

Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Danilo Augusto Cleto Souto

Análise da Probabilidade de Fuga nos Presídios do Paraná

**Curitiba
2019**

Danilo Augusto Cleto Souto

Análise da Probabilidade de Fuga nos Presídios do Paraná

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Fernando de Pol Mayer

Curitiba
2019

Análise da Probabilidade de Fuga nos Presídios do Paraná

Criação de um modelo logístico para identificar os detentos com maior risco de fuga

Danilo Augusto Cleto Souto¹

¹Especialização em Data Science & Big Data, Universidade Federal do Paraná, Curitiba, PR, Brasil*

Resumo

O Brasil possui uma das maiores populações carcerárias do mundo, e ainda em crescimento. Com presídios superlotados, em situações insalubres e precárias e ainda o baixo investimento, as fugas são cada vez mais constantes. Não há como manter todos os detentos sob vigilância constante com o atual efetivo de agentes carcerários, sem um olhar diferenciado, a ponto de impedir que tentativas de fugas ocorram. O objetivo foi propor um modelo para identificar os presos com maior risco de fuga e as características que mais exercem influência neste risco. A análise de sobrevivência foi utilizada para descobrir os fatores que mais influenciam no tempo até a fuga e a regressão logística utilizada para identificar os presos com maior risco de fuga. Foram constatados como fatores de maior influência no risco de fuga a condição do preso (com aumento no risco de fuga) e o estabelecimento penal (atuando como fator de proteção, menor risco de fuga). Também foi encontrada uma diferença significativa no risco em relação ao sexo. O modelo criado para classificação identificou 34% dos fugitivos com uma precisão de 86%.

Palavras-chave: Risco de fuga, fuga, evasão, predição, análise de sobrevivência, regressão logística

Abstract

Brazil has one of the largest prison populations in the world and still growing. With overcrowded prisons, in unsanitary and precarious situations and few investments, escapes are more and more constant. There is no way to keep all prisoners under constant surveillance with the current number of prison officers, without looking from a different point of view, to the point of preventing escape attempts. The objective was to propose a model to identify the prisoners with the highest risk of escape and the most influential risk characteristics. Survival analysis was used to find the factors that most influence the time to escape and the logistic regression used to identify the prisoners with the highest chance of escaping. The prisoner's condition (with an increase in the risk of escape) and the penal establishment (acting as a protection factor, less risk of escape) were found to be more influential in the risk of escape. A significant difference in risk was also found in relation to sex. The model created for classification identified 34% of fugitives with an accuracy of 86%.

Keywords: Escape risk, escape, evasion, predict, survival analysis, logistic regression

1. Introdução

O Brasil possui a terceira maior população carcerária do mundo com mais de 494 mil presos [11] e apresenta presídios superlotados e em péssimas condições [2]. A precariedade e insalubridade existentes no sistema com a falta de higiene contribuem para a disseminação de doenças e epidemias associadas a má alimentação, ócio e violência, cometida tanto por agentes quanto por presos.

A falta de segurança [1] torna as condições tão terríveis que os presos passam a querer escapar desse martírio de qualquer forma. Há ainda o anseio irre-

primível do ser humano pela liberdade que jamais se conforma com o estado de confinamento [1]. Assim as fugas são fatos reais que, quando reprimidas, podem levar até a outro problema ainda mais grave as rebeliões.

O pouco investimento brasileiro no sistema prisional, segundo dados do Ministério da Justiça, ocasiona uma carência de mais de 185 mil vagas. Cada preso custa cerca de 4,5 salários mínimos mensais aos cofres públicos e a previsão de crescimento anual do número de detentos está entre 10 a 12% [4].

Estima-se que 40% dos presos estejam em Delegacias Cíveis, cumprindo tanto penas provisórias quanto definitivas, devido a ausência ou insuficiência de ca-

*danilosouto@gmail.com

definitivas, devido a ausência ou insuficiência de cadeias públicas [1]. Só no Paraná estima-se que 80% dos investigadores estejam ligados à custódia de presos[3].

Como há número insuficiente de vagas em presídios e de agentes prisionais [2], além do despreparo, desvio de função e instalações inseguras, torna-se extremamente difícil evitar que tentativas fugas ocorram. Sem mencionar ainda a atuação de facções criminosas e grupos organizados.

Desta maneira, conforme exposto e dada sua condição deficitária, é impossível manter um nível alto de vigilância sobre todos os detentos contidos no sistema prisional. Então nota-se uma necessidade de uma forma diferente de abordar o problema pois, com os investimentos atuais, não há previsão de atuação para diminuir as tentativas de fugas.

