Companhia de Planejamento do Distrito Federal

Texto para discussão

SEGURANÇA PÚBLICA E INTELIGÊNCIA ARTIFICIAL: UM ESTUDO GEORREFERENCIADO PARA O DISTRITO FEDERAL

Ana Julia Akaishi Padula
Fernanda Santos Amorim
Gustavo Monteiro Pereira
Jader Martins Camboim de Sá
Marcelo Fernando Felix de Oliveira
Mariana Rosa Montenegro
Matheus Facure Alves
Peng Yaohao

nº 33/dezembro de 2017
ISSN 2446-7502

codeplan
COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL

Texto para Discussão

nº 33/dezembro de 2017

SEGURANÇA PÚBLICA E INTELIGÊNCIA ARTIFICIAL: UM ESTUDO GEORREFERENCIADO PARA O DISTRITO FEDERAL

Ana Júlia Akaishi Padula[1]
Fernanda Santos Amorim[2]
Gustavo Monteiro Pereira[3]
Jader Martins Camboim de Sá[4]
Marcelo Fernando Félix de Oliveira[5]
Mariana Rosa Montenegro[6]
Matheus Facure Alves[7]
Peng Yaohao[8]

Brasília-DF, dezembro de 2017

[1] Ana Julia Akaishi Padula - Graduada em Administração pela Universidade de Brasília (UnB).
[2] Fernanda Santos Amorim - Graduanda em Administração pela Universidade de Brasília (UnB).
[3] Gustavo Monteiro Pereira - Graduando em Economia pela Universidade de Brasília (UnB).
[4] Jader Martins Camboim de Sá - Graduando em Ciência da Computação pela Universidade de Brasília (UnB).
[5] Marcelo Fernando Felix de Oliveira - Graduado em Administração pela Universidade de Brasília (UnB).
[6] Mariana Rosa Montenegro - Graduada em Administração e Mestrado em Administração (Finanças e Métodos Quantitativos) pela Universidade de Brasília (UnB).
[7] Matheus Facure Alves - Graduado em Economia pela Universidade de Brasília (UnB).
[8] Peng Yaohao - Graduado em Relações Internacionais e Mestrado em Administração (Finanças e Métodos Quantitativos) pela Universidade de Brasília (UnB).

Texto para Discussão

Veículo de divulgação de conhecimento, análises e informações, sobre desenvolvimento econômico, social, político, gestão e política públicas, com foco no Distrito Federal, na Área Metropolitana de Brasília (AMB) e na Região Integrada de Desenvolvimento do Distrito Federal e Entorno (RIDE) e estudos comparados mais amplos, envolvendo os casos acima.

Os textos devem seguir as regras da **Resolução 143/2014,** que regem o Comitê Editorial da Codeplan, e não poderão evidenciar interesses econômicos, político-partidários, conteúdo publicitário ou de patrocinador. As opiniões contidas nos trabalhos publicados na série Texto para Discussão são de exclusiva responsabilidade do(s) autor(es), não exprimindo, de qualquer maneira, o ponto de vista da Companhia de Planejamento do Distrito Federal - Codeplan.

Companhia de Planejamento do Distrito Federal - Codeplan

Texto para Discussão

TD - n. 33 (2017) - . - Brasília: Companhia de Planejamento do Distrito Federal, 2016.

n. 33, dezembro, 29,7 cm.

Periodicidade irregular.
ISSN 2446-7502

1. Desenvolvimento econômico-social. 2. Políticas Públicas 3. Área Metropolitana de Brasília (AMB). 4. Região Integrada de Desenvolvimento do Distrito Federal e Entorno (RIDE). I. Companhia de Planejamento do Distrito Federal. II. Codeplan.

CDU 338 (817.4)

RESUMO

O presente estudo realizou um estudo de georreferenciamento para a ocorrência de crimes violentos e contra o patrimônio com base em dados do ano de 2013 da Pesquisa Distrital por Amostra de Domicílios (PDAD), mediante metodologias inovadoras de aprendizado de máquinas – uma área do conhecimento recente e que vem, cada vez mais, sendo explorada por pesquisas científicas de ponta em nível global em aplicações diversas. Realizou-se um levantamento bibliográfico sistemático da literatura acadêmica correlata, culminando na seleção e fundamentação de 50 variáveis explicativas presentes na base de dados que subsidiam a previsão de ocorrências futuras de crimes nas 31 Regiões Administrativas do Distrito Federal. Para o cômputo das previsões, foram consideradas as metodologias de regressão logística, classificador Bayes ingênuo (*Naive-Bayes*), árvore de decisão *random forest* e máquina de vetores de suporte (*Support Vector Machine*). Cada conjunto de previsões foi avaliado mediante três métricas de desempenho (acurácia, precisão e revocação) e georreferenciadas com a malha digital do Distrito Federal, gerando mapas que ilustram diferentes gradações da propensão à criminalidade de cada Região Administrativa. Os resultados mostram que as previsões atingiram alta aderência aos dados reais, evidenciando que, com base nas variáveis explicativas escolhidas, a ocorrência de crimes no Distrito Federal possui um grau razoável de previsibilidade, visto pelo mapeamento de zonas de alta propensão a crimes. Os resultados obtidos são de grande valia para a sociedade civil e gestores de políticas públicas, os quais podem utilizá-los como insumo para fundamentar a formulação e execução de medidas de segurança pública de intervenção *ex-post* e políticas de desenvolvimento social de intervenção *ex-ante* para combater a criminalidade e promover o bem-estar da população nessa região.

Palavras-chave: Aprendizado de Máquina; Criminalidade; Desenvolvimento Social; Georreferenciamento.

SUMÁRIO

1. INTRODUÇÃO

Tem-se observado um aumento de crime violentos no Distrito Federal e Entorno, dado o crescimento da taxa de homicídios nessa região, a qual subiu 39,3% entre 2004 e 2014, tendência esta que apresenta evidências de continuidade, visto pelo crescimento dessa taxa em 7,5% de 2013 para 2014. (WAISELFSZ, 2016). Dessa forma, de um lado, o mapeamento espacial adequado da ocorrência de práticas criminosas é desejável para a sociedade civil, cuja integridade e segurança devem ser plenas e perenes; de outro, o tema possui grande relevância para governantes e gestores de políticas públicas, que possuem como responsabilidade incorporar as demandas sociais e garantir à população seus direitos inalienáveis.

Sendo assim, o presente estudo realiza uma previsão das zonas de crime das 31 Regiões Administrativas (RAs) do Distrito Federal mediante quatro métodos de previsão, tendo como base variáveis explicativas extraídas de microdados da Pesquisa Distrital por Amostra de Domicílios (PDAD) do ano de 2013. As previsões foram então mapeadas utilizando técnicas de georreferenciamento, gerando quatro mapas da região, escalonados pela frequência de crimes registrados.

Nessa pesquisa, dois prismas analíticos foram analisados. Estes podem auxiliar a construção de políticas públicas eficazes em prol da segurança pública e do desenvolvimento social no Distrito Federal: uma dimensão *ex-ante-facto* e uma dimensão *ex-post-facto*, elucidadas a seguir:

Por “dimensão *ex-post-facto*”, entende-se como o conjunto de ações intervenientes tomadas **após** a ocorrência de práticas criminosas. Ou seja, com base em dados históricos de criminalidade na região, busca-se gerar previsões para futuras ocorrências com base nas variáveis explicativas levantadas, de modo que essas previsões irão informar localidades de alta periculosidade nas quais recomenda-se uma intervenção policial redobrada, de modo a coibir a ação de criminosos e assegurar a segurança e o bem-estar da população;

Por “dimensão *ex-ante-facto*”, entende-se como o conjunto de fatores relevantes para a determinação da criminalidade de uma localidade, mas que exercem uma influência **antes** de o fato gerador “crime” de fato se concretize. Caso sejam identificados quais fatores são relevantes para justificar as **causas** dos crimes, há uma margem para que se façam intervenções nessas “dimensões ocultas” que possam impedir a própria ocorrência de crimes futuros, dispensando parcialmente assim a necessidade de qualquer tipo de intervenção *ex-post*. Ao levantar variáveis independentes que forneçam com precisão um diagnóstico das possíveis causas para a escalada ou para a coibição da atividade criminosa, o presente estudo pode fornecer *insights* valiosos para fatores alternativos que afetam a ocorrência de crimes – como preferências culturais e recreativas – e que podem ser exploradas por formuladores de políticas públicas como uma intervenção *ex-ante*: tornar o ambiente social menos propício à germinação de futuros meliantes.

Dessa forma, focando nessas duas dimensões igualmente pertinentes, o presente estudo tem como objetivos primordiais:

1. Realizar uma busca exaustiva da literatura científica correlata às causas determinantes do crime; e, tendo esses estudos especializados como fundamento, levantar uma lista de variáveis relevantes para esse fenômeno, as

quais serão utilizadas como insumos para a previsão de futuras atividades criminosas, mediante a aplicação de quatro métodos de classificação e aprendizado de máquinas. Com base na qualidade preditiva apresentada por esses classificadores, espera-se que um bom desempenho preditivo represente uma alta significância das variáveis levantadas para o poder explicativo do modelo especificado, de modo que uma análise cuidadosa dessas variáveis pode revelar fatores estruturais para a formulação de boas políticas públicas com cunho de intervenção *ex-ante*;

2. Com base nas previsões geradas pelos algoritmos de aprendizado de máquina, confeccionar mapas que ilustrem a propensão das RAs do DF à criminalidade, delimitando-se claramente zonas de alta necessidade de intervenção *ex-post* (uma vez que a criminalidade é intrinsecamente dependente da consumação da atividade criminosa em si). Os resultados podem subsidiar decisivamente na otimização da alocação de forças policiais, unidades de patrulha e de futuras instalações de segurança, tais como delegacias e postos de atendimento comunitário.

