinha).

Os imóveis dominiais são organizados e mantidos por sistema estruturante da Secretaria do Patrimônio da União (SPU). As finalidades e competências da SPU estão definidas no anexo I, art. 30 ao art. 33 do Decreto 8.818, de 21/7/2016, e no seu Regimento Interno, aprovado pela Portaria 220/2014 de 25/6/2014. Ademais, destacam-se a seguir duas das principais competências legais da SPU são:

a) alienar imóveis da União (art. 23 da Lei 9.636/1998 c/c o art. 1º, inciso I, do Decreto 3.125/1999);

b) organizar e manter sistema unificado de informações sobre os bens da União (art. 3-A da Lei 9.636/1998);

O sistema Siapa (Sistema Integrado de Administração Patrimonial), é quem cumpre, na SPU, a competência destacada no item b). O Siapa consiste em uma ferramenta de apoio à administração do patrimônio imobiliário da União voltado para imóveis dominiais e tem como objetivos:

a) identificar os imóveis dominiais da União, em que locais estão e quais são as suas características;

b) identificar os usuários dos imóveis dominiais da União, que imóveis estão ocupando, quais são os regimes de utilização e período de ocupação dos imóveis;

c) agilizar a cobrança e aprimoramento dos controles sobre os devedores omissos e fornecer dados para o encaminhamento dos processos para inscrição em dívida ativa da União e a competente execução judicial;

d) integrar os procedimentos da SPU e de suas Superintendências; e

e) dispor à SPU informações que possam apoiar os esforços de combate à sonegação e à moralização no trato da coisa pública.

Entre outras funcionalidades, o Siapa permite o gerenciamento da arrecadação de receitas patrimoniais devidas pelo uso dos imóveis dominiais da União e a padronização dos procedimentos operacionais das Superintendências do Patrimônio da União nas unidades da federação.

Auditoria realizada pelo TCU em 2016 (TC 011.609/2016-8) apontou diversas fragilidades na base de dados do sistema SIAPA. Algumas inconsistências são apresentadas a seguir (BRASIL., 2018):

> a) seis pessoas físicas e 97 pessoas indefinidas (pessoas não classificadas como pessoa física ou jurídica) com o nome formado por apenas um único termo;
> b) quatro pessoas físicas e 53 pessoas indefinidas com nomes que incluem dígitos numéricos;
> c) duas pessoas físicas e 83 pessoas indefinidas cujos nomes contêm termos como: “esposa”, “mulher”, “marido”, “mulher”, “conjuge”, “cônjuge”, “esposo” e “família”;
> d) 12.423 pessoas indefinidas sem informação de número de CPF ou de CNPJ – todos preenchido com o valor “000000000”;

> e) 145.719 pessoas físicas e 12.419 pessoas indefinidas (física ou jurídica) sem a data de nascimento preenchida;
> f) 1.122 pessoas físicas com século da data de nascimento diferente de 19XX, ou seja, ano de nascimento menor do que 1900 ou maior do que 1999, acusando idade maior do que 116 anos ou menor do que 16 anos.
>
> Dos números apresentados acima, destaca-se que quase metade das pessoas físicas responsáveis por imóveis ativos (145.719 de um total de 292.647) está sem a data de nascimento preenchida e apenas 4 pessoas indefinidas (de um total de 12.423) responsáveis por imóveis ativos estão com essa data preenchida.

Está em curso na SPU um Programa de Modernização da Gestão do Patrimônio da União (PGMPU). Segundo o relatório de auditoria (BRASIL., 2018), o programa abrange diversos projetos voltados para melhoria da gestão dos imóveis da União:

> Dentre estes projetos, pode-se mencionar o desenvolvimento do Sistema Unificado de Gestão do Patrimônio Público Federal, que atua como pilar principal de diversos projetos.
>
> Esse novo sistema é motivado principalmente pela necessidade de reconstrução dos atuais sistemas de gestão de imóveis existentes na SPU/MP, os quais demandam alto custo para manutenção, apresentam diversas limitações arquiteturais e cujos níveis de aderência aos processos de gestão e à legislação patrimonial precisam ser elevados e melhorados.
>
> Ainda de acordo com informações da SPU/MP, tais problemas foram responsáveis por algumas inconsistências existentes atualmente na base de dados do Siapa.

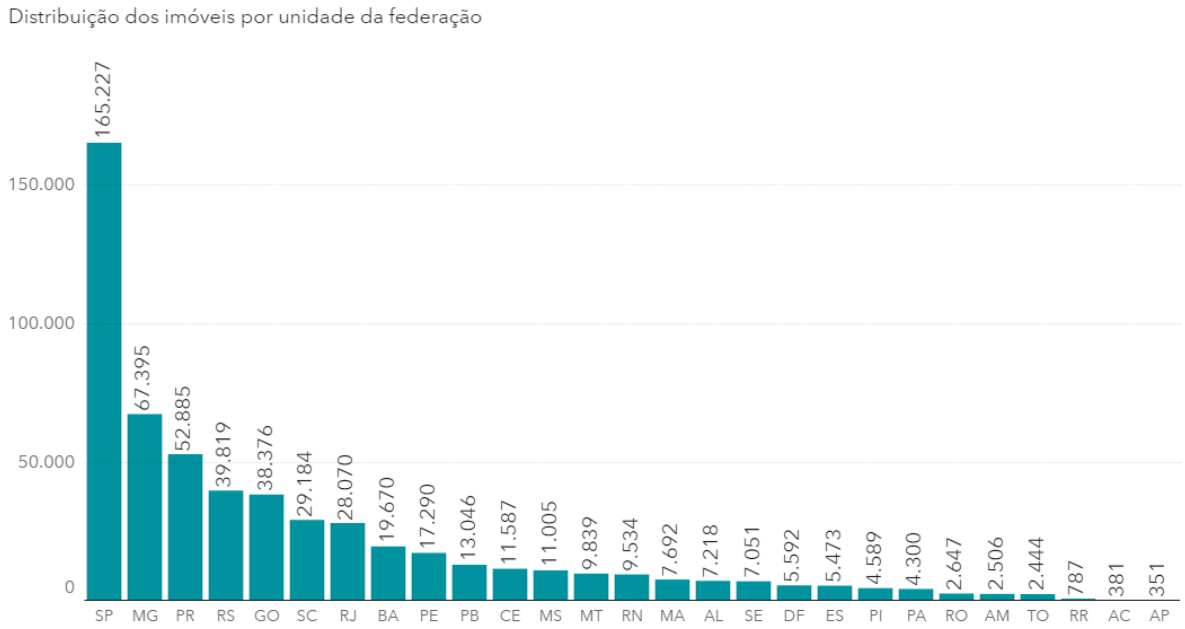
Conforme mencionado na seção 1.1.1, o objetivo principal do presente trabalho é o desenvolvimento de modelo preditivo para determinar o valor venal de imóveis da SPU, dispostos na base de dados do SIAPA. Para permitir a construção do modelo, foi solicitado à Caixa Econômica Federal (CEF), acesso à sua base de dados contendo imóveis residenciais financiados. Para cada imóvel da base foi realizado procedimento de avaliação por meio de especialistas, baseado em formulário com critérios objetivos. Após a avaliação, o valor de venda foi associado a cada imóvel. Na seção 2.2.2 exploraremos detalhes da base de dados fornecida pela CEF.

2.2.2 Entendimento dos dados

Nesta seção apresentaremos análise exploratória realizada para conhecer a base de dados cedida pela Caixa Econômica Federal contendo dados de imóveis residenciais financiados. Trata-se de base cujos dados são protegidos por sigilo pela CEF. Assim, quando apresentados, os dados serão anonimizados. Ademais, diversas colunas do arquivo original, recebido em formato .CSV, foram removidas da análise por representarem dados sensíveis em termos de sigilo e irrelevantes para as análises e construção dos modelos.