Já que não é possível evitar que tentativas de fugas aconteçam, podemos ao menos criar uma forma de identificar os casos mais propensos previamente. Outra forma de atuação é identificar os motivos mais relevantes que levam às tentativas e os locais onde ocorrem, para que possam ser tomadas medidas específicas e esses casos tratados com mais atenção, seja de forma mais restritiva ou diminuir o risco ao atender as necessidades individuais.

Sendo assim, o objetivo do estudo é propor um modelo para identificar os presos com maior risco de fuga e as características que mais influenciam neste risco, com base nas informações já contidas nos sistemas existentes para que ajude os responsáveis na escolha de medidas específicas para cada caso.

Para cumprir o objetivo principal, diversas etapas devem ser executadas como: Separar entre os dados os eventos de interesse (Fuga); Identificar os fatores mais importantes relacionados aos eventos de fuga; Analisar a diferença entre a população carcerária masculina e feminina quanto ao risco de fuga.

É importante ressaltar que nem todo detento está obrigatoriamente o tempo todo dentro da prisão. Alguns cumprem um regime diferenciado (aberto ou semi-aberto) ou tem saídas autorizadas. Neste caso, quando não retornam ao presídio ou deixam de cumprir alguma obrigação legal, são considerados *evadidos*. Enquanto os que fogem são aqueles que forçam a saída pela transposição de muros ou perímetro delimitado [5]. Para simplificar serão considerados como fuga tanto os fugitivos quanto os evadidos.

2. Materiais e Métodos

2.1. Linguagem R

Como ferramenta estatística para realizar os tratamentos necessários e o ajuste dos modelos de interesse, foi utilizada a linguagem R versão 3.6.0 [12].

A linguagem R tem por base a linguagem S e foi lançada no começo dos anos 1990 sobre a licença pública GNU. Sendo *open source* qualquer pesquisador pode contribuir com pacotes ou bibliotecas complementares. Assim é uma linguagem interessante para projetos e pesquisas estatísticas [6] pois tornam muitos métodos estatísticos tanto do mais simples aos mais avançados, de leitura amigável e uso fácil.

2.2. Dos Dados

Os dados cedidos para o estudo são do sistema de gerenciamento de presos do Departamento Penitenciário do Paraná (Depen). Apesar dos dados utilizados serem reais, não permitem identificar alguém ou alguma situação em especial. Os critérios de seleção foram: O detido constar no censo do Depen e ter idade superior a 18 anos. Destes, cerca de 17% já experimentaram alguma fuga ou evasão, mesmo que pequena. Das fugas ou evasões registradas, pelo menos 18% tiveram retorno voluntário.

O evento fuga, foco do estudo, pode ser visto sob duas óticas distintas: (a) olhar somente para o evento em si, a fuga. Neste caso separa-se de maneira dicotômica (sim e não) com detentos que fugiram até a data do estudo; ou (b) relacionado ao tempo até a fuga. Assim podem ser aplicadas técnicas estatísticas de análise de sobrevivência para uma análise mais detalhada dos fatores de risco.

Nem todos os presos fugiram até a data do estudo, isso acontece pois podem ter ocorrido restrições como por exemplo o cumprimento de toda a pena ou morte; ou por motivo de tempo, onde a data de coleta das informações para o estudo tenham ocorrido antes do evento.

Segundo Borges [7], quando essas restrições ocorrem, temos o que chamamos de censura à direita. Pois não sabemos se o indivíduo irá experimentar o evento devido ao fim do estudo, em nosso caso, a data de coleta dos dados. Há também a censura independente ou aleatória, quando o indivíduo sai do estudo independentemente do evento ter ocorrido ou não [7]. São os casos de liberdade e morte.

Outro detalhe importante para se observar é que não podemos simplesmente eliminar o indivíduo que não experimentou o evento antes do fim do estudo pois eles fornecem alguma informação sobre o tempo de sobrevivência [7], o que para o nosso interesse nada mais é do que o tempo de permanência do preso na unidade de detenção.

Portanto o tempo de sobrevivência para o estudo é contado em dias a partir da data de detenção, sendo inteiro, contínuo e sempre positivo, assim pode ser medido de forma independente da data que tenha ocorrido a prisão para cada preso.