2. REFERENCIAL TEÓRICO

Os estudos científicos de Becker (1968) na área de criminalidade deram início às discussões acerca das variáveis que influenciam a decisão de cometer um crime, abordando um raciocínio econômico entre os custos e benefícios de um crime. Muitos outros estudos foram derivados da problemática apresentada por Becker (1968), que seria verificar como o comportamento de um indivíduo influencia a decisão de cometer um crime. Para analisar os padrões de comportamento, geralmente são estudadas variáveis como renda, escolaridade, moradia, entre outros. Artigos de estudos semelhantes mostram que as interações familiares, sociais e educacionais, ou seja, com o meio em que o indivíduo está inserido, são determinantes na criminalidade (GLAESER; SACERDOTE; SCHEINKMAN, 1995), e também que a criminalidade é proveniente de desigualdades sociais e carência de recursos para sustentar certos padrões de consumo (MERTON, 1959).

Em estudos feitos no Rio Grande do Sul (OLIVEIRA, 2008), evidenciou-se que além dos aspectos gerais citados acima, também se deve considerar um aspecto moral que é construído pelo decisor, como barreira para a escolha de cometer o crime. Este aspecto cultural abrange um processo de construção psicológica do indivíduo, desde a infância, que remete a relações familiares, até a vida adulta, remetendo a relações com a sociedade, ou seja, o ambiente em que possui grande influência nas perspectivas do indivíduo. (BRONFENBRENNER, 1979).

O fenômeno de altos índices de criminalidade ainda é realidade nas cidades do Distrito Federal, de acordo com o Balanço feito pela Secretaria de Segurança Pública do DF do último trimestre (julho/agosto/setembro 2016). O presente estudo procura fazer uma análise espacial dos índices de criminalidade do DF relacionando com o padrão de comportamento de cada RA de acordo com educação, renda, cultura e aspectos pessoais dos entrevistados. Para isso, serão utilizados os microdados da Pesquisa Distrital por Amostra de Domicílios (PDAD) de 2013, versão mais atual publicada. De forma a simplificar, elas foram agrupadas em quatro categorias: pessoa, cultura, educação e renda.

A Econometria do Crime e estudos econométricos de forma geral possuem interesse maior na explicação dos fatos, isto é, parte-se de uma teoria, desenvolve-se um modelo e testa-se esse modelo. Nos modelos de aprendizagem de máquina, há um interesse maior na previsão gerada pelo modelo e na sua capacidade de generalização e utilização dos dados como início (VARIAN, 2014). Com isso, este trabalho busca avaliar quais são os principais locais de crime do Distrito Federal e avaliar se o ambiente em que o indivíduo está inserido influencia na prática criminosa, a fim de se propor políticas públicas mais eficazes.

Segundo Santos e Kassouf (2008), as principais variáveis utilizadas na literatura têm sido socioeconômicas, entre elas renda, taxa de desemprego, nível de escolaridade, pobreza, desigualdade de renda e urbanização. Além disso, informações demográficas têm papel relevante, como questões de raça, gênero, estado civil e escolaridade. Como forma de considerar estas questões, as variáveis escolhidas estão descritas abaixo, bem como a motivação para a escolha delas.

Composta por 31 Regiões Administrativas (RAs), os dados das RAs serão cruzados com informações da base de dados disponibilizada pela Polícia Civil do Distrito Federal a fim de prever locais de crime do Distrito Federal por meio de georreferenciamento. Desta forma, a variável de localização busca mapear os perfis dos indivíduos das regiões.

Pela junção dos conceitos de geografia e *marketing*, tendo o primeiro como sendo a distribuição territorial dos fenômenos e o segundo como "o ato de conhecer o mercado de atuação de uma organização, para posteriormente oferecer, de forma inovadora e criativa, produtos e serviços que esse mercado deseja" (JUNIOR, 2007). Define-se *geomarketing* como "a disciplina que estuda as relações existentes entre as estratégias e políticas de *Marketing* e o território ou espaço, onde a instituição, seus clientes, fornecedores e pontos de distribuição se localizam" (JUNIOR, 2007).

A base do sucesso dessa metodologia se dá pela identificação, em detalhes, de grupos de clientes específicos que possuem características homogêneas. Para identificar esses segmentos, é necessário um número muito grande de informação para, assim, conhecer as particularidades de cada grupo e poder satisfazer as suas necessidades. O processo de segmentação é longo e complexo, pois exige a confirmação de que os segmentos existem, a determinação das suas características e localização para que, a partir dessas informações, se possa adotar medidas específicas para cada região com base nas características da população (SHEPARD, 1993).

Os Sistemas de Informação Geográfica (SIG) podem ser de grande utilidade para esse tipo de tarefa. Os SIG integram três tipos de arquivos: banco de dados; arquivos geográficos; e arquivos de pontos. O banco de dados contém as informações puramente externas à empresa, por exemplo, dados econômicos, demográficos e sociais do mercado. Os arquivos geográficos contêm as entidades geográficas definidas por suas coordenadas e servem para a produção dos mapas. O terceiro tipo de arquivo é a união dos dois primeiros, onde os dados coletados ficam associados à sua localização geográfica. A junção desses três arquivos torna possível a criação de mapas e a aplicação de cores, padrões e símbolos, representando simultaneamente diversos tipos de dados. O resultado final é a análise de potencial de mercado, segmentação, localização de clientes ou grupos de foco e até mesmo a identificação de regiões que necessitam, como é o caso deste trabalho, de ações governamentais e políticas públicas (ARANHA, 1996).

As variáveis Número de Moradores do Domicílio, Condição no Domicílio, Estado Civil são utilizadas como *proxies* para a estrutura familiar. O interesse está em identificar se há alguma relação entre famílias com maior número de moradores e falta de planejamento familiar. Santos e Kassouf (2008) indicam que há uma relação entre núcleo familiar estável e menos prática de crimes violentos, para isso, a variável de Estado Civil também funciona como uma *proxy*.

No que concerne aos fatores demográficos básicos, as variáveis Sexo, Raça e Idade buscam contemplar estes aspectos. Espera-se uma relação inversa entre a quantidade de crimes cometidos e a idade da população (ALMEIDA, 2007). Além disso, a maior parte dos crimes é cometida por indivíduos do sexo masculino. A variável raça busca complementar a análise de desigualdade social (MACEDO *et al*., 2001).

Já as variáveis Naturalidade, Ano de Chegada ao DF, Motivo da Mudança, Tempo na RA atual e Tempo no Domicílio buscam incorporar os efeitos econômicos de migrações, isto é, indivíduos que se mudam para o DF à procura de novas oportunidades de emprego e têm esses interesses frustrados. Foote (2015) indica haver uma relação entre taxas de crime da cidade e migrações dentro de um mesmo país. Desta forma, essas variáveis buscam englobar estes fatores na previsão. Estas variáveis também têm o papel de identificar pontos de instabilidade familiar, seja por renda, seja por relacionamento entre os membros da família.

As variáveis relacionadas à cultura presentes na base de dados utilizada foram: Frequência em Museus, Frequência em Cinema, Frequência em Biblioteca, Frequência em

Teatro, Frequência em Shows, Frequência em Parques e Jardins, Hábito de Leitura e Participação em Atividades Extracurriculares. Este grupo de variáveis é utilizado para analisar o nível de relação com cultura que o indivíduo possui como forma de associar com o fato de que o processo de integração do indivíduo com a sociedade por meio da cultura pode ser visto como formador do custo moral citado acima (OLIVEIRA, 2008). As variáveis de religião podem também funcionar como *proxies* para aspectos culturais (FAJNZYLBER; LEDERMAN; LOAYZA, 2002). Desta forma, duas variáveis de religião presentes na base buscam considerar estes elementos. Além disso, elas podem ter papel ao incorporar o custo moral indicado por Oliveira (2008). Assim, indivíduos que frequentam algum tipo de religião podem ter o custo moral de se realizar o ato criminoso mais alto que indivíduos que não frequentam ou estes podem variar entre religiões diferentes.

As variáveis Pratica Atividade Esportiva e Frequência em Espaços desportivos também se relacionam com o custo moral, trazendo um pressuposto de estudos já feitos mostrando que a prática de atividades esportivas é uma maneira de integração e inclusão social (MALINA; CESARIO, 2009). Além disso, estudos feitos por Peres (2013) em favelas do Rio de Janeiro mostram que o esporte também pode ser usado como forma de afastamento da criminalidade e das drogas. É importante ressaltar que as atividades esportivas incentivadas por projetos do Governo em comunidades carentes podem não só auxiliar no afastamento da criminalidade e drogas, como também no aumento das perspectivas profissionais dos indivíduos contemplados (PERES, 2013).