A base recebida contém 563.958 observações (linhas) de imóveis financiados em todo o país. A Figura 3 oferece o primeiro contato com a base. Trata-se da distribuição dos imóveis por unidade da federação. Para as UFs com maior quantidade de imóveis, são apresentados o número total de imóveis e o respectivo percentual.

Figura 3 - Distribuição dos imóveis por unidade da federação.



Fonte: Elaborada pelo autor (2020).

Duas outras *features* relevantes da base são o número de dormitórios e o padrão de acabamento do imóvel. A Tabela 1 apresenta a distribuição dos imóveis conforme seu padrão de acabamento. Cerca de 98% dos imóveis são de padrão “baixo” ou “normal”. Em etapas

posteriores do trabalho este conhecimento será importante, haja vista sua natureza desbalanceada.

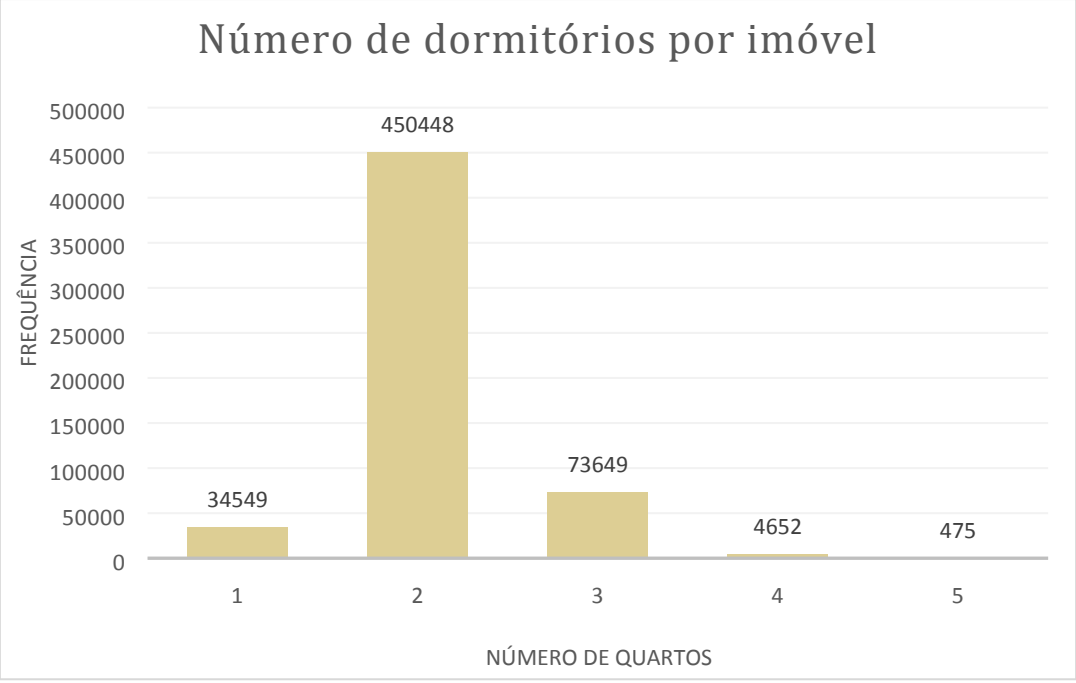
Tabela 1 - Quantidade de imóveis por padrão de acabamento.

Padrão de acabamento	Quantidade
Alto	7.703
Normal	439.066
Baixo	115.013
Mínimo	2.175

Fonte: Elaborada pelo autor (2020).

Os dados de distribuição por UF e padrão de acabamento têm boa qualidade na base recebida. Não há valores nulos nem valores fora do domínio definido. O mesmo não ocorre com a coluna que trata do número de quartos. Embora todos os valores estejam preenchidos, é clara a existência de *outliers* entre os dados. A base contém 4 casas com 0 quartos e outras 180 com número de quartos bastante superior a cinco. Assim, apresentamos a seguir, por intermédio da Figura 4, a distribuição dos imóveis que possuem de um a cinco dormitórios.

Figura 4 - Distribuição do número de dormitórios por imóvel.



Fonte: Elaborada pelo autor (2020).

A informação mais rica e relevante da base de dados é o valor de avaliação e venda dos imóveis. Conforme mencionado no início desta seção, trata-se de informações sensíveis e que não podem ser detalhadas ao longo do trabalho. Entretanto, em se tratando de uma análise exploratória, podemos conhecer a natureza do dado de forma agregada. Assim, a Tabela 2 demonstra a análise dos valores de avaliação e venda dos imóveis.

Tabela 2 - diferenças percentuais (em módulo) entre os valores de avaliação e venda dos imóveis

Descrição da variação	Percentual
Diferença entre valor de avaliação e venda menor que 5%	24,25%
Diferença entre valor de avaliação e venda menor que 10% e maior que 5%	10,45%
Diferença entre valor de avaliação e venda maior que 10%	11,62%
Valor da avaliação é igual ao valor de venda	53,68%

Fonte: Elaborada pelo autor (2020).

Além disso cabe destacar análise adicional. Ocorre que apenas 49.345 imóveis (8,75%) foram vendidos com preço superior ao da avaliação. Nos outros 91,25% (514.591 imóveis), o valor de venda foi igual ou inferior ao da avaliação.

Por fim, vale explorar dados acerca da área dos imóveis, seja a área total, seja a área de uso privado. O valor médio de área total dos imóveis da base é de 153,07 m^2, enquanto a área de uso privado tem valor médio de 111,27 m^2. Novamente, cabe apresentar a existência de alguns *outliers*: 166 imóveis foram detectados com área privada de 0.01 m^2 além de 3 imóveis com área total de 99999 m^2.

2.2.3 Preparação dos dados

As tarefas apresentadas nesta seção, bem como o trabalho de modelagem descrito na seção 2.2.4 é completamente agnóstico em relação às ferramentas e tecnologias adotadas. Optou-se pelo uso da ferramenta *SAS Enterprise Miner 14.3* em virtude da facilidade de uso, da capacidade de comparação entre modelos preditivos, dos recursos sofisticados para preparação de dados e, em especial, pela capacidade de realizar volumes substancialmente representativos de processamentos em paralelo. Nas seções subsequentes serão apresentados

possíveis trabalhos futuros e cabe adiantar que a realização de trabalhos semelhantes como, por exemplo, as linguagens de programação Python e R para comparação das abordagens de implementação seria de grande interesse.

Conforme apresentado na seção 2.2.2, os imóveis da base em análise estão distribuídos por todas as unidades da federação. Ademais, cada unidade da federação tem especificidades que fazem com que a previsão de valor de venda de imóveis seja tarefa complexa.

O Lago Norte, por exemplo, é um bairro nobre e de alto poder aquisitivo da capital do país. Uma via e cerca de 100 m o separam de outra região, o Varjão, que tem realidade complemente oposta. A diferença no valor de venda de imóveis nessa região é superior a 200%. Trata-se de exemplo simples, mas em um país com dimensões continentais, há dezenas de outras regiões com características semelhantes. Há regiões no Rio de Janeiro, por exemplo, que têm favelas e bairros nobres praticamente conectados. Diante deste cenário, o primeiro grande desafio deste trabalho foi: como criar um modelo que consiga levar em consideração os pormenores decorrentes da localização dos imóveis?

Consoante descrito no parágrafo anterior, o uso de localização (latitude + longitude) não seria adequado haja vista que imóveis em um raio inferior a 5 km podem ter valores de venda complemente distintos. Embora não tenha sido realizado nenhum experimento, acredita-se que a probabilidade de um modelo conseguir se adaptar a uma quantidade absolutamente grande de nuances entre regiões limítrofes é baixa e que o modelo teria pouca acurácia. Novamente, cabe adiantar que trabalho futuro poderia seguir esta linha de pesquisa, especialmente por meio de uso de redes neuronais.

