UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE LORENA

LEANDRO KELLERMANN DE OLIVEIRA

Análise de um modelo preditivo de aprendizagem de máquina criado com dados de boletins de ocorrência registrados na Grande São Paulo entre 2007 e 2014.

Lorena
2020

LEANDRO KELLERMANN DE OLIVEIRA

Análise de um modelo preditivo de aprendizagem de máquina criado com dados de boletins de ocorrência registrados na Grande São Paulo entre 2007 e 2014.

Trabalho apresentado à Escola de Engenharia de Lorena da Universidade de São Paulo como parte da disciplina Trabalho de Graduação.

Orientador: Fabiano F. Bargos

Lorena

2020

Ficha catalográfica elaborada pelo Sistema Automatizado
da Escola de Engenharia de Lorena,
com os dados fornecidos pelo(a) autor(a)

Oliveira, Leandro Kellermann de
Análise de um modelo preditivo de aprendizagem de máquina criado com dados de boletins de ocorrência registrados na Grande São Paulo entre 2007 e 2014 / Leandro Kellermann de Oliveira; orientador Fabiano Fernandes Bargos. - Lorena, 2020.
95 p.

Monografia apresentada como requisito parcial para a conclusão de Graduação do Curso de Engenharia Física - Escola de Engenharia de Lorena da Universidade de São Paulo. 2020

1. Aprendizagem de máquina. 2. Inteligência artificial. 3. Análise preditiva. 4. Floresta aleatória. I. Título. II. Bargos, Fabiano Fernandes, orient.

AGRADECIMENTOS

Eu acho muito difícil escolher uma ou outra pessoa para citar aqui nos agradecimentos por que eu não tenho capacidade de julgar o grau de importância que as pessoas tiveram na minha vida. Eu fui presenteado com momentos na companhia de pessoas e não-pessoas que me trouxeram até aqui, que me levarão até algum outro lugar. Pra mim, tudo o que aconteceu até hoje foi importante, mas eu não vou lembrar de todos e com certeza, se lembrasse, esse trecho do trabalho seria o maior deles. Então, gostaria de expor essa limitação da minha memória e pedir a compreensão a todos que não cito aqui, mas saiba que eu sempre reconhecerei a importância de todo gesto já direcionado a mim e como isso me levou até aqui. Mas vamos lá: aos agradecimentos.

Tem uma pessoa que não teria como não citar: a dona Tereza, vulgo minha mãe. Eu tive a sorte de tê-la ao meu lado sempre que precisei, por sempre ter acreditado em mim, por perdoar as minhas falhas e comemorar comigo os meus acertos. Sem o apoio dela, a minha "ideia mirabolante"de estudar para trabalhar nunca teria saído do papel. Pensar "ou trabalha, ou estuda"é uma ideia comum em famílias pobres, e a ideia toma mais força ainda quando todas as pessoas ao seu redor pensam o mesmo.

Outra pessoa que gostaria de agradecer é a um velho amigo de infância, Ruan, e a toda família dele. Os pais dele incentivavam ele desde pequeno a estudar algo pra "ser alguém na vida". Um dia o Ruan de 14 anos, por incentivo dos pais, resolveu ir fazer uma prova para o curso de mecânico de usinagem no SENAI-SP e botou a mesma ideia na minha cabeça de 14 anos. Um ano depois dessa ideia dele, e que ele seguiu, segui também. Uns anos depois, quando estava por decidir faculdade do que eu ia fazer, lembrei de uma conversa que tive com um grupo de colegas do SENAI antes de ir para a aula. Aí lembrei de um dos meus colegas, Djavan, falando do curso de engenharia física. Sou grato a ele por ter falado do curso naquele dia.

Lá no SENAI eu tive a oportunidade de estudar, conseguir meu primeiro emprego registrado, e um dia pensar em estudar em uma faculdade. Eu pensei nisso no fim de 2012, quando terminei meu curso técnico, iniciado depois do curso de mecânico de usinagem. A ideia era fazer um tecnólogo e continuar no emprego que estava graças ao curso técnico, mas na empresa em que eu trabalhava existiam pessoas como Océlio, Vinícius, Carlos, Patrícia, Everton, Adriano, Sandro, Daniel, Paulo... e muitas outras que eu até esqueci o nome. Ter trabalhado com eles foi importante por que eu aprendi, com 19 anos, que dá pra fazer um trabalho sério e ser bem-humorado e desde então eu tenho eles como exemplo de profissionais e pessoa. Mas o que me faz lembrar deles agora é que eles colocaram na minha cabeça de fazer um bacharelado, eles disseram que eu tinha cabeça pra fazer um curso mais "puxado". O Vinícius foi quem me falou que dava pra fazer um curso integral na USP, mesmo sem trabalhar, graças aos programas de apoio à permanência da

universidade. Além disso, quando o expediente acabava ele me dava carona até o Tietê quando eu ia pro cursinho e ele ia para casa. O dinheiro que eu economizava de passagem, ia pro cursinho.

Tive que sair do emprego em 2014 quando passei no vestibular e isso ia pesar na renda de casa. Eu tinha dinheiro só para viver uns três meses fora de casa na época e não sabia se ia dar certo. As notas baixas em quase todas as primeiras provas, uma ex-gerente da empresa em que trabalhei me chamando para trabalhar em outro lugar e a incerteza sobre conseguir um meio de me manter na faculdade nos próximos anos me tentaram a sair. Mas aqui na EEL tinha o prof. Shigue. No dia que eu ia desistir do curso de engenharia física, ele me entregou um Arduíno e disse para eu estudar aquilo, por que ele ia preencher uma bolsa do programa de Cultura e Extensão e ia me selecionar para aquilo. Eu não entendi foi nada, mas aquilo era uma chance de eu continuar na faculdade. Acabou que ali foi um primeiro passo que eu dei em direção à eletrônica e programação. Isso me ajudou depois a participar de outro projeto de pesquisa em parceria com o prof. Fernando Vernilli, onde aprendi mais ainda como estudar sozinho, pude desenvolver experimentos e que também me ajudou a me manter na faculdade.