Outro grupo de variáveis presentes no estudo são as variáveis relacionadas à renda. Algumas proxies foram utilizadas para avaliar a renda: Situação de Atividade, Setor da Atividade Remunerada, Posição na Ocupação, Contribuição para a Previdência, Dispor de Plano de Saúde, Motivos dos Aposentados que voltaram a Trabalhar e Valor do Benefício Social. Neste caso, as variáveis apresentadas para relacionar com a renda do indivíduo são importantes para medir o bem-estar das famílias (SOMAVILLA, 2015) e também para explicar como os fatores econômicos influenciam na decisão do crime (MENDONÇA, 2001).

Considerando que a educação tem fator bastante relevante na determinação da execução de crimes, foram escolhidas três variáveis que levam isso em conta: Nível de Escolaridade, Local Onde Estuda e Se Estudava, no momento da pesquisa. Lochner e Moretti (2004) apontam que a educação pode reduzir a taxa de criminalidade por diversos fatores: reduzir a taxa de criminalidade por diversos fatores, entre eles, e o principal, pode ser o aumento dos salários, o que representa um custo maior ao se cometer um crime e de passar algum período na prisão, pois o ganho econômico por meio de atividades ilícitas tenderia a oferecer menos incentivos em relação a uma maior remuneração legítima, entre eles, e o principal, pode ser o aumento dos salários. Neste mesmo estudo, os autores identificaram que um aumento de 1% na taxa de formação no ensino médio economizaria R$1,4 bilhão por ano com os crimes sofridos pela sociedade em geral. Isso indica que políticas públicas neste âmbito poderiam trazer grandes retornos à sociedade.

Lobo e Fernandez (2003) verificaram na Região Metropolitana de Salvador que o nível de educação contribui significativamente para a redução da atividade criminosa. Desta forma, o modelo tratado neste trabalho busca incorporar as variáveis de nível de escolaridade, local onde estuda e também se o indivíduo é estudante de escola pública ou particular para considerar estes efeitos.

Em relatório apresentado para o Congresso dos Estados Unidos, (SHERMAN *et al.*, 1998) realizaram uma análise de diversos tipos de políticas públicas voltadas para a redução do crime. Essas políticas eram divididas para diferentes públicos: crianças, pré-adolescentes em situação de risco, escolas, ex-detentos e zonas de risco. As soluções apresentadas para cada um podem ser sintetizadas como: aulas com visita semanal dos(as) professores(as) às crianças; terapia familiar; estabelecimento de normas claras e

consistentes e *coaching*; treinamento vocacional; aumento de patrulha policial. Estas políticas aparentam ter alto custo, entretanto, como dito anteriormente, os valores que poderiam ser economizados com crimes ultrapassaram um bilhão de dólares nos Estados Unidos (LOCHNER; MORETTI, 2004).

Considerando a realidade brasileira, pode ser citado o Programa Nacional de Segurança Pública com Cidadania (Pronasci), promovido pelo Ministério da Justiça em 2007 e que busca a melhoria da segurança aliando medidas preventivas, atacando a causa dos crimes desde o princípio. Todavia, devido ao seu caráter descentralizador, ele não pode ser plenamente estabelecido devido a algumas fragilidades dos municípios (MADEIRA; RODRIGUES, 2015). Por não ter a adesão obrigatória, não foi plenamente adotado tendo em vista algumas medidas não valerem o custo-benefício. Considerando isso, este estudo busca aliar a tecnologia às necessidades sociais e mostrar informações contidas nos dados que podem não ser vistas tão facilmente para, assim, mostrar potenciais soluções para as medidas como as indicadas pelo Pronasci sejam eficazes.

3. METODOLOGIA

Como disposições gerais, foram escolhidos métodos de aprendizado de máquinas para as previsões do presente trabalho, dado que se trata de uma área do conhecimento que apresentou grande expansão de popularidade nos últimos anos, figurando-se como uma das agendas de pesquisa mais promissoras, para as próximas décadas, conforme indicam LeCun, Bengio e Hinton (2015). A abordagem de aprendizado de máquinas é notória por sua versatilidade. Vide aplicações recentes com sucesso em diversas áreas do conhecimento, tais como: previsão de séries temporais (ABIDIN; JAAFAR, 2012; LIN; HU; TSAI, 2012), previsão de taxas de câmbio (LI; SUOHAI, 2013; SERMPINIS *et al.*, 2015), análise de sentimentos (BALAHUR; TURCHI, 2014), diagnóstico de doenças (BIBAULT; GIRAUD; BURGUN, 2016; KOTSAVASILOGLOU *et al.*, 2017), classificação de imagens (DORNAIKA *et al.*, 2016), sustentabilidade ambiental (PHAM *et al.*, 2016), e – específico para a temática do presente estudo – estudos de criminalidade (BERK; BLEICH, 2013; BRENNAN; OLIVER, 2013).

A base de dados utilizada foi a PDAD de 2013, sua versão mais atualizada até a conclusão deste trabalho. Foram selecionadas 50 variáveis independentes extraídas dessa base, cujos fundamentos teóricos na literatura científica foram elucidados na seção anterior do referencial teórico. A variável dependente é binária e relativa à ocorrência de atividades criminosas, assumindo valor *1* para presença de crime, e *0* para ausência de crime.

A base de dados é constituída por 85.797 observações, entre as quais foram selecionadas 70% (60.056 observações do total de 85.797) para se realizar o treinamento dos algoritmos classificadores, enquanto os 30% restantes (25.739 observações do total de 85.797) foram utilizados para realizar as previsões e conferir o desempenho preditivo. Esse mecanismo objetiva mitigar o viés de *data-snooping* (WHITE, 2000), o qual diz respeito a uma inflação artificial da qualidade das previsões oriunda do uso repetido de um dado tanto para definir as curvas de decisão quanto para avaliar seu poder explicativo – ao segmentar a base em dois conjuntos mutuamente excludentes, a qualidade dos modelos levantados pode ser melhor verificada, ao se alimentar as funções de decisão com dados que não foram utilizados para definir suas respectivas formas funcionais, de modo a fornecer uma medida mais robusta para a capacidade de previsão dos modelos propostos.

Dado que a variável dependente predita é binária, é evidente que há quatro cenários possíveis para os valores observado e predito da dita variável, apresentados no Quadro 1 a seguir:

Quadro 1 - Classes das previsões confrontadas com dados observados

Previsão\Observação	Não ocorreu crime	Ocorreu crime
Não se previu crime	**VN**:Não-crimes classificados corretamente	**FN**: Crimes não previstos – Erro tipo I (falso negativo)
Previu-se crime	**FP**: Previsão de crime que não ocorreu (“alarme falso”) – Erro tipo II (falso positivo)	**VP**: Crimes classificados corretamente

O presente trabalho considerou três métricas de avaliação das previsões: acurácia, precisão e revocação (*recall*), definidas a seguir:

$$\text{Acurácia} = \frac{\text{classificações corretas}}{\text{total de observações}} = \frac{\boldsymbol{VP} + \boldsymbol{VN}}{\boldsymbol{VP} + \boldsymbol{FP} + \boldsymbol{VN} + \boldsymbol{FN}} \quad (1)$$

$$\text{Precisão} = \frac{\text{crimes preditos corretamente}}{\text{total de crimes preditos}} = \frac{\boldsymbol{VP}}{\boldsymbol{VP} + \boldsymbol{FP}} \quad (2)$$

$$\text{Revocação} = \frac{\text{crimes preditos corretamente}}{\text{total de crimes observados}} = \frac{\boldsymbol{VP}}{\boldsymbol{VP} + \boldsymbol{FN}} \quad (3)$$

A acurácia fornece a taxa de acerto do classificador em relação ao total de observações, e em geral é a métrica padrão para a avaliação da qualidade de previsões em variáveis binárias. No entanto, tendo em vista que grande parte dos entrevistados **não** sofreu crimes, o desbalanceamento das classes "crime" e "não crime" tende a elevar artificialmente o valor da acurácia, motivando assim o uso de outras métricas que consideram a qualidade da previsão relativo ao total de elementos observados e preditos da classe de interesse (nesse caso, indivíduos que sofreram crime), o que define a precisão – que retorna a qualidade da previsão em relação ao total de previsões realizadas; e o *recall* – que fornece a proporção de crimes que aconteceram de fato e que foram corretamente previstos. Ao usar as três métricas, o estudo tratou com maior robustez a avaliação da real qualidade das previsões, bem como logrou êxito em identificar vantagens e fraquezas de cada algoritmo classificador considerado.

3.1. Georreferenciamento

Para a construção dos mapas, foi utilizado o programa Qgis. Este software permite a junção, de forma prática, entre a previsão de crimes e as malhas digitais do Distrito Federal. Como resultado desse cruzamento de dados, foram criados quatro mapas, refletindo os resultados provenientes das metodologias abordadas, coloridos de acordo com o nível de previsão de crimes encontrado em cada região de análise.

Apesar das pesquisas nos sites governamentais responsáveis e de algumas tentativas de contato com funcionários do IPEA, não foi possível encontrar nenhuma malha digital do Distrito Federal dividida entre as 31 Regiões administrativas. Portanto, a solução foi utilizar a malha digital disponibilizada no site do IBGE que está dividida por setores censitários, estabelecidos pelo próprio Instituto Brasileiro de Geografia e Estatística para a realização de pesquisas, e que está dividida entre os 19 subdistritos do Distrito Federal.