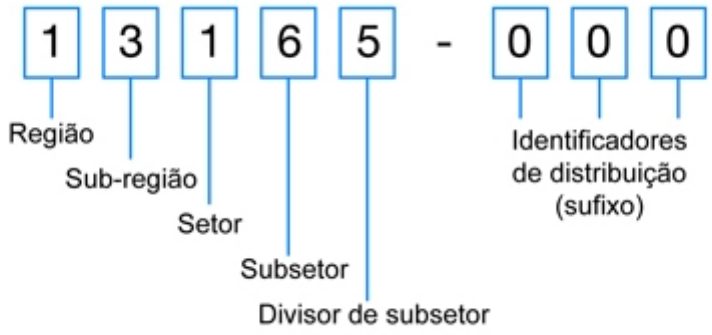
O uso do endereço dos imóveis foi aventado, mas rapidamente descartado. Apenas para o bairro de Águas Claras, em Brasília, foram detectadas mais de 10 variações na representação do nome do bairro (Águas Claras, Ag. Claras, AGUASCLARAS *etecetera)*. Em nível de avenidas e quadras há dezenas de outras mutações. Extrapolando a situação supradita para todas as unidades da federação, seria bastante complexo realizar saneamento de dados para cada bairro e região do país.

Uma hipótese de sucesso foi o uso do CEP (Código de Endereçamento Postal), informação presente na base de dados. Neste caso, para avançar na discussão acerca da preparação dos dados, faz-se necessário detalhar a estrutura do CEP. Este detalhamento é realizado na seção 2.2.3.1.

2.2.3.1 O Código de Endereçamento Postal (CEP)

O detalhamento apresentado a seguir foi obtido a partir do sítio dos Correios (Estrutura do CEP, 2020) e demonstra como o CEP está estruturado segundo o sistema decimal. Ele é composto de Região, Sub-região, Setor, Subsetor, Divisor de Subsetor e Identificadores de Distribuição. A Figura 5 ilustra a estruturação do CEP.

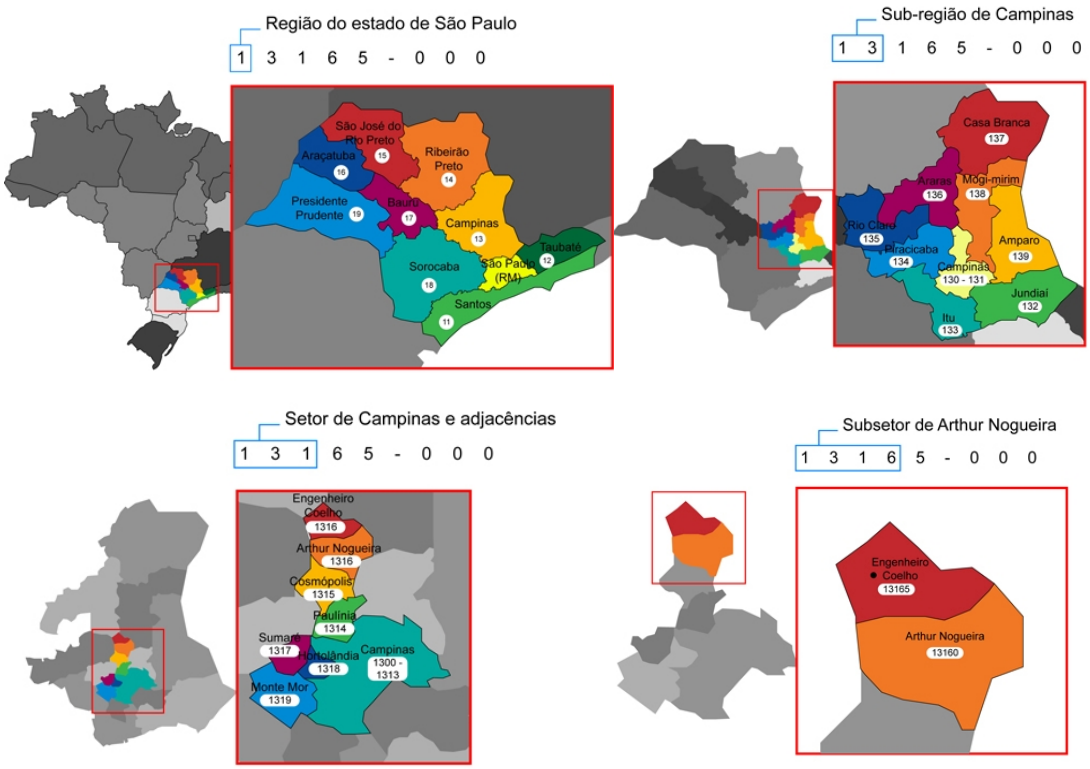
Figura 5 - Estrutura numérica do CEP



Fonte: Correios (2020)

Baseado no desenvolvimento sócio-econômico e fatores de crescimento demográfico, os Correios (Estrutura do CEP, 2020) dividiram o país em regiões postais (primeiro algarismo da estrutura numérica). Cada Região Postal foi dividida em 10 sub-regiões, que são indicadas pelo segundo algarismo do CEP. Cada Sub-Região foi dividida em 10 Setores, que são representados pelo terceiro algarismo. Cada Setor foi dividido em 10 subsetores, que são representados pelo quarto algarismo. Cada Subsetor foi dividido em 10 divisores de subsetor, que são representados pelo quinto algarismo. Os três algarismos após o hífen são denominados de SUFIXO e destinam-se à identificação individual de Localidades, Logradouros, Códigos Especiais e Unidades dos Correios. A Figura 6 ilustra, por meio de um caso concreto no estado de São Paulo, a divisão da estrutura numérica do CEP até o nível de subsetor.

Figura 6 - Exemplo de divisão em região, sub-região, setor e subsetor para uma localidade na cidade de Campinas - SP.



Fonte: Correios (2020)

A seção 2.2.3.2 descreve a importância do uso do CEP no processo de preparação dos dados.

2.2.3.2 Sobre o uso do CEP como elemento básico para preparação dos dados

Tendo como base a estrutura do CEP apresentada na seção 2.2.3.1, a Tabela 3 ilustra a quantidade máxima de divisões possíveis para cada elemento de estruturação do CEP. Ademais, ao agrupar os dados da base descrita na seção 2.2.2, obtivemos o número concreto de CEPs nos quais há imóveis da União. Este agrupamento também foi incluído na Tabela 3. Para exemplificar, seja o agrupamento "Setor". Conforme indica a tabela, há 1.000 diferentes possibilidades de CEP para este agrupador e a base de dados usada neste trabalho apresenta 869 em uso.

Tabela 3 – Quantidade máxima de CEPs por divisão e número de CEPs em uso na base de dados fornecida pela Caixa Econômica Federal.

Descrição	Quantidade máxima de divisões	Número de divisões em uso
Região	10	10
Sub-região	100	100
Setor	1.000	869
Subsetor	10.000	5.344
Divisor de subsetor	100.000	13.563

Fonte: Elaborada pelo autor (2020).

Conhecendo a organização do CEP e a quantidade de total de CEPs diferentes utilizados na base da Caixa, a próxima pergunta a ser respondida era: como estão distribuídos os imóveis da base de dados por estes CEPs? Em outros termos, a questão seria: quantos setores, subsetores ou divisores de subsetores temos na base com, pelo menos, 10 imóveis? Este conhecimento é fundamental para viabilizar a modelagem dos dados realizada na seção 2.2.4 haja vista a necessidade de uma amostra representativa de imóveis em cada divisão.