2.3. Modelos de Sobrevivência

Para funções de sobrevivência, devemos descartar os modelos estatísticos de sobrevivência que não trabalham com censura [7]. Dentre os modelos que consideram a censura temos a estimativa não paramétrica de Kaplan-Meier [7], onde podemos obter a taxa de sobrevivência com o passar do tempo ou, em nosso caso, a taxa de fuga com o passar do tempo. Outros modelos como o modelo semi-paramétrico de Cox e os paramétricos Weibull e Log-logístico que podem ser utilizados para a regressão com censura. Para nosso estudo vamos considerar o modelo de Kaplan-Meier para analisar a diferença entre os grupos ao utilizar o teste Log-Rank, e o modelo de Cox para verificar a importância das covariáveis sobre o tempo de permanência até a fuga do sistema prisional.

2.4. Modelos Lineares Generalizados

Como os presos podem ser separados dicotomicamente entre os que fugiram e os que ainda não fugiram (0 e 1) o estudo pode ser separado em duas etapas, adotando os Modelos Lineares Generalizados (MLG). Os MLG são "uma parte importante de toda pesquisa em modelagem estatística e envolve a procura de um modelo que seja o mais simples possível e que descreva bem os dados observados" [9]. Os modelos MLG foram propostos por Nelder e Wedderburn em 1972 e reúnem uma série de técnicas em uma classe de modelos de regressão constituídos principalmente de três partes: uma variável resposta ou componente aleatório; variáveis explanatórias com entrada linear; e, por último, a função de ligação, responsável por fazer a ligação entre o componente aleatório e o sistemático.

Dessa maneira, dentre os modelos pertencentes à família dos MLG utilizamos o modelo de regressão logístico, um caso especial dos Modelos Lineares Gene-

ralizados utilizados para quando a variável de resposta é binária [8], apresentando-se como uma alternativa interessante pois obtemos uma regressão linear múltipla para separar os detentos com maior risco de fuga. Onde a variável resposta varia entre 0 e 1, sendo que os valores próximos a zero indicam os presos com menor risco de fuga enquanto os próximos a 1 indicam os presos com maiores riscos.

Foram testados diversas combinações de covariáveis de forma aditiva para se obter a melhor resposta possível buscando sempre um modelo mais simples.

Para escolher o modelo foi utilizado o critério de Akaike (AIC) que "foi desenvolvido para estender o método de máxima verossimilhança para a situação de ajustes de muitos modelos com diferentes números de parâmetros" [9]. Assim temos um modelo parcimonioso com bom ajuste [9].

3. Resultados e Discussões

O trabalho na preparação do conjunto de dados analisado foi intenso, primeiro para separar as diversas variáveis que estavam vazias ou que não continham quase nenhuma informação. Após, houve a necessidade de uma atenção especial para obter somente os eventos relacionados à fuga e ou evasão dos detentos.

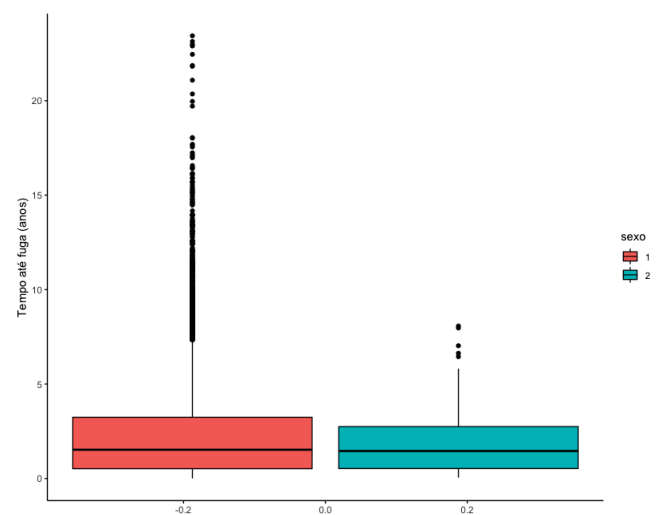


Figura 1: Boxplot do tempo da prisão até a data de fuga separados por sexo

Como pode ser observado na Figura 1, não há muita diferença no tempo da prisão até fuga quando comparado os gêneros exceto por uma cauda maior à direita do sexo 1. Ambos com mediana igual a 1 mas com média maior (em anos) para o sexo 1: 1,91 contra 1,45.

Então somente pelo tempo médio de fuga não pode-se dizer que pertencem a grupos distintos.