3.2. *Support Vector Machine*

O *Support Vector Machine* - SVM (BOSER *et al.*, 1992; CORTES & VAPNIK, 1995) é um algoritmo de aprendizagem supervisionada que fornece uma função de decisão de classificação que discrimina os dados observados em duas classes complementares, $+1$ e -1, de modo a maximizar a margem entre as classes. Basicamente, dadas n observações com p variáveis explicativas e a respectiva classe à qual a n-ésima observação pertence, o SVM é um método que permite computar a expressão de uma curva em $\mathbb{R}^p$ que separa as

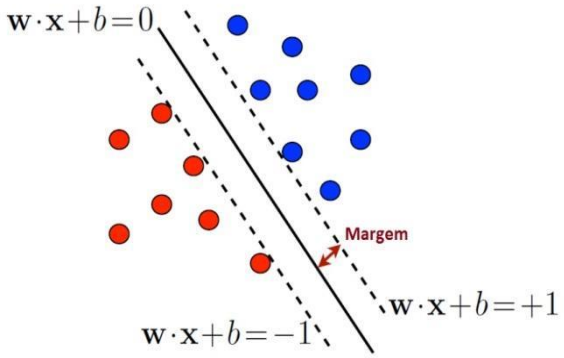
duas classes (assumindo-se que os dados são separáveis em $\mathbb{R}^p$), permitindo realizar inferências em relação à $n+1$–ésima observação das classes p variáveis explicativas, predizendo a classe à qual essa nova observação irá pertencer.

Apontado pela literatura como um dos melhores métodos para classificação, o SVM tem sido amplamente utilizado em diversas áreas do conhecimento científico na medida em que consegue realizar a separação de dados considerando estruturas complexas de não linearidade, ao mesmo tempo em que exige poucos parâmetros na sua estimação (JUNG & KIM, 2014; MOGHADDAM & HAMIDZADEH, 2016). O objetivo do SVM é encontrar o hiperplano que maximiza a margem entre as classes, condicionado às observações **x** tomadas.

$$\sum_{i=1}^{p} \mathrm{w_i\, x_i} - w_0 = \mathbf{w}\cdot\mathbf{x} - \mathrm{w_0} = 0 \qquad (4)$$

Considerando que as observações pertencentes à classe $+1$ estão acima do hiperplano $\mathbf{w}\cdot\mathbf{x} - \mathrm{w_0} = +1$ e que as observações da classe -1 estão abaixo de $\mathbf{w}\cdot\mathbf{x} - \mathrm{w_0} = -1$, (representados pelas retas tracejadas na figura abaixo). A função de decisão para a predição da $n+1$–ésima observação será o hiperplano representado pela reta cheia da Figura 1, paralela a $\mathbf{w}\cdot\mathbf{x} - \mathrm{w_0} = +1$ e $\mathbf{w}\cdot\mathbf{x} - \mathrm{w_0} = -1$. A classe prevista de novas observações será dada tendo esse hiperplano como referencial.

Figura 1 - Representação do SVM para dados linearmente separados



Fonte: Adaptado de (Mohri; Rostamizadeh; Talwalkar, 2012)

O hiperplano de separação é obtido solucionando-se o seguinte problema de otimização quadrática:

$$\text{Minimizar:} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{Sujeito a:} \quad \mathbf{D}(\mathbf{A}\mathbf{w} - \mathrm{w_0}\mathbf{1}) \geq 1 \qquad (5)$$

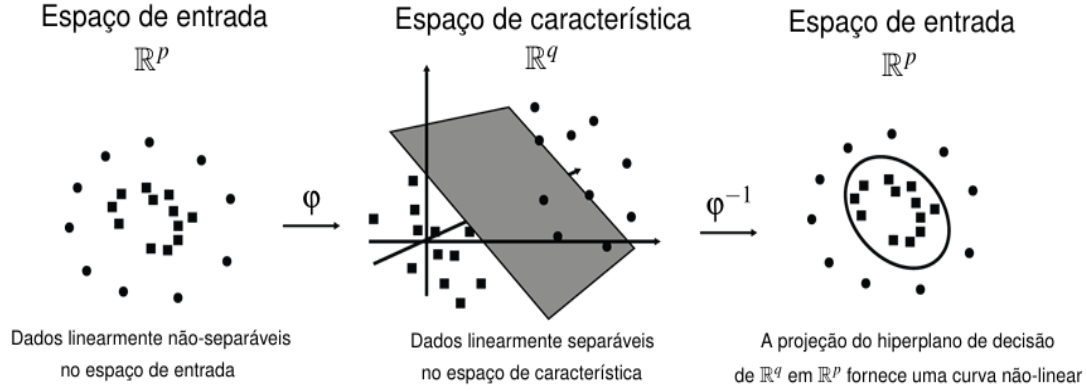
$$w_0 \in \mathbb{R}, \boldsymbol{w} \in \mathbb{R}^p$$

onde $\boldsymbol{w}$ é o vetor coluna $pX1$ referente aos coeficientes angulares do hiperplano de decisão; w_0 é um escalar real relativo ao intercepto (coeficiente linear) do hiperplano de decisão; $\boldsymbol{D}$ é

a matriz diagonal nXn com os valores de $\boldsymbol{y}$; e $\mathbf{1}$ é o vetor coluna $nX1$ com todos os valores iguais a 1.

Para o caso em que as observações em $\mathbb{R}^p$ **não** podem ser separadas por uma função linear, realiza-se uma transformação $\varphi(\boldsymbol{x}) \in \mathbb{R}^q$ nas variáveis explicativas para uma dimensão superior na qual os dados se tornam linearmente separáveis. O espaço original $\mathbb{R}^p$ é denominado "espaço de entrada" (*input space*), e o espaço induzido por φ, $\mathbb{R}^q$, é denominado "espaço de característica" (*feature space*). Assim, a matriz de observações A_{nXp} será mapeada em uma dimensão superior como Φ_{nXq}. No espaço de característica $\mathbb{R}^q$, procede-se da mesma forma para o SVM do caso com dados linearmente separáveis: constroem-se dois hiperplanos que separem as duas classes, dos quais se deriva o hiperplano de decisão para realizar predições para observações novas. Ao retornar para o espaço de entrada, a curva que separa as classes poderá ser uma função não linear, dado que as curvas de nível da secção transversal do hiperplano em $\mathbb{R}^q$ com a transformação φ, projetadas na dimensão inferior $\mathbb{R}^p$, podem assumir formas não lineares. A Figura 2 sintetiza essa ideia, ilustrando a redução do problema de classificação não linear para um problema de classificação linear mediante a aplicação da transformação φ.

Figura 2 - Intuição do SVM para dados não lineares



Fonte: Modificado a partir de (Soman; Loganathan e Ajay, 2011)

Ademais, a fim de evitar que a função de decisão seja demasiadamente volátil – o que prejudica a capacidade de generalização do modelo – introduz-se um parâmetro fixo de custos que atribui uma penalização às classificações errôneas (ou seja, elementos da classe +1 classificados como –1, e vice-versa), ponderados pela distância dos pontos classificados erradamente, a qual também é conhecida e dada pelo vetor $\boldsymbol{\xi}$. Dessa forma, o algoritmo irá buscar o "meio-termo" entre a maximização de pontos classificados corretamente e a preferência por curvas de decisão com comportamento mais estável possível (ou seja, mais lineares possíveis no espaço de entrada).

Dessa forma, realizando-se a transformação das observações originais $\boldsymbol{x} \longrightarrow \varphi(\boldsymbol{x})$ e introduzindo-se o parâmetro de custo C, o problema de otimização do SVM para separação não linear -- conhecido como "SVM de margem suave" é dado por:

$$\text{Minimizar:} \qquad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\mathbf{1}^T\boldsymbol{\xi} \tag{6}$$

Sujeito a: $\mathbf{D}(\mathbf{\Phi w} - \mathrm{w}_0\mathbf{1}) + \boldsymbol{\xi} \geq \mathbf{1}$

$$\boldsymbol{\xi} \geq \mathbf{0}$$

$$w_0 \in \mathbb{R}, \boldsymbol{w} \in \mathbb{R}^q$$

Derivando o dual de Wolfe do SVM não linear, a matriz $K = \mathbf{\Phi}\mathbf{\Phi}^T$, que representa os produtos internos das transformações φ, aparece na função objetivo do dual. Dessa forma, caso possa computar diretamente os produtos internos de φ mediante uma **função *Kernel***, não haverá necessidade de se aplicar explicitamente a transformação φ aos dados originais, o que torna a otimização do problema significantemente menos onerosa. A função *Kernel* adotada no presente estudo foi o *Kernel* Gaussiano, dado por:

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp\left(-\gamma\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2\right) \quad (7)$$

onde $\exp(.)$ é a função exponencial e $\|.\|$ é a norma euclidiana.

O *Kernel* Gaussiano é a função *Kernel* mais utilizada pela literatura de aprendizado de máquinas (LI & SUOHAI, 2013; GONG *et al.*, 2016), dado que é capaz de computar um espaço de características de dimensão **infinita** (STEINWART *et al.*, 2006), apesar de exigir apenas um parâmetro $\boldsymbol{\gamma}$ para sua estimação. A capacidade do *Kernel* Gaussiano em abarcar sinteticamente uma gama de interações não lineares entre as variáveis explicativas e a elevada acurácia das previsões de algoritmos que fazem uso dele se estabeleça que essa função seja a escolha padrão para estudos de aprendizado de máquinas (KAMRUZZAMAN *et al.*, 2003; SUN & LI, 2012).