Diante deste cenário, foi desenvolvido programa em linguagem Python para responder à questão supradita. O programa, cujo código fonte encontra-se no APÊNDICE A, apresenta para as divisões de setor, subsetor e divisor de subsetor, a quantidade mínima e o percentual de CEPs diferentes para valores de 1 a 100, ou seja, para um valor *i* de 1 a 100, o programa responde quantos CEPs diferentes possuem, ao menos, *i* imóveis em dada região e o percentual correspondente na base de dados. A Tabela 4 ilustra os 20 primeiros valores gerados pelo programa.

Com o produto gerado pela Tabela 4, poder-se-ia selecionar um "ponto de corte" com o número mínimo de imóveis por divisão para que um divisor participasse do modelo a ser definido na etapa 2.2.4. Sem embargo, antes de avançar, resta compreender um comportamento fundamental: é viável construir um único modelo para toda a base de dados?

Tabela 4 – Primeiros 20 valores gerados pelo programa Python que analisou a quantidade de CEPs distintos e o respectivo percentual para cada quantidade de imóveis, de 1 a 100.

Quantidade de imóveis	CEPs diferentes por setor	% de CEPs por setor	CEPs diferentes por subsetor	% de CEPs por subsetor	CEPs diferentes por divisor subsetor	% de CEPs por divisor de subsetor
1	869	100,00	5344	100,00	13563	100,00
2	846	97,24	4849	90,72	11317	83,43
3	833	95,75	4585	85,78	10034	73,98
4	828	95,17	4338	81,16	9118	67,22
5	823	94,60	4188	78,35	8443	62,25
6	820	94,25	4019	75,19	7927	58,44
7	815	93,68	3868	72,37	7507	55,35
8	813	93,45	3744	70,05	7126	52,54
9	810	93,10	3623	67,78	6783	50,01
10	806	92,64	3540	66,23	6501	47,93
11	803	92,30	3426	64,10	6243	46,03
12	798	91,72	3349	62,66	5997	44,21
13	794	91,26	3258	60,95	5781	42,62
14	791	90,92	3192	59,72	5600	41,29
15	788	90,57	3132	58,60	5399	39,80
16	785	90,23	3066	57,36	5201	38,34
17	783	90,00	3000	56,13	5040	37,16
18	779	89,54	2937	54,95	4874	35,93
19	776	89,20	2874	53,77	4726	34,84
20	775	89,08	2804	52,46	4584	33,80

Fonte: Elaborada pelo autor (2020).

Para responder à pergunta, foram executadas 5 tarefas, por meio da ferramenta *SAS Enterprise Miner*:

1. carga da base de dados;
2. remoção de 20 colunas irrelevantes para análise;
3. divisão da base em 70% para treinamento e 30% para teste;
4. criação da coluna "CEP_SUBSETOR" (produto da expressão "INT(ED_CEP_IMOVEL / 10000)"); e
5. criação de uma coluna nova coluna para cada subsetor (*one hot encoding)* .

A Figura 7 ilustra a organização das tarefas na ferramenta SAS Enterprise Miner.

Figura 7 – Fluxo de tarefas para preparação dos dados de geração de subsetor de CEP para base da CEF

Fonte: Elaborada pelo autor (2020).

Conforme apresentado na seção 2.2.2, a base recebida contém 563.958 observações (linhas). Como a operação de *one hot encoding* gera uma nova coluna (com zero ou um) para cada CEP diferente e sabendo que há 5.344 diferentes subsetores na base, teríamos, pelo menos, uma matriz da ordem de 3 bilhões de elementos (563.958 x 5.344 = 3.013.791.552).

A despeito do volume de dados expressivo, o *SAS Enterprise Miner* executou a criação da referida matriz em menos de 4 minutos.

Diante deste resultado, temos os dados preparados para iniciar a etapa mais relevante do trabalho: a modelagem dos dados. Esta etapa será detalhada na seção 2.2.4.

2.2.4 Modelagem

Inicialmente, foram criados três modelos-base: um baseado no setor do CEP, outro no subsetor e finalmente um terceiro no menor nível granular, o divisor de subsetor. Note que cada modelo deve aprender e responder individualmente pelos dados de todo o país. Dadas as especificidades de cada região – detalhadas na seção 2.2.2 – há enorme desafio para obter resultados aceitáveis e proveitosos em uma aplicação de produção.

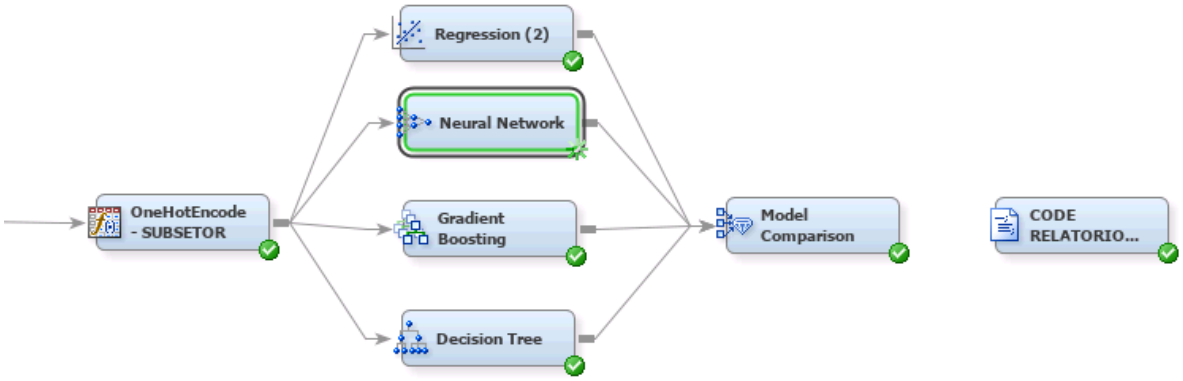
Para cada modelo-base, usando o *SAS Enterprise Miner,* foram aplicados 4 modelos preditivos (com seus hiperparâmetros *default)* a saber:

- regressão,
- árvore de decisão,
- *gradient boosting, e*
- rede neuronal.

Após a aplicação dos modelos preditivos, foi utilizado o recurso de *ensemble* do *SAS Enterprise Miner* com o objetivo de otimizar o modelo-base e aproveitar o comportamento mais

efetivo de cada modelo preditivo. Ato contínuo, foi utilizado o recurso de comparação de modelos do *SAS Enterprise Miner* para aferir o que obteve melhor resultado, conforme a métrica de erro quadrático médio. Ademais, foram coletados os tempos de execução e foi mensurado o volume de dados mantido em memória ao longo da modelagem. A Figura 8 ilustra o modelo-base criado para subsetor.

Figura 8 - Modelo-base construído com divisão do CEP em subsetores



Fonte: Elaborada pelo autor (2020).

Para avaliar os modelos-base, foi criada uma métrica de fácil entendimento e comunicação. Ao término da execução do modelo-base, para cada valor de imóvel previsto, calculou-se a diferença entre o valor real de venda e o valor previsto pelo modelo. Ato contínuo, verificou-se o percentual de casos com erro inferior a 10%, a 15% e a 20%. Os resultados para os três modelos-base são apresentados na Tabela 5.

Tabela 5 - Resultados dos modelos-base com Setor, Subsetor e Divisor de subsetor do CEP

Descrição	**Tempo de execução (horas)**	**Tamanho da matriz (bilhões de elementos)**	**Erro inferior a 10%**	**Erro inferior a 15%**	**Erro inferior a 20%**
Setor	1	0,5	47,17	63,00	73,98
Subsetor	54	3	36.02	50,83	63,27
Divisor de subsetor	168+	7,2	-	-	-

Fonte: Elaborada pelo autor (2020).