Outros professores que só o jeito deles darem aula me estimularam a estudar foram a Bertha, o Fabiano e a Mariana. Quando a professora Bertha falava sobre mecânica ondulatória, eletromagnetismo e relatividade nas aulas de Física IV, eu queria que ela nunca parasse. Eu queria aprender mais daquilo. A "paranóia"foi tão grande que depois de uma aula dela eu decidi que ia dar um jeito de viver mais só pra estudar mais daquelas coisas que ela ensinava, nisso passei a viver um estilo de vida mais saudável e emagreci 40 kg. Com o Fabiano, eu lembro dele ter projetado um código durante a aula que eu achei bonito: compacto e cheio de vetor. Então desde essa aula passei a querer fazer códigos sempre desse jeito. A professora Mariana, por sua vez, me deixava indignado durante as aulas de Estatística quando ela contava alguma coisa muito contraintuitiva que tinha base estatística. Uma delas foi o problema de Monty Hall, que ela tentou explicar durante a aula e eu não entendi. Quando percebi que pra entender o problema a gente tinha que pensar na abertura da porta como uma informação e não como um evento, tomei gosto por querer descobrir como a estatística poderia medir as coisas.

Fora da sala de aula, também recebi ajuda de alguns colegas - que viraram amigos. A faculdade não foi fácil, mas foi bem mais fácil com a Amanda, Humberto, Paulo e Vinícius. Sou grato ao Humberto aguentar as brincadeiras que fazemos com ele e por não aparecermos em uma matéria do Cidade Alerta. Sou grato à Amanda pelos bolos e outros doces que eventualmente ela dividia com a gente. Ao Paulo eu só agradeço por ter me dado a ideia de um dia ir trabalhar em uma consultoria e desagradeço profundamente pela qualidade dos filmes sugeridos a nós. Ao Vinícius eu sou grato pelos *insights* para

inventar uma piada nova e por estimular todo o grupo a seguir ideias fadadas ao fracasso, mas que no fim iam render boas risadas.

Eu queria agradecer muitas outras pessoas que compartilharam comigo momentos de leveza. Muitas delas eu conheci em filas de shows, ou caminho deles; outras eu nem conheci e elas nem me conhecem, mas o trabalho artístico feito por elas tornam a minha vida e a de outros mais fácil. Outras ainda conseguiam me alegrar só com a presença e com conversas despretensiosas que um dia acabaram definitivamente. E ainda tem aquelas que eu nem sei que me ajudaram, mas o fizeram mesmo assim, conscientes ou não.

"Não existe algo como o 'self-made man'. Nós somos feitos de milhares de outros."
George Matthew Adams (1878-1962)

RESUMO

OLIVEIRA, L. K. **Análise de um modelo preditivo de aprendizagem de máquina criado com dados de boletins de ocorrência registrados na Grande São Paulo entre 2007 e 2014.**. 2020. Número de páginas 95p. Monografia (Trabalho de Graduação) - Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena, 2019.

A inteligência artificial (IA) é um conjunto de tecnologias cujo objetivo é objetivo criar mecanismos que permitem computadores mimetizar a inteligência humana e no centro dessa tecnologia temos a aprendizagem de máquina ou *machine learning* (ML). Neste trabalho apresentamos os conceitos básicos por trás do algoritmo de floresta aleatória para um modelo preditivo de classificação e avaliaremos este modelo por diferentes métricas, esperando que o modelo acerte pelo menos 50% das previsões e que as métricas de avaliação apresentem coerência entre si. O modelo será construído em um computador pessoal, utilizando recursos como os *softwares* R, Microsoft Excel e SQL Server. Os dados utilizados foram extraídos de boletins de ocorrência registrados entre 2007 e 2014 na Grande São Paulo e o modelo tentará prever qual tipo de crime uma pessoa é mais propensa a sofrer, considerando sua cor de pele, sexo, faixa etária, localização no momento do crime, bem como período do dia e dia da semana e uma outra variável, não identificada no dicionário de dados. O modelo construído acertou $69,74\%$, com erro *OOB* de $30,26\%$, das previsões nos valores de teste e um índice Kappa de Fleiss igual a $0,551$, o que significa que o modelo realiza previsões de concordância moderada com os resultados reais. O modelo também mostrou que, no geral, o decrescimento médio de Gini foi maior para a variável que representa o local em que a pessoa está, indicando que esta é a variável mais importante para o modelo. Contudo, notou-se a partir da matriz de confusão que o modelo apresentou-se enviesado em favor dos valores mais frequentes da variável alvo. É sugerido, por fim, que o modelo seja melhorado tratando os dados de treino a fim de compensar os valores da variável alvo que são menos frequentes no conjunto em questão. Isso pode ser alcançado através de técnicas computacionais tais como a sobreamostragem de dados de treino.

Palavras-chave: aprendizagem de máquina, inteligência artificial, análise preditiva, floresta aleatória.

ABSTRACT

OLIVEIRA, L. **Analysis of a predictive machine learning model built with police reports data recorded in Greater São Paulo between 2007 and 2014**. 2020. Number of pages 95 p. Monograph