3.3. Regressão logística

Regressão logística é uma técnica estatística utilizada para modelar distribuições binomiais. Nesse sentido, em termos práticos, é mais próxima aos modelos de classificação supervisionada do que de uma regressão. No presente estudo, as variáveis são binárias. O modelo tem seu nome derivado da função logística, dada por:

$$\kappa F(x) = \frac{e^x}{1 + e^x} \quad (8)$$

Essa função varia entre 0 e 1, sendo o seu resultado facilmente interpretado como probabilidade, daí o seu uso difundido em problemas de classificação.

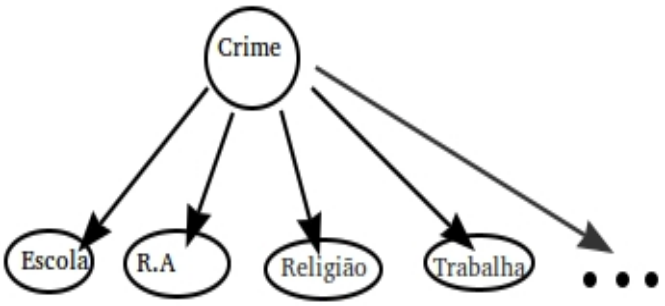
O modelo logístico está na classe de modelos lineares generalizados e pode ser obtido da seguinte forma:

$$F(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}} \quad (9)$$

3.4. Bayes Ingênuo

O algoritmo Bayes Ingênuo assume que estamos trabalhando com uma rede Bayesiana cujos nós na base são condicionalmente independentes. No nosso caso, a variável não observada da rede seria o crime, e as outras características são as variáveis observadas. Com informação dessas últimas então, o algoritmo deduz os parâmetros da rede Bayesiana, usando aprendizagem supervisionada, por meio de estimadores de máxima verossimilhança e teorema de Bayes. A rede que estamos tentando estimar pode ser representada na Figura 3.

Figura 3 - Rede Bayesiana



Fonte: Elaborado pelos autores

A rede acima pode parecer artificial por não fluir nos sentidos da causalidade, mas tem a vantagem de representar a distribuição de probabilidade conjunta com uma quantidade muito menor de parâmetros do que se tivéssemos que combinar cada variável exaustivamente para estabelecer essa mesma distribuição. Isso tudo, no entanto, vem ao custo da pressuposição ingênua do algoritmo, que é assumir independência condicional entre os nós na base da rede. Isso obviamente é falso, e podemos tecer vários argumentos de por que o fato de o indivíduo trabalhar, por exemplo, é dependente de ele estar na escola, mesmo dada a variável de crime para que a dependência seja condicional. Com esse pressuposto e como o teorema de Bayes, temos:

$$P(y|x_{1,}\dots,x_n) = \frac{P(y)P(x_{1,}\dots,x_n|y)}{P(x_{1,}\dots,x_n)} \tag{10}$$

Usando a hipótese de independência condicional:

$$P(x_i|y,x_{1,\dots},x_{i-1},x_{i+1},\dots,x_n) = P(x_i|y) \tag{11}$$

Portanto:

$$P(y|x_{1,}\dots,x_n) = \frac{P(y)\prod_{i=1}^{n} P\,(x_i|y)}{P(x_{1,}\dots,x_n)} \tag{12}$$

Como o denominador será idêntico para todas as classes de y, e estamos apenas interessados em comparar qual delas tem maior probabilidade dado o vetor de características, podemos ignorar o denominador e trabalhar com pseudoprobabilidades.

$$P(y|x_{1,}\ldots,x_n) \propto P(y)\prod_{i=1}^{n} P\,(x_i|y) \tag{13}$$

$$\hat{y} = \underset{y}{argmax} P(y)\prod_{i=1}^{n} P\,(x_i|y) \tag{14}$$

Como exemplo, considere que os dados sejam como na Tabela 1.

Tabela 1 - Exemplo hipotético para análise de ocorrência do crime

Sofreu Crime	Trabalha	Sexo	Casado(a)
0	1	1	1
0	1	0	0
0	1	0	0
0	0	1	1
1	1	0	1
1	0	1	1
1	0	1	0

Fonte: Elaborado pelos autores

Suponha agora que recebemos o vetor de características de um novo indivíduo e queremos saber se ele vai ou não sofrer um crime. Suponha o vetor $\boldsymbol{x} = (\mathbf{1}, \mathbf{1}, \mathbf{0})$. A pseudoprobabilidade do indivíduo sofrer um crime é:

$$P'(y|x_{1,}x_{2,}x_3) = P(y)P(x_1|y)P(x_1|y)P(x_3|y) \tag{15}$$

Substituindo os valores para o caso de sofreu crime e não sofreu crime, temos:

$$P'(y|x_{1,}x_{2,}x_3) = \frac{3}{7}\frac{1}{3}\frac{2}{3}\frac{1}{3} = 0.031 \tag{16}$$

$$P'(Noty|x_{1,}x_{2,}x_3) = \frac{4}{7}\frac{3}{4}\frac{2}{4}\frac{2}{4} = 0.107 \tag{17}$$

Assim, o classificador preveria que o nosso novo indivíduo ainda não visto não irá sofrer um crime, dado que a pseudoprobalidade desse caso é bem maior. Esse exemplo mostra um problema que pode surgir: a pseudoprobabilidade depende da probabilidade a priori P(y) e quando ela é muito baixa, o classificador simplesmente preverá que ninguém é do caso y. Quando aplicarmos esse algoritmo nos dados da PDAD, devemos levar esse problema muito a sério, uma vez que menos de 10% das pessoas pesquisadas sofreram crimes.

Algumas observações finais a serem feitas é que a explicação acima é bem mais simplificada do que o algoritmo que de fato usaremos. Um exemplo é que uma das classes de y não tivesse nenhum exemplo positivo de uma das características x, a pseudoprobabilidade seria automaticamente zero, independente de quão para cima as outras características a puxassem. Outro exemplo é que pressupomos acima que todas são características não binárias, o que nem sempre é o caso. Esses problemas são corrigidos pelo algoritmo utilizado, mas a explicação de como isso é feito é demasiado extensa, e acreditamos que o que foi tratado acima já é suficiente para entender como o algoritmo funciona.

3.5. Árvore de Decisão (*Decision Tree*)

O algoritmo *Decision Tree* é um método não paramétrico de aprendizado supervisionado usado para classificação e regressão, este tem por objetivo criar uma árvore de decisão que permita fazer previsões sobre valores e classes aprendendo sobre os atributos dos dados e inferindo regras sobre o procedimento a ser utilizado para a classificação/regressão. Matematicamente, dado um vetor de treino $x_i \in R^n$, i=1,2,3… e um vetor de classe $y \in R^l$, a árvore de decisão recursivamente particiona o espaço de forma que índices de mesma classe sejam agrupados juntos arranjando um candidato no vetor de treino para essa separação.

Neste trabalho, tal algoritmo foi aplicado para nos expor seus procedimentos e regras de classificação sobre dados demográficos de cidadãos do Distrito Federal, prevendo para cada indivíduo se sofreria violência ou não. Em outras palavras, será exposta uma sequência de operações ramificadas baseadas em comparações de quantidades a respeito dos atributos, sendo as comparações atribuídas por fim alguma classe-alvo. Sendo o algoritmo otimizado, ele trata de escolher os aspectos mais impactantes para separações de classe, assim, teremos quais os fatores principais que podem levar um indivíduo a sofrer de algum ato de violência.

4. RESULTADOS

4.1. Regressão logística

Como aproximadamente 93% das pessoas na pesquisa não sofreram nenhum tipo de crime, a base de dados de crime pode ser considerada severamente desbalanceada. Assim, qualquer classificador conseguirá uma acurácia de 93% simplesmente prevendo que ninguém sofrerá crime. Para lidar com esse problema, estamos dispostos a sacrificar a acurácia e medirmos a capacidade do modelo com precisão e revocação, com a primeira medindo a capacidade de identificar as vítimas, e a segunda, a quantidade de verdadeiros-positivos sobre o total de positivos previstos.

Para tanto, mudamos os pesos das classes positivas (as que sofreram crime) e negativas (não sofreram) de tal forma que o classificador é penalizado dez vezes mais por errar a classificação de uma vítima do que a de alguém que não sofreu nenhum tipo de crime. Conseguimos no set de teste uma acurácia de 54%, uma precisão de 75% e uma revocação de 9%. O mapa georreferenciado das previsões de criminalidade nas RAs do DF geradas pela regressão logística está nas Figuras 4 e 5.