Embora pareça simples, há bastantes elementos que se depreendem da Tabela 5 e que requerem discussão adicional:

a) Mesmo no caso de maior acurácia (setor), os resultados são absolutamente frágeis e de pouca aplicabilidade em um modelo de produção. Na seção 2.3 serão discutidos os motivos que fizeram com que os resultados fossem tão limitados.
b) Não foi possível obter resultados para o divisor de subsetor. Após 1 semana completa de processamento, a tarefa foi interrompida. A matriz sob a qual os algoritmos de predição trabalhavam possuía cerca de 7,2 bilhões de elementos e não havia como prever o tempo para conclusão do trabalho de processamento.
c) Os resultados para Setor e Subsetor foram bastante próximos. Isso mostra que os algoritmos não conseguiram aproveitar o fato de cada CEP ter sido separado em sua própria coluna indicativa.

A despeito do fracasso na construção de modelo único capaz de prever valores para a base de dados com imóveis de todas as regiões do país, avaliou-se uma segunda estratagema: em vez de se criar um único modelo capaz de aprender o comportamento de 13.563 regiões, o plano foi criar 13.563 submodelos diferentes, um para cada região (divisor de subsetor), e gerar uma comunidade (associação) capaz de detectar qual submodelo mais apropriado para cada situação de produção.

O método supramencionado foi aplicado a um subgrupo de imóveis do país. Inicialmente, foram selecionados apenas os imóveis do Distrito Federal para análise. Ademais, foram implementadas duas novas transformações:

a) o atributo “padrão de acabamento” dos imóveis estava com faixa de valores que variava de forma decrescente. Foi aplicada regra (PADRAO_ACABAMENTO * (-1) + 5) para permitir que os dados fossem corretamente interpretados pelos algoritmos de predição. A Tabela 6 apresenta os valores originais e os valores após a transformação.
b) o atributo “Tipo imóvel” apresentava dois valores válidos: 1 para “Casa” e 2 para “Apartamento”. Neste caso, foi aplicado o *one hot encoding* e criadas colunas específicas para cada um dos casos.

Tabela 6 - Transformação da coluna padrão de acabamento dos imóveis

Descrição	Valores da base original	Valores transformados
Alto	1	4
Normal	2	3
Baixo	3	2
Mínimo	4	1

Fonte: Elaborada pelo autor (2020).

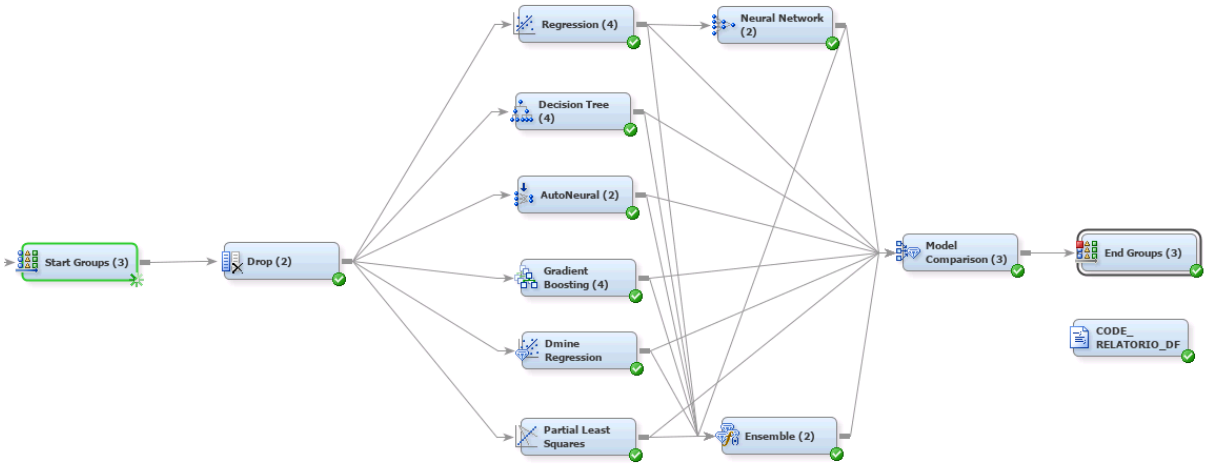
A Figura 9 ilustra a filtragem dos dados do DF na ferramenta *SAS Enterprise Miner*,

Figura 9 - Preparação dos dados para modelagem dos imóveis do DF

Fonte: Elaborada pelo autor (2020).

A separação por divisor de subsetor gerou 78 CEPs diferentes (com ao menos 10 imóveis) para análise no Distrito Federal. Como serão gerados 78 diferentes modelos, decidiu-se aproveitar o potencial da ferramenta *SAS Enterprise Miner* e, para cada CEP, verificar entre 8 diferentes algoritmos, aquele que obteria melhor desempenho em termos da métrica estabelecida. Foram utilizados: regressão linear com múltiplas variáveis, árvore de decisão, rede neuronal, *autoneural* – rede neuronal autoconfigurada –, *gradient boosting*, *partial least squares*, *Dmine Regression* e um *emsemble* de todas as opções. A Figura 10 ilustra trecho do diagrama que emprega o uso desses modelos.

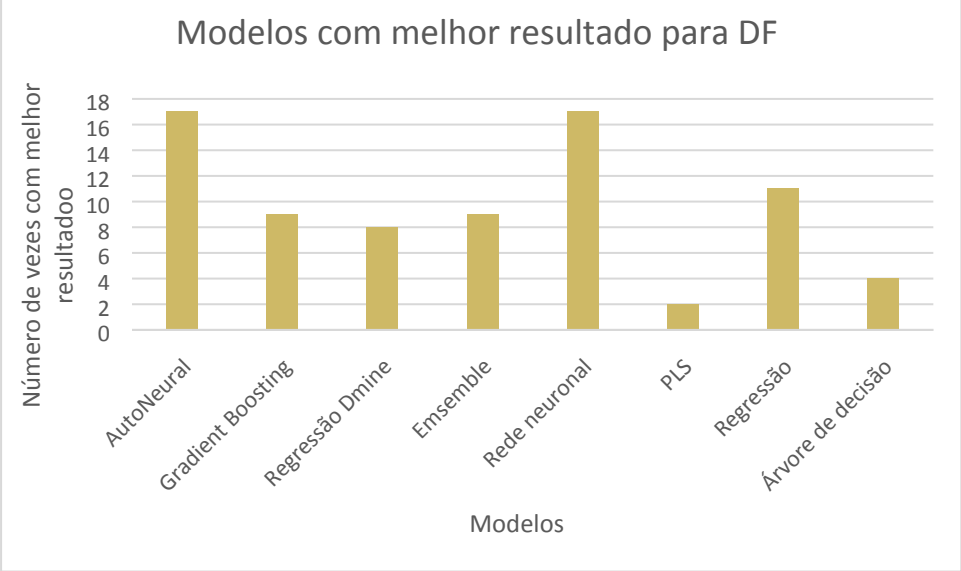
Figura 10 - Diagrama da ferramenta SAS Enterprise Miner com os algoritmos empregados para treinamento do modelo.



Fonte: Elaborada pelo autor (2020).

Em cerca de 2h, foram computados os 78 modelos para base de dados referente ao Distrito Federal. A Figura 11 apresenta gráfico com o número de vezes que cada um dos 8 (oito) modelos se sobressaiu.

Figura 11 – Número de vezes que cada modelo se sobressaiu no DF (78 CEPs com divisor de subsetor).



Fonte: Elaborada pelo autor (2020).

Acerca dos resultados obtidos, cabe destacar que:

a) O critério utilizado para escolher o modelo de melhor desempenho foi o menor erro quadrático médio.
b) É patente a superioridade dos modelos baseados em redes neuronais. Em 44% dos CEPs, elas obtiveram melhor resultado.
c) Todos os modelos testados se sobressaíram em, pelo menos, duas ocasiões. Este resultado deixa claro que não há modelo único que atenda a todas as regiões. As especificidades de cada CEP fazem com que, para cada região, um algoritmo tenha melhor resultado.