3.1. Análise Sobrevivência

Ao utilizar a função de sobrevivência de Kaplan-Meier estratificado por sexo, é possível notar a diferença entre os grupos, visto que o p-valor ($p < 0,0001$) muito pequeno com o teste log-rank rejeita a hipótese de igualdade entre os grupos.

Na Figura 2 nota-se um risco de fuga muito maior para o sexo 1 em relação ao 2. Além de possuir uma taxa de sobrevivência mais alta, > 90% para o grupo sexo=2, contra uma taxa de sobrevivência < 70% do sexo=1. O grupo sexo=2 tem um ritmo menos acentuado, praticamente estabilizado com poucas fugas a partir do quinto ano de pena. No grupo sexo=1, apesar do ritmo diminuir a partir do quinto ano de pena, só há uma estabilidade para os que estão presos há mais de 15 anos.

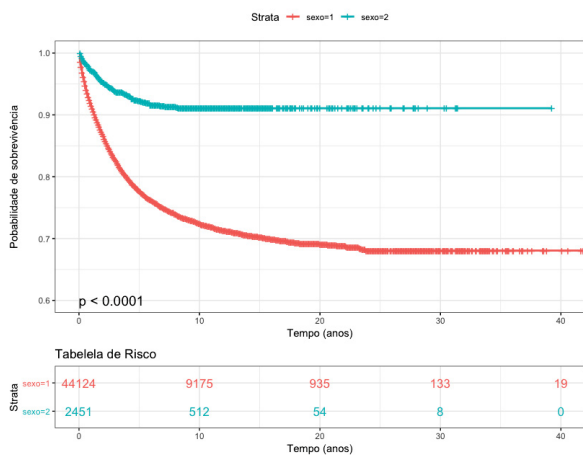


Figura 2: Curva de sobrevivência dos presos em relação ao tempo de permanência estratificado por sexo

Somente com essa análise percebe-se um risco maior de fuga para o grupo sexo=1 apontando a necessidade de uma atenção maior, ou de mais investimentos para manutenção da segurança e vigilância, ou a melhoria da situação precária nas unidades que os levam a buscar mais as fugas [1].

Como modelo de regressão utilizamos o modelo de Cox, que permite comparar o fator de risco para as diferentes covariáveis do modelo que influenciam o tempo de sobrevivência.

Com o exponencial dos coeficientes gerados pelo modelo de regressão de Cox obtém-se o fator de relevância das covariáveis quanto ao tempo de fuga. A linha vertical da Figura 3 indica o ponto neutro. Assim

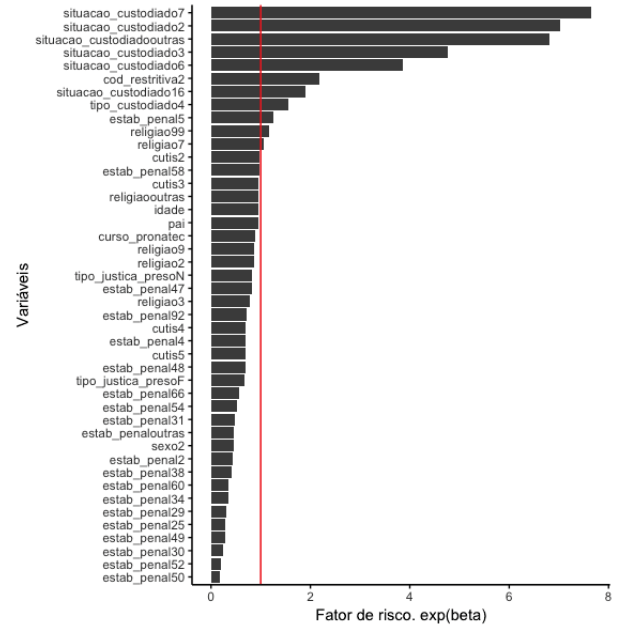


Figura 3: Exponencial dos coeficientes do modelo de regressão de Cox

as barras que se encontram a esquerda da linha vertical atuam como fator de proteção diminuindo o risco de fuga. Percebe-se que o sexo=2 e alguns estabelecimentos se encontram nessa área, indicando um risco de fuga menor. Por outro lado, as covariáveis à direita da linha vertical indicam aumento no fator de risco, aonde se evidencia algumas situações do custodiado que aumentam consideravelmente o risco de fuga.

3.2. Modelos de Regressão

Como pode ser visto na Tabela 1, o Modelo 4 apresenta menor AIC representando uma boa seleção mesmo sem utilizar tantas covariáveis quanto o modelos 1 e 2.