Figura 4 - Mapa georreferenciado das previsões de criminalidade pelo classificador regressão logística

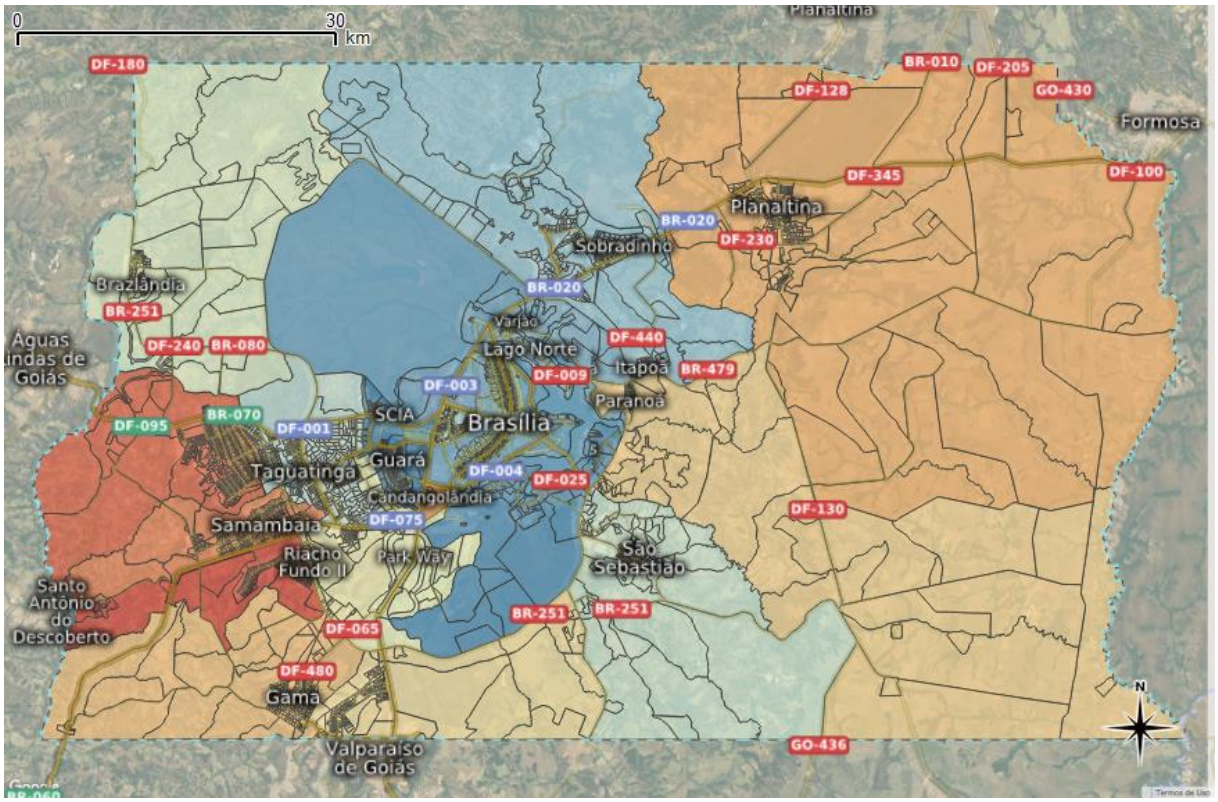


Figura 5 - Legenda do mapa georreferenciado das previsões de criminalidade pelo classificador regressão logística

4.2. Bayes Ingênuo

Para resolver o problema da base de dados desbalanceada, forçamos o classificador a considerar a probabilidade a priori de sofrer crime como sendo 50%. O resultado obtido foi de 60% de acurácia, 10% de revocação e 71% de precisão. A baixa revocação, em contraponto à elevada precisão, mostra que esse classificador priorizou a minimização de falsos positivos, o que, por outro lado, comprometeu a proporção de falsos negativos, assim como na regressão logística. A acurácia geral e a precisão foram satisfatórias. O mapa georreferenciado das previsões de criminalidade nas RAs do DF geradas pelo classificador Bayes ingênuo está nas Figuras 6 e 7.

Figura 6 - Mapa georreferenciado das previsões de criminalidade pelo classificador Bayes ingênuo

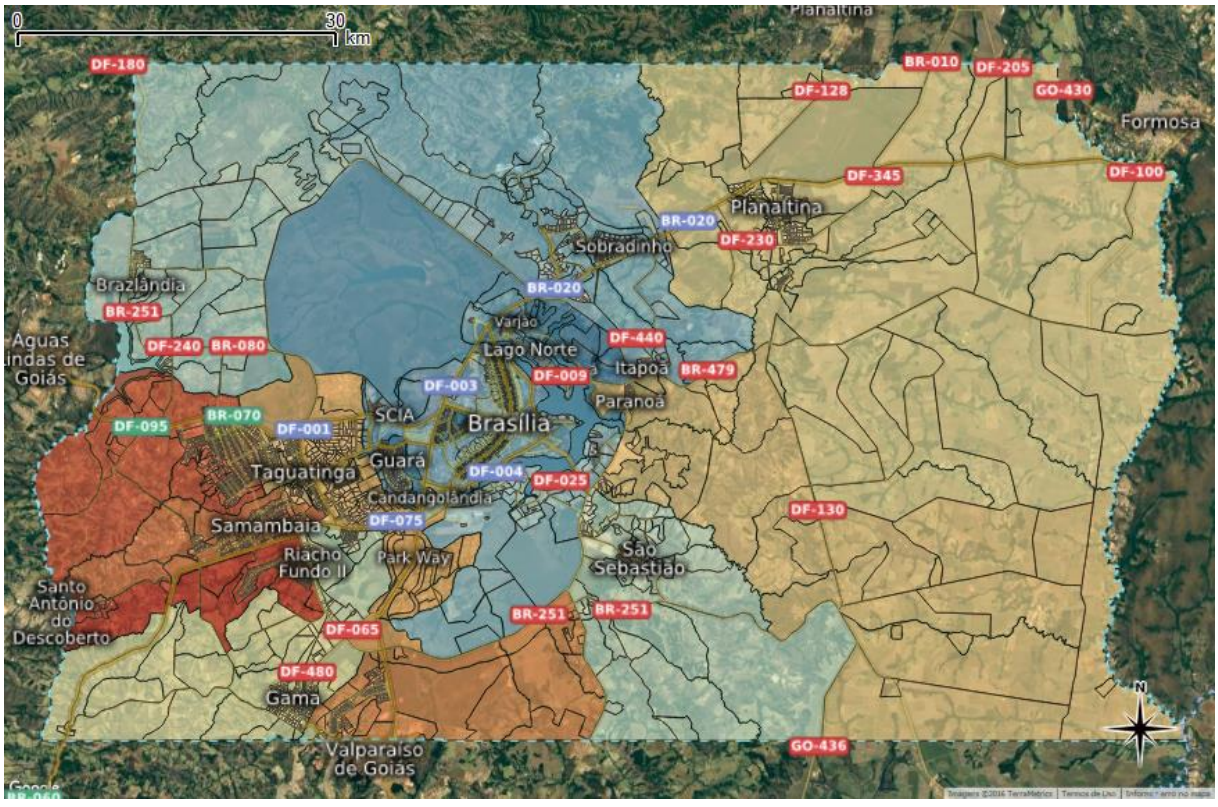
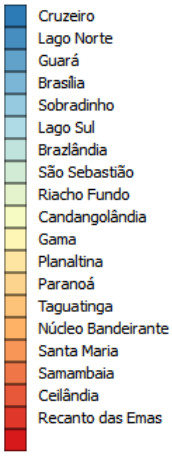


Figura 7 - Legenda do mapa georreferenciado das previsões de criminalidade pelo classificador Bayes ingênuo



4.3. SVM

O classificador SVM de margem suave retornou uma elevada acurácia de 70,23%, demonstrando a boa eficiência desse classificador. A precisão alcançada foi de 47,24%, e a revocação foi de 67,70%, evidenciando que, para os parâmetros ótimos encontrados pelo *grid search* (C=2 e $\gamma = 0.1$), a função de separação apresentou menor proporção de falsos negativos em relação a falsos positivos, ou seja, o algoritmo foi mais "parcimonioso" que o Bayes ingênuo, prevendo grande parte dos crimes que ocorreram de fato, com o custo de um número maior de "alarmes falsos". A revocação satisfatória (acima de 50%) reforça a robustez do algoritmo. O mapa georreferenciado das previsões de criminalidade nas RAs do DF geradas pelo SVM está nas Figuras 8 e 9.

Figura 8 - Mapa georreferenciado das previsões de criminalidade pelo classificador SVM

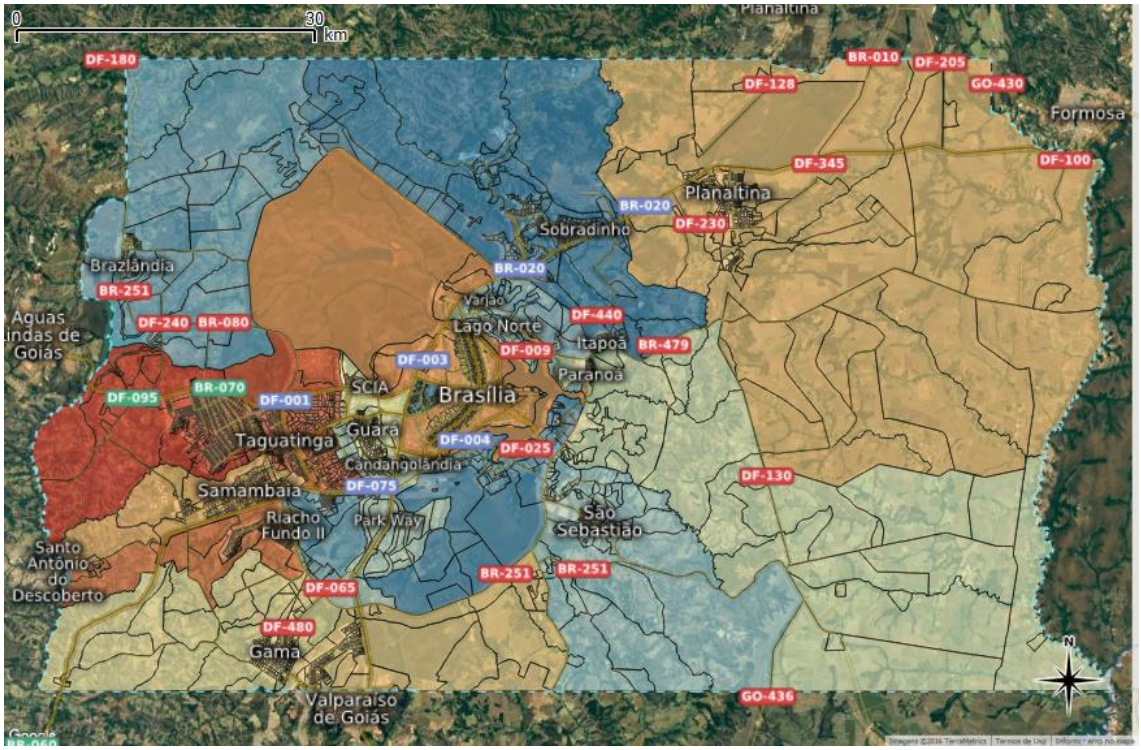


Figura 9 - Legenda do mapa georreferenciado das previsões de criminalidade pelo classificador SVM