Os resultados obtidos na execução estão listados na Tabela 7 e são muito superiores aos colhidos com a estratégia anterior.

Tabela 7 - Resultados dos modelos-base com Divisor de subsetor do CEP para imóveis do Distrito Federal

Descrição	**Tempo de execução (horas)**	**Erro inferior a 10%**	**Erro inferior a 15%**	**Erro inferior a 20%**
Divisor de subsetor	2h	80,49	89,30	93,91

Fonte: Elaborada pelo autor (2020).

Na seção a seguir, os resultados acima serão comparados a *benchmarks* de mercado e o modelo gerado será aplicado em um conjunto de dados de teste para avaliação.

2.3 RESULTADOS

O Zestimate (https://www.zillow.com/zestimate/), da *Zillow Group* (https://www.zillow.com/), é um modelo de avaliação de valor venal de imóveis similar ao gerado na seção 2.2.4. Trata-se de *benchmark* reconhecido mundialmente pela acurácia e maturidade. Desde 2006, a *Zillow* investe milhões de dólares no aprimoramento de seus modelos e no uso de novos atributos.

A organização disponibiliza em seu sítio na internet tabelas contendo a acurácia do modelo por região dos Estados Unidos. A Tabela 8 apresenta exemplos de regiões dos EUA e o percentual de erro nas estimativas.

Tabela 8 - Exemplos de regiões dos EUA e o respectivo percentual de erro nas estimativas realizadas pelo grupo *Zillow*.

Região	Erro inferior a 5%	Erro inferior a 10%	Erro inferior a 20%
San Francisco, CA	62,7	86,1	97,6
Seattle, WA	78,0	94,4	99,4
Detroit, MI	78,9	94,3	98,4
Boston, MA	77,8	94,7	99,1

Fonte: (GROUP, 2020)

Fazendo uso de apenas quatro atributos (padrão de acabamento, número de dormitórios, área total privativa, área total) o percentual de erro inferior a 10% no experimento realizado para o DF foi de 79,73% (vide Tabela 7). Trata-se de resultado bastante expressivo, especialmente por ainda haver diversas oportunidades de melhoria dos modelos.

No tocante aos resultados com dados de teste, por meio do portal de Alienação de Imóveis da União (ECONOMIA, 2020), foi obtido edital com dados de imóveis residenciais pela SPU/MP. Em especial, foi utilizada a Concorrência Pública SPU/MP 01/2017 haja vista conter diversos imóveis no Distrito Federal, local usado como base na modelagem implementada na seção 2.2.4.

O edital ofertou 25 (vinte e cinco) imóveis, mas apenas 11 (onze) deles estavam em regiões (pelo divisor de subsetor do CEP) nas quais também havia ao menos um imóvel das bases de treino e validação mencionada na seção 2.2.2. A Tabela 9 apresenta esses imóveis com seus respetivos atributos, obtidos por meio de consulta à base do SIAPA no Labcontas e por meio do anexo do próprio edital.

Os 11 (onze) imóveis listados na Tabela 9 estão divididos em três diferentes divisores de subsetor 70733, 70736 e 70747. Para o primeiro deles havia 2 imóveis na base de treino e para o segundo e o terceiro, apenas um imóvel. Os resultados apresentados na Tabela 7, conforme mencionado anteriormente, foram baseados em casos com pelo menos 10 imóveis para treino do modelo. Logo, não faz sentido utilizar o modelo construído na seção 2.2.4.

Tabela 9 - Dados de imóveis negociados na Concorrência Pública SPU/MP 01/2017

Endereço	CEP (divisor de subsetor)	Área construída	Dormitórios	Padrão acabamento
SQN 104, BL. I, AP.607	70733	125,3	3	Alto
SQN 104, BL. C, AP.607	70733	111,72	3	Alto
SQN 104, BL. F, AP.302	70733	111,72	3	Alto
SQN 104, BL. F, AP.502	70733	111,72	3	Alto
SQN 304, BL. D, AP.415	70736	134,46	3	Alto
SQN 304, BL. D, AP.606	70736	131,93	3	Alto
SQN 304, BL. B, AP.607	70736	134,41	3	Alto
SQN 308, BL.A, AP.604	70747	122,48	3	Alto
SQN 308, BL.B, AP.306	70747	122,48	3	Alto
SQN 108, BL.E, AP.208	70747	123,93	3	Alto
SQN 108, BL.H, AP.501	70747	123,93	3	Alto

Fonte: (ECONOMIA, 2020) – com adaptações

No caso concreto, embora a base cedida pela Caixa Econômica Federal (vide seção 2.2.2) seja bastante rica e contenha mais de meio milhão de imóveis, a localização destes bens não tem grande intersecção com os bens disponíveis para venda pela Secretaria de Patrimônio da União. Esta observação é um grande limitador de toda a proposta deste trabalho.

Em virtude do problema supramencionado, o modelo produzido na seção 2.2.4 foi novamente gerado, desta vez usando o setor como elemento agrupador em detrimento ao divisor de subsetor. Na prática, em vez de usarmos três diferentes modelos para os CEPs 70733, 70736 e 70747, foi criado um único modelo baseado no setor 707, contendo 76 (setenta e seis) imóveis no bairro Asa Norte e, portanto, passível de ser utilizado com os 11 (onze) imóveis detalhados na Tabela 9.

O modelo gerado, conforme previsto, obteve resultados muito inferiores aos produzidos com o separador “divisor de subsetor”. Os valores estão dispostos na Tabela 10.

Tabela 10 - Resultados dos modelos-base com Setor do CEP para imóveis do Distrito Federal

Divisor do CEP	Erro inferior a 10%	Erro inferior a 15%	Erro inferior a 20%
Setor	65,42	77,34	83,10

Fonte: Elaborada pelo autor (2020).

A Tabela 11 reapresenta os 11 (onze) imóveis negociados na Concorrência Pública SPU/MP 01/2017, com o valor mínimo disposto no edital, o valor obtido por meio do modelo preditivo e a diferença percentual entre os dois. Em todos os casos, o erro foi inferior a 15% e em mais da metade dos casos, o erro foi inferior a 10%. Trata-se de resultado absolutamente promissor, especialmente no caso de utilizarmos o setor em vez do divisor de subsetor para treinar o modelo preditivo.

Tabela 11 - Preço mínimo dos imóveis vendidos na Concorrência Pública SPU/MP 01/2017, valores previstos pelo modelo e a diferença percentual entre os valores.

Identificador do imóvel no edital	Endereço	Preço mínimo (edital) em R$	Valor predito pelo modelo em R$	Diferença (percentual)
5	SQN 104, BL. I, AP.607	950.000	1.043.126,04	9,80
7	SQN 104, BL. C, AP.607	880.000	939.806,35	6,79
8	SQN 104, BL. F, AP.302	836.000	939.806,35	12,41
9	SQN 104, BL. F, AP.502	836.000	939.806,35	12,41
11	SQN 304, BL. D, AP.415	1.011.000	1.122.602,72	11,03
15	SQN 304, BL. D, AP.606	989.000	1.098.759,72	11,09
23	SQN 304, BL. B, AP.607	1.119.000	1.122.602,72	0,32
18	SQN 308, BL.A, AP.604	945.000	1.027.230,70	8,70
19	SQN 308, BL.B, AP.306	928.000	1.027.230,70	10,69
20	SQN 108, BL.E, AP.208	1.002.000	1.035.178,37	3,31
21	SQN 108, BL.H, AP.501	986.000	1.035.178,37	4,98

Fonte: Elaborada pelo autor (2020).

3 CONCLUSÃO

Ao longo deste trabalho foram desenvolvidos modelos preditivos para determinar o valor venal de imóveis da Secretaria de Patrimônio da União. A estratégia de produzir diversos modelos, regionalizados e categorizados por meio do CEP, produziu resultados superiores à estratégia com um modelo único para todo o país.