Modelo	GL	AIC
Modelo 1	54	37.207,24
Modelo 2	53	37.209,40
Modelo 3	52	37.207,71
Modelo 4	48	37.205,15
Modelo 5	37	37.280,75

Tabela 1: Índices AIC para classificação seleção do modelo

Para validar a qualidade do Modelo 4, foram separados aleatoriamente 30% (13.972) dos registros para validação. Mas como as classes estão desbalanceadas (78,7% não fugiram), foram escolhidos os dados de forma aleatória para que cada classe ficasse com 50% totalizando 5.950 registros. Para maximizar a acurácia

do modelo o limiar escolhido para separar os fugitivos foi calculado pelo conjunto de dados de treinamento e ficou em 0.4973162. Todos com previsão maior que esse valor foram classificados como fugitivos.

Assim podemos fazer a tabela de confusão (Tabela 2), onde há o cruzamento entre os valores preditos e observado. Nas colunas estão os dados observados (reais) e nas linhas os preditos. Na diagonal principal temos os dados em que o modelo acertou a predição, também conhecidos como Verdadeiros Negativos e Verdadeiros Positivos.

	Observado Não	Observado Fugiu
Predito Não	2.810	1.958
Predito Fugiu	165	1.017

Tabela 2: Matriz de Confusão

Com a matriz de confusão podemos obter algumas informações interessantes sobre o modelo como Acurácia, Sensibilidade e Especificidade.

A Acurácia é definida como a soma dos valores (positivos ou negativos, classificados corretamente dividida pelo total de registros. Como a acurácia resume muitos números em um só, são necessárias outras medidas para interpretar a performance de um modelo [10]. Para o modelo escolhido, a acurácia calculada com uma base de testes balanceada, foi de 64,32%. Ela é a razão entre os valores verdadeiros preditos dividido pelo total de registros. É uma medida importante pois indica quantos registros foram preditos conforme observados.

Outra medida mais informativas é a Sensibilidade, também conhecida como *Recall*, que neste modelo alcançou 34,18%. Ela nos informa a quantidade de valores positivos verdadeiros encontrados no estudo (os fugitivos). Ou seja, do total de fugitivos reais, o modelo indicou cerca de 34%.

Mas a sensibilidade não diz tudo, outra medida importante é a Especificidade. É a mesma informação mas para valores negativos [10] e com os 94,45% alcançado pelo modelo, indica que foram classificados corretamente quase todos os valores negativos.

A precisão é a razão dos positivos verdadeiros classificados corretamente pelo número de positivos encontrados [10]. Assim, para o estudo, essa medida passa a ser importante pois informa a taxa de acerto/erro dos fugitivos. Ou seja, quanto maior for esse valor, maior a certeza de indivíduos irão realmente fugir. A precisão foi de 86,04%.

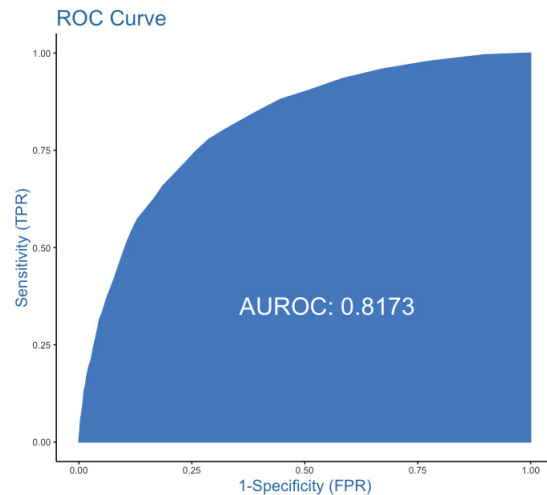


Figura 4: Curva ROC

Na Figura 4 vemos a curva ROC do modelo gerado. "A diagonal representa exatamente a probabilidade, então qualquer coisa acima da diagonal é melhor que a probabilidade, sendo que quanto mais distante da linha, melhor." [10]. A área da curva ROC foi de 0,817 bem acima da probabilidade de detecção simplesmente ao acaso.

4. Conclusões

O resultado da análise de sobrevivência indica que as questões levantadas pelos pesquisadores sobre a situação dos custodiados e a importância da qualidade dos presídios aparentam ter grande influência na função de sobrevivência do tempo até a fuga dos presos, pois foram identificadas no modelo de Cox (Fig. 3) como as mais influentes.