4.4. *Random Forest*

A amostra de 60.057 observações conta com um grupo de 56.402 que não sofreu e um grupo de 3.655 que sofreu violência. O principal atributo selecionado é a TP_MOR_ESCOLARIDADE que representa o nível de escolaridade do observado, sendo 1 para analfabeto e maior que os níveis mais altos de escolaridade.

Esse parâmetro já retornou grande inferência sobre as classes, ou seja, 100% dos indivíduos que sofreram violência observados por esse estudo são analfabetos, porém 911 são analfabetos e não sofreram violência, ficando a cargo do segundo nó de separação de classes, FQ_MOR_EXTRA - categoria sobre se o entrevistado frequenta atividade extracurricular, como informática, língua, entre outros. Os resultados mostram que essa variável exerce uma influência significativa para a previsão da criminalidade na base de dados proposta, de modo que maiores níveis de atividade extracurricular induzem uma diminuição da propensão à atividade criminosa, evidenciando que uma análise mais aprofundada de elementos do âmbito sociocultural pode revelar vicissitudes adicionais para o melhor entendimento das causas de violência em grandes centros urbanos. A Figura 10 ilustra o encadeamento da árvore de decisão. O mapa georreferenciado das previsões de criminalidade nas RAs do DF geradas pelo *random forest* está nas Figuras 11 e 12.

Figura 10 - Modelo de árvore de decisão

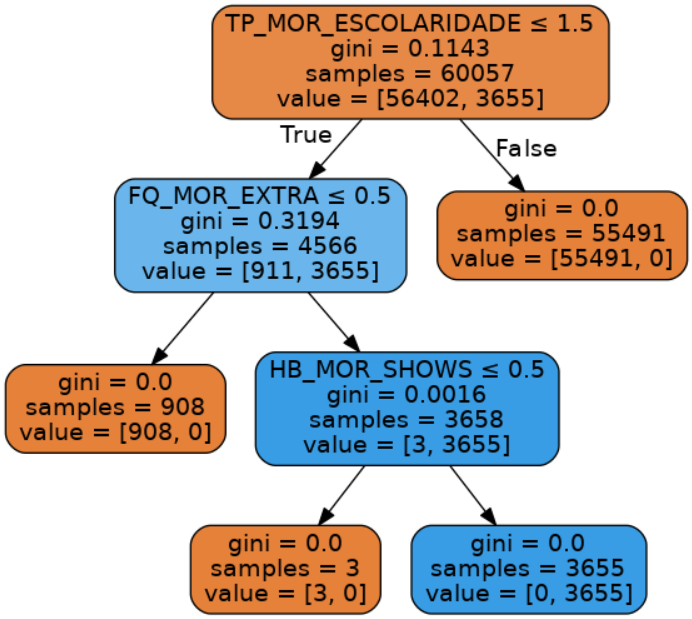


Figura 11 - Mapa georreferenciado das previsões de criminalidade pelo classificador *random forest*

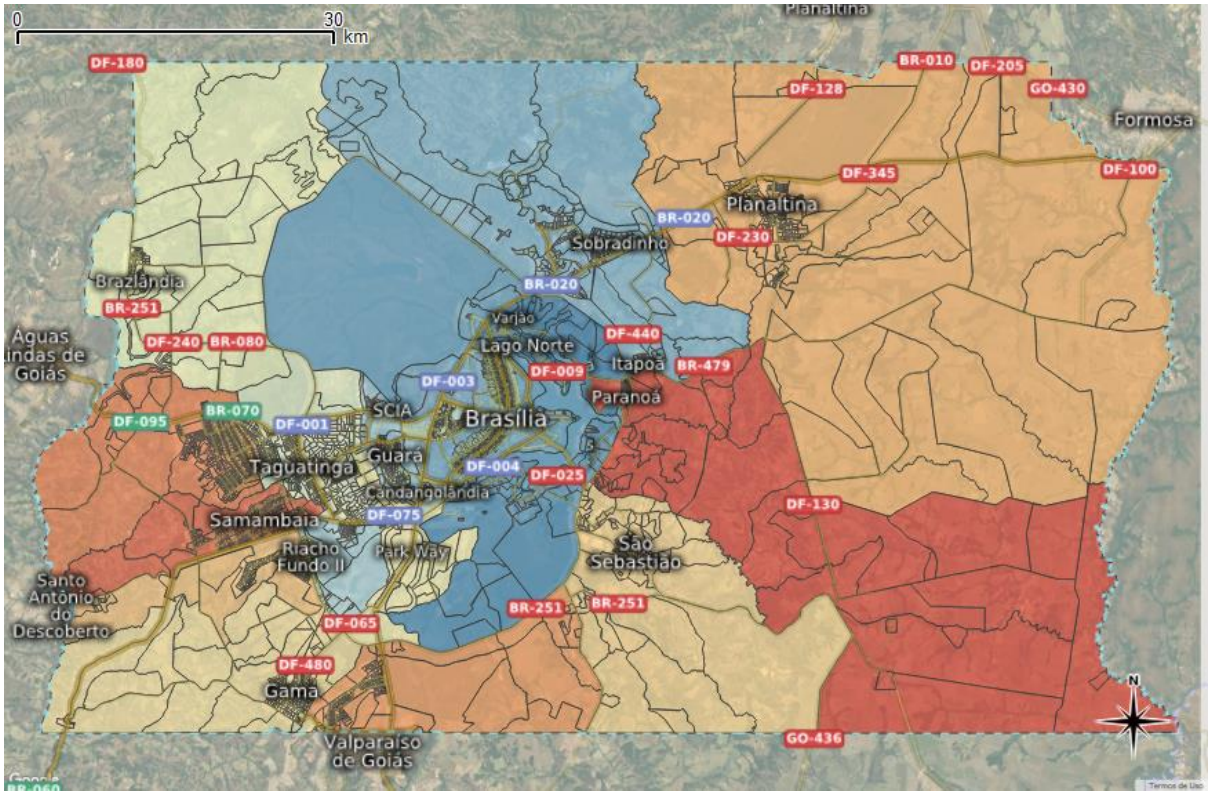
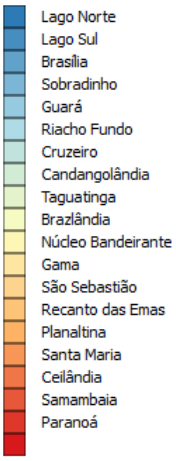


Figura 12 - Legenda do mapa georreferenciado das previsões de criminalidade pelo classificador *random forest*



5. CONCLUSÃO

Os mapas gerados pelo georreferenciamento vão ao encontro da realidade observada no Distrito Federal: zonas de alta periculosidade identificadas pelas previsões possuem grande aderência com os níveis reais de criminalidade, evidenciando o bom desempenho que metodologias de aprendizado de máquinas são capazes de lograr, identificando padrões de alta complexidade com boa acurácia.

Em geral, os resultados foram favoráveis, visto pelo desempenho preditivo das metodologias consideradas, o que apresenta evidências robustas de que as variáveis explicativas levantadas possuem forte poder explicativo para a previsão de ocorrências de atividades criminosas no Distrito Federal. Os resultados são de grande valia para subsidiar a formulação e instrumentalização de políticas públicas relativas à segurança, na medida em que as boas previsões das localidades de alta propensão para crimes podem auxiliar diretamente:

- Na intervenção *ex-post* a criminalidade, reforçando o policiamento em zonas identificadas como tendo alta propensão a crimes, bem como indicando pontos estratégicos para a alocação de contingente de patrulha e de futuras instalações físicas como delegacias e postos de vigilância comunitária;
- Na intervenção *ex-ante* a criminalidade, dada a significância satisfatória lograda pelas variáveis explicativas levantadas, o que sugere a existência de uma correlação entre o nível de criminalidade com fatores que permeiam aspectos mais profundos e complexos, indo muito além de informações socioeconômicas, as quais certamente exercem influência decisiva para a criminalidade, mas que são insuficientes para englobar a totalidade de nuanças sociais e psicológicas que perpassam a construção social de uma atividade criminosa; dimensões “latentes” como as atividades culturais e recreativas podem fornecer novos horizontes de articulação para o combate ao crime sem a necessidade de se consumar o crime de fato.