Não houve um único algoritmo de aprendizado de máquina que se sobressaísse em todas as setorizações por CEP implementadas. Embora as redes neuronais tenham obtido bons resultados na maioria dos casos, houve diversas situações nas quais as árvores de decisão, as regressões, o *emsemble* e o *gradient boosting* apresentaram resultados superiores. Assim, o modelo implementado neste trabalho utiliza o algoritmo com melhor resultado na região na qual se deseja realizar uma previsão de valor de venda.

A aplicação do modelo gerado a imóveis da Concorrência Pública SPU/MP 01/2017 produziu resultados bastante promissores (100% das avaliações com erro inferior a 15%), mas que devem ser analisados com comedimento. Há poucos editais publicados para análise por parte da SPU e seria necessária uma massa de testes mais substancial para avaliar o uso do modelo em ambiente de produção.

Embora a base de dados cedida pela Caixa Econômica Federal seja bastante rica e contenha mais de meio milhão de imóveis, há casos nos quais a localização destes bens não tem intersecção com os bens disponíveis para venda pela Secretaria de Patrimônio da União, o que obriga o uso de modelos de menor acurácia, baseados no subsetor e no setor do CEP de uma região. Essa particularidade é o grande limitador de toda a proposta deste trabalho.

Há ainda muito por fazer. Trabalhos futuros podem experimentar o uso de redes neuronais de forma mais extensiva. A ferramenta *SAS Enterprise Miner* facilita a criação dos modelos preditivos, sem embargo, é possível que ajustes em hiperparâmetros de redes neuronais gere resultados ainda mais expressivos, especialmente se fosse realizado um treinamento prévio. Na mesma linha de trabalho, reimplementar todo o código em linguagem *Python* permitiria comparar os resultados obtidos por meio da ferramenta gráfica e, quiçá, aprimorar os resultados. Um estudo pormenorizado dos motivos pelos quais cada modelo se sobressaiu ou fracassou em cada CEP estudado seria valioso para novas implementações. Outra oportunidade de estudo seria o uso da informação de setor censitário, disponibilizada pelo IBGE. Nesse caso, o conceito de "área de ponderação" poderia ser de grande valia para determinar regiões com similaridades em termos, por

exemplo, da renda média da população. Por fim, seria igualmente salutar uma análise indicativa do nível de confiança de cada previsão, baseada na comparação de métricas – como o erro quadrático médio – produzidas para cada região modelada.

REFERÊNCIAS

BRASIL. Tribunal de Contas da União. **Acórdão de relação nº 244/2018 - Plenário:** Relator - Ministro Aroldo Cedraz. Sessão de 07/02/2018., 2018. Disponivel em: <https://contas.tcu.gov.br/sagas/SvlVisualizarRelVotoAcRtf?codFiltro=SAGAS-SESSAO-ENCERRADA&seOcultaPagina=S&item0=584942>. Acesso em: 20 setembro 2019.

BREIMAN, L. et al. **Classification and Regression Trees**. California: CRC, 1984. 368 p.

ECONOMIA, M. D. Alienação de imóveis da União. **Alienação de imóveis da União**, 2020. Disponivel em: <http://www.planejamento.gov.br/acesso-a-informacao/licitacoes-e-contratos/licitacoes/alienacao-de-imoveis-da-uniao/>. Acesso em: 2 Março 2020.

ESTRUTURA do CEP. **Correios**, 2020. Disponivel em: <https://www.correios.com.br/enviar-e-receber/ferramentas/cep/estrutura-do-cep>. Acesso em: 11 Janeiro 2020.

FACELI, K. E. A. **Inteligência Artificial:** Uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: Livros Técnicos e Científicos Editora Ltda., 2015.

FREUND, Y.; SCHAPIRE, R. A decision-theoretic generalization of online learning and an application to boosting. **Journal of computer and system sciences**, v. 1, n. 55, p. 119-139, 1997.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. **Japanese Society For Artificial Intelligence** , n. 14, p. 771-780, 1999.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, 2001. 1189-1232.

GROUP, Z. Zestimate. **Zillow - Reimagine Home**, 2020. Disponivel em: <https://www.zillow.com/zestimate/>. Acesso em: 2 Março 2020.

HAIKIN, S. **Redes Neurais:** princípios e prática. 2. ed. Porto Alegre: Bookman, 2001. 900 p.

HAIR, J. et al. **Análise Multivariada de Dados**. Porto Alegre: Bookman, 2009.

HAN, ; KAMBER, ; PEI,. **Data mining:** concepts and techniques. [S.l.]: Elsevier, 2011.

HASTIE, T. J.; TIBSHIRANI, R. J.; FRIEDMAN, J. H. **The elementos of statistical learning:** data mining, inference and prediction. Springer Series in Statistics. New York: Springer, 2009.

KASS, G. An exploratory technique for investigating large quantities of categorical data. **Applied Statistics,** 29, n. 2, 1980. 119-127.

KONAR, A. **Artificial intelligence and soft computing:** behavioral and cognitive modeling of the human brain. [S.l.]: CRC Press, 1999.

MAIMON, O. Z.; ROKACH, . **Data Mining and Knowledge Discovery Handbook**. 2. ed. [S.l.]: Springer, 2005.

MCCULLOCH, W. S.; PITTS, W. A Logical Calculus of the Ideas Imminent in Nervous Activity. **Bulletin of Mathematical Biophysics**, v. 17, p. 115-133, 1943.

MULLER, M.; FILL, H. D. **Redes Neurais aplicadas**. Curitiba: [s.n.]. 2003.

PEREIRA, T. **Uso de Inteligência Artificial para estimativa da capacidade de suporte de carga do solo**. Santa Maria: [s.n.], 2017. 179 p.

QUINLAN, J. R. Introduction of decision trees. **Machine Learning**, 1, 1986. 81-106.

QUINLAN, J. R. **C4.5:** Programs for machine learning. San Mateo: Morgan Kaufmann Publishers, 1993. 302 p.

RAGSDALE, C. T. **Spreadsheet Modeling and Decision Analysis**. 3. ed. Cincinnati: South-Western College Publishing, 2001.

STERNBERG, J. C.; STILLO, H. S.; SCHWENDEMAN, R. H. Spectrophotometric Analysis of Multicomponent Systems Using Least Squares Method in Matrix Form. **Analytical Chemestry**, n. 32, 1960. 84-90.

TACONELI, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia. Tese (Doutorado)**. Piracicaba: USP, 2008.