Outro fator demonstrado no estudo é que apesar de não haver grande diferença do tempo até a fuga entre os grupos de estudos, como mostrado pela gráfico de caixas (Fig. 1), há sim uma diferença significativa quanto ao risco de fuga entre os gêneros, comprovado pelo teste de log-rank, vide figura 2. Portanto o risco de fuga é maior quando estamos nos referindo ao sexo=1, indicando a necessidade de maior atenção por parte das autoridades.

E por fim, o modelo logístico (modelo 4) criado para identificar individualmente os casos com maior risco de fuga, obteve sensibilidade relativamente baixa (34,18%) e precisão de 86%. Isto implica que, apesar de identificar somente cerca de um terço dos presos, tem-se um índice alto de certeza daqueles que foram classificados como possíveis fugitivos irão realmente fugir.

Logo o modelo logístico (modelo 4), pode sim ser utilizado como ferramenta para auxílio na identificação de casos com alto risco de fuga.

5. Trabalhos Futuros

O estudo possibilita análises complementares da situação dos presídios com maior risco fuga. Já que tanto a situação do custodiado como a situação do estabelecimento penal têm grande influência nas tentativas de fuga.

Também é possível comparar a performance do modelo logístico criado para classificação do risco de fuga com outras técnicas de classificação, como por exemplo os algoritmos de Aprendizado de Máquina.

Um complemento para o estudo seria a continuação de uma segunda parte após a identificação dos possíveis fugitivos. Há a possibilidade de utilizar outro modelo para fazer a regressão do tempo até a fuga. Uma opção seria considerar o Modelo Gamma já que a variável tempo não tem um comportamento normal padrão e possui uma cauda muito longa à direita.

6. Agradecimentos

Gostaria de agradecer a Secretaria Segurança Pública e o Maj. Rodrigo Perim de Lima pelo apoio no estudo, a Manoel Fávio Leal pelo intermédio, a Mário Sanchez Rodrigues Filho pela ajuda na obtenção das informações e a Everton Osnei Cesário pela ajuda na revisão do trabalho.

Referências

- [1] Assis, Rafael Damaceno de. *A realidade do atual sistema penitenciário brasileiro*, (Direito Net, 2007). Disponível em: <<https://www.direitonet.com.br/artigos/exibir/3481/A-realidade-atual-do-sistema-penitenciario-brasileiro>> Acesso em: 05 mai. 2019.
- [2] Salla, Fernando. *As rebeliões nas prisões: novos significados a partir da experiência brasileira*. 2006.
- [3] Folha de Londrina. *PR registra 12 fugas em 2 meses*. Disponível em: <<https://www.folhadelondrina.com.br/reportagem/sistema-carcerario-pr-registra-12-fugas-em-dois-meses-840562.html>>. Acesso em: 05 maio 2019.
- [4] Dullius, Aladio Anastacio; Hartmann, Jackson André Müller. *Análise do sistema prisional brasileiro*. (Revista Âmbito Jurídico). Disponível em: <http://ambito-juridico.com.br/site/?artigo_id=10878&n_link=revista_artigos_leitura>. Acesso em: 05 de maio 2019.
- [5] Departamento Penitenciário do Paraná. Disponível em: <<http://www.depen.pr.gov.br/modules/conteudo/conteudo.php?conteudo=147>>. Acesso em: 05 maio 2019.
- [6] Davies, Tilman M. *The Book Of R*. (No Starch Press, San Francisco, 2016).
- [7] Borges, Alexandra Isabel Monteiro *Análise de Sobrevida com o R*. (2014).
- [8] Bradley, Efron; Hastie, Trevor *Computer Age Statical Inference. Algorithms, Evidence, and Data Science*. (Cambridge University Press, New York, 2016).
- [9] Cordeiro, Gaus Moutinho; Demétrio, Glárisse G.B. *Modelos Lineares Generealizados e Extensões*. (2008).
- [10] Marsland, Stephen. *Machine Learning. An Algorithmic Perspective*. (CRC Press, 2014).
- [11] Instituto Brasileiro de Ciências Criminais. (Fugas nos presídios) Disponível em: <<https://ibccrim.jusbrasil.com.br/noticias/2515229/fugas-nos-presidios>>. Acesso em: 05 maio 2019.
- [12] R Core Team. *textitR: A language and environment for statistical computing*. R Foundation for Statistical Computing (Vienna, Austria, 2019). Disponível em: <<https://www.R-project.org/>>