Estudos futuros são encorajados a replicar a pesquisa para demais localidades e para bases de dados mais atualizadas. Em especial, é de grande atratividade replicar o presente trabalho para uma escala temporal longitudinal, a fim de verificar a pertinência das variáveis escolhidas para captar tendências temporais – como efeitos de defasagem (isto é, identificação de variáveis que surtem efeito a curto prazo na criminalidade, como renda; ou a longo prazo, como escolaridade).

Ademais, o desempenho de cada classificador foi avaliado, e os resultados mostraram que cada classificador possui suas virtudes e debilidades. É recomendável uma extensão ao presente estudo que considere ponderações e modelos de mistura entre as metodologias apresentadas (fenômeno conhecido como *boosting*), de modo a buscar um classificador “híbrido” que englobe as vantagens de cada método individual, ou que dilua as deficiências inerentes a cada abordagem preditiva considerada.

REFERÊNCIAS BIBLIOGRÁFICAS

ABIDIN, Siti Nazifah Zainol; JAAFAR, Maheran Mohd. Surveying the best volatility measurements in stock market forecasting techniques involving small size companies in Bursa Malaysia. IEEE **Symposium on Humanities, Science and Engineering Research**, p. 975-979, 2012.

ALMEIDA, Marco Antônio Silveira De. **Análise exploratória e modelo explicativo da criminalidade no Estado de São Paulo: interação espacial** (2001). Universidade Estadual Paulista, 2007.

ARANHA, F. **Sistemas de informação geográfica: Uma arma estratégica para o Database Marketing**. v. 36, n. 2, p. 12-16, 1996.

BALAHUR, Alexandra; TURCHI, Marco. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. **Computer Speech and Language**, v. 28, n. 1, p. 56-75, 2014.
Disponível em: http://dx.doi.org/10.1016/j.csl.2013.03.004.

BECKER, Gary S. **Crime and Punishment: An Economic Approach**, v. 76.1968.

BERK, Richard A.; BLEICH, Justin. Overview of: "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment". **Criminology and Public Policy**, v. 12, n. 3, p. 511-511, 2013.

BIBAULT, Jean-Emmanuel; GIRAUD, Philippe; BURGUN, Anita. Big data and machine learning in radiation oncology: state of the art and future prospects. **Cancer Letters**, v. 382, n. 1, p. 110-117, 2016.
Disponível em: http://www.sciencedirect.com/science/article/pii/S0304383516303469.

BRENNAN, Tim; OLIVER, William L. The Emergence of Machine Learning Techniques in Criminology. **Criminology & Public Policy**, v. 12, n. 3, p. 551-562, 2013.
Disponível em: http://doi.wiley.com/10.1111/1745-9133.12055.

BRONFENBRENNER, U. **A ecologia do desenvolvimento humano: experimentos naturais e planejados**. Porto Alegre: Arte Médicas, 1979.

DORNAIKA, Fadi *et al.* Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors. **Expert Systems with Applications**, v. 58, p. 130-142, 2016.
Disponível em: http://dx.doi.org/10.1016/j.eswa.2016.03.024.

FAJNZYLBER, Pablo; LEDERMAN, Daniel; LOAYZA, Norman. **What causes violent crime?** v. 46. 2002.

FOOTE, Andrew. Decomposing the Effect of Crime on Population Changes. **Demography**, v. 52, n. 2, p. 705-728, 2015.

GLAESER, Edward L.; SACERDOTE, Bruce; SCHEINKMAN, Jose A. **Crime and social interactions**, no w5026. Cambridge, MA, 1995.

JUNIOR, A. C. **Módulo geomarketing**. NGeo-DECiv-UFSCAR, 2007.

KOTSAVASILOGLOU, C.; KOSTIKIS, N.; HRISTU-VARSAKELIS, D.; ARNAOUTOGLOU, M. Machine learning-based classification of simple drawing movements in Parkinson's disease. **Biomedical Signal Processing and Control**, v. 31, p. 174-180, 2017.
Disponível em: http://dx.doi.org/10.1016/j.bspc.2016.08.003.

LECUN, Yann; BENGIO, Yoshua; HINTON, **Geoffrey; Deep learning**. **Nature**, v. 521, n. 7553, p. 436-444, 2015.

LI, Ma; SUOHAI, Fan. **Forex prediction based on SVR optimized by artificial fish swarm algorithm**. p. 47-52, 2013.

LIN, Wei Yang; HU, Ya Han; TSAI, Chih Fong. Machine learning in financial crisis prediction: A survey. **IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews**, v. 42, n. 4, p. 421-436, 2012.

LOBO, Luiz Fernando; FERNANDEZ, José Carrera. **A criminalidade na região metropolitana de Salvador**. ANPEC - Associação Nacional dos Centros de Pósgraduação em Economia [Brazilian Association of Graduate Programs in Economics], 2003.

LOCHNER, Lance; MORETTI, Enrico. The Effect of Education on Crime : Evidence from Prison Inmates, Arrest, and Self-Reports. **American Economic Review**, v. 94, n. 1, p. 155-189, 2004.

MACEDO, Adriana C; *et al.* Violência e desigualdade social : mortalidade por homicídios e condições de vida em Salvador, Brasil. **Revista de Saúde Pública**, v. 35, n. 6, p. 515-522, 2001.
Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&amp\nnrm=iso&amp\nlng=pt&amp\ntlng=pt&amp\npid=S0034-89102001000600004\nhttp://www.scielosp.org/scielo.php?script=sci_arttext&amp.

MADEIRA, Lígia Mori; RODRIGUES, Alexandre Ben. Novas bases para as políticas públicas de segurança no Brasil a partir das práticas do governo federal no período 2003-2011. **Revista de Administração Pública**, v. 49, n. 1, p. 3-22, 2015.
Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122015000100003&lng=pt&nrm=iso&tlng=pt .

MALINA, André; CESARIO, Sebastiana. **Esporte: Fator de Integração e Inclusão Social?** Campo Grande, 2009.

MENDONÇA, Mário Jorge Cardoso De. **Um Modelo de Criminalidade para o Caso Brasileiro**, 2001.

MERTON, Robert K. Social Conformity, Deviation, and Opportunity-Structures: A Comment on the Contributions of Dubin and Cloward. **American Sociological Review**, p. 177-189, 1959.

MOHRI, M. M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of Machine Learning.: MIT press**, 2012.

OLIVEIRA, Cristiano Aguiar. Análise espacial da criminalidade no Rio Grande do Sul **Spatial analysis of criminality in Rio**. v. 34, n. 3 (ano 32), p. 35-60, 2008.

PERES, Paulo Roberto Monteiro. **A perspectiva do esporte como elemento responsável pelo afastamento de crianças e adolescentes das drogas e da criminalidade na cidade do Rio de Janeiro**, 2013.

PHAM, BINH THAI; PRADHAN, B., BUI, D. T., PRAKASH, I., e DHOLAKIA, M. B. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). **Environmental Modelling & Software**, v. 84, p. 240-250, 2016.
Disponível em: http://linkinghub.elsevier.com/retrieve/pii/S1364815216303139.

SANTOS, Marcelo Justus Dos; KASSOUF, Ana Lúcia. **Economia e criminalidade no Brasil: evidências e controvérsias empíricas**. 2008.

SERMPINIS, Georgios; *et al.* Modeling, forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms - Support vector regression forecast combinations. **European Journal of Operational Research**, v. 247, n. 3, p. 831-846, 2015. Disponível em: http://dx.doi.org/10.1016/j.ejor.2015.06.052.

SHEPARD, D.; **Database Marketing: O novo marketing direto**. Rio de Janeiro, 1993.

SHERMAN, L. W.; *et al.* **Preventing crime: What works, what doesn't, what's promising**. Washington, DC, 1998. Disponível em: http://search.ebscohost.com/login.aspx?direct=true&db=sih&AN=SM160881&lang=es&site=ehost-live.

SOMAN, K. P.; LOGANATHAN, R.; AJAY, V. **Machine Learning with SVM and Other Kernel Methods**.: PHI Learning Private Limited, 2011.

SOMAVILLA, Luana Maria. **Fatores determinantes dos latrocínios na região metropolitana de Porto Alegre: uma análise econométrica**. 2015. 0-48 f. Universidade do Vale do Rio dos Sinos - UNISINOS, 2015.

VARIAN, Hr. Big data: New tricks for econometrics. **The Journal of Economic Perspectives**, v. 28, n. June 2013, p. 1-36, 2014.

WAISELFSZ, Julio Jacobo. **Mapa da violência 2016. Flacso Brasil**., 2016. Disponível em: http://www.mapadaviolencia.org.br/pdf2016/Mapa2016_armas_web.pdf. Acesso em: 7 nov. 2016.

Revisão e Copidesque
Eliane Menezes

Editoração Eletrônica
Maurício Suda

Companhia de Planejamento
do Distrito Federal - Codeplan

Setor de Administração Municipal
SAM, Bloco H, Setores Complementares
Ed. Sede Codeplan
CEP: 70620-080 - Brasília-DF
Fone: (0xx61) 3342-2222
www.codeplan.df.gov.br
codeplan@codeplan.df.gov.br

codeplan
COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL

Secretaria de
Planejamento,
Orçamento e Gestão

Governo do Distrito Federal