APÊNDICE A – Código-fonte do caderno Jupyter de análise de CEPs por setor, subsetor e divisor de subsetor

```
import pandas as pd

import numpy as np

from pandas import ExcelWriter

from pandas import ExcelFile

pd.set_option('display.max_columns', 500)


df = pd.read_csv('Imovel_Demanda_TCU.csv', delimiter=';', decimal=",", encoding='UTF-8', low_memory=False)

df.head(10)

df = pd.read_excel('IMOVEIS_CAIXA.xlsx')

df.head(10)

# Adicionar colunas com partes do CEP

df['CEP_SETOR'] = (df['ED_CEP_IMOVEL'].replace({' ': '-72110210'}).astype(int))//100000

df['CEP_SUBSETOR'] = (df['ED_CEP_IMOVEL'].replace({' ': '-72110210'}).astype(int))//10000

df['CEP_DIVISOR_SUBSETOR'] = (df['ED_CEP_IMOVEL'].replace({' ': '-72110210'}).astype(int))//1000

df['CEP_SUBREGIAO'] = (df['ED_CEP_IMOVEL'].replace({' ': '-72110210'}).astype(int))//1000000

df['CEP_REGIAO'] = (df['ED_CEP_IMOVEL'].replace({' ': '-72110210'}).astype(int))//10000000

df['CEP_SETOR'] = df['CEP_SETOR'].astype('category')

df['CEP_SUBSETOR'] = df['CEP_SUBSETOR'].astype('category')

df['CEP_DIVISOR_SUBSETOR'] = df['CEP_DIVISOR_SUBSETOR'].astype('category')

df['CEP_SUBREGIAO'] = df['CEP_SUBREGIAO'].astype('category')

df['CEP_REGIAO'] = df['CEP_REGIAO'].astype('category')

df = df.drop(['ED_CEP_IMOVEL'], axis=1)
```

```
# Manter apenas as colunas mais relevantes

df =
df.drop(['NU_REGISTRO_GARANTIA','DT_REGISTRO_GARANTIA','NU_GRAU_GAR
ANTIA_CONSITUIDA','ED_LOGRADOURO_IMOVEL','ED_NUMERO_IMOVEL','ED_
COMPLEMENTO_IMOVEL','ED_BAIRRO_IMOVEL','NU_MUNICIPIO_IMOVEL_IBGE
'], axis=1)

df =
df.drop(['NU_TIPO_IMPLANTACAO','QT_VAGA_GARAGEM_PRIVATIVA','NU_AREA
_TESTADA','NU_AREA_TERRENO','NU_ESTADO_CONVERVACAO_CONDOMINIO','
NU_ESTADO_CONSERVACAO_IMOVEL','DT_AVALIACAO_IMOVEL','VR_AVALIA
CAO_IMOVEL','DT_COMPRA_VENDA','VR_COMPRA_VENDA','DT_MOVIMENTO'],
axis=1)

df = df.drop(['NU_CARTORIO','NU_MATRICULA'], axis=1)

df_totais_setor = pd.DataFrame(columns=['TOTAL_CEPS_SETOR'])

df_cep_setor = pd.DataFrame({'count' :
df.groupby(['CEP_SETOR'])['CEP_SETOR'].count()}).reset_index()

for i in range(1,101) :

    df_totais_setor = df_totais_setor.append({'TOTAL_CEPS_SETOR' :
df_cep_setor[df_cep_setor['count'] >= i].count()['count']} , ignore_index=True)

df_totais_setor['TOTAL_PERCENTUAL_CEPS_SETOR'] =
(df_totais_setor['TOTAL_CEPS_SETOR']/df_totais_setor['TOTAL_CEPS_SETOR'][0])*100

df_ceps = df_totais_setor

df_totais_subsetor = pd.DataFrame(columns=['TOTAL_CEPS_SUBSETOR'])

df_cep_subsetor = pd.DataFrame({'count' :
df.groupby(['CEP_SUBSETOR'])['CEP_SUBSETOR'].count()}).reset_index()

for i in range(1,101) :

    df_totais_subsetor = df_totais_subsetor.append({'TOTAL_CEPS_SUBSETOR' :
df_cep_subsetor[df_cep_subsetor['count'] >= i].count()['count']} , ignore_index=True)

df_totais_subsetor['TOTAL_PERCENTUAL_CEPS_SUBSETOR'] =
(df_totais_subsetor['TOTAL_CEPS_SUBSETOR']/df_totais_subsetor['TOTAL_CEPS_SUBS
ETOR'][0])*100

df_ceps = pd.concat([df_ceps, df_totais_subsetor], axis = 1)

df_totais_divisor_subsetor =
pd.DataFrame(columns=['TOTAL_CEPS_DIVISOR_SUBSETOR'])
```

```
df_cep_divisor_subsetor = pd.DataFrame({'count' :
df.groupby(['CEP_DIVISOR_SUBSETOR'])['CEP_DIVISOR_SUBSETOR'].count()}).reset_
index()

for i in range(1,101) :

    df_totais_divisor_subsetor =
df_totais_divisor_subsetor.append({'TOTAL_CEPS_DIVISOR_SUBSETOR' :
df_cep_divisor_subsetor[df_cep_divisor_subsetor['count'] >= i].count()['count']} ,
ignore_index=True)

df_totais_divisor_subsetor['TOTAL_PERCENTUAL_CEPS_DIVISOR_SUBSETOR'] =
(df_totais_divisor_subsetor['TOTAL_CEPS_DIVISOR_SUBSETOR']/df_totais_divisor_subs
etor['TOTAL_CEPS_DIVISOR_SUBSETOR'][0])*100

df_ceps = pd.concat([df_ceps,df_totais_divisor_subsetor], axis = 1)

df_totais_regiao = pd.DataFrame(columns=['TOTAL_CEPS_REGIAO'])

df_cep_regiao = pd.DataFrame({'count' :
df.groupby(['CEP_REGIAO'])['CEP_REGIAO'].count()}).reset_index()

for i in range(1,101) :

    df_totais_regiao = df_totais_regiao.append({'TOTAL_CEPS_REGIAO' :
df_cep_regiao[df_cep_regiao['count'] >= i].count()['count']} , ignore_index=True)

df_totais_regiao['TOTAL_PERCENTUAL_CEPS_REGIAO'] =
(df_totais_regiao['TOTAL_CEPS_REGIAO']/df_totais_regiao['TOTAL_CEPS_REGIAO'][0]
)*100

df_ceps = pd.concat([df_ceps,df_totais_regiao], axis = 1)

df_totais_subregiao = pd.DataFrame(columns=['TOTAL_CEPS_SUBREGIAO'])

df_cep_subregiao = pd.DataFrame({'count' :
df.groupby(['CEP_SUBREGIAO'])['CEP_SUBREGIAO'].count()}).reset_index()

for i in range(1,101) :

    df_totais_subregiao = df_totais_subregiao.append({'TOTAL_CEPS_SUBREGIAO' :
df_cep_subregiao[df_cep_subregiao['count'] >= i].count()['count']} , ignore_index=True)

df_totais_subregiao['TOTAL_PERCENTUAL_CEPS_SUBREGIAO'] =
(df_totais_subregiao['TOTAL_CEPS_SUBREGIAO']/df_totais_subregiao['TOTAL_CEPS_S
UBREGIAO'][0])*100

df_ceps = pd.concat([df_ceps,df_totais_subregiao], axis = 1)

df_ceps.head()

df_ceps.to_csv('analise_cep.csv')
```

APÊNDICE B – Código-fonte SAS de geração da métrica de erro dos modelos

```
ods listing;

data RESULTADOS_DF;

SET EMWS1.EndGrp3_VALIDATE;

DIF_PERCENTUAL = abs(R_VALOR_COMPRA_VENDA) * 100 /
VALOR_COMPRA_VENDA;

IF DIF_PERCENTUAL = . then RESULTADO_PREVISAO_10 = .;

ELSE IF DIF_PERCENTUAL < 10 then RESULTADO_PREVISAO_10 = 1;

ELSE RESULTADO_PREVISAO_10 = 0;

IF DIF_PERCENTUAL = . then RESULTADO_PREVISAO_15 = .;

ELSE IF DIF_PERCENTUAL < 15 then RESULTADO_PREVISAO_15 = 1;

ELSE RESULTADO_PREVISAO_15 = 0;

IF DIF_PERCENTUAL = . then RESULTADO_PREVISAO_20 = .;

ELSE IF DIF_PERCENTUAL < 20 then RESULTADO_PREVISAO_20 = 1;

ELSE RESULTADO_PREVISAO_20 = 0;

run;

proc print data=RESULTADOS_DF (obs=30);

var VALOR_COMPRA_VENDA DIF_PERCENTUAL RESULTADO_PREVISAO_15;

run;


proc freq data=RESULTADOS_DF;

TABLE RESULTADO_PREVISAO_10 RESULTADO_PREVISAO_15
RESULTADO_PREVISAO_20 ;

run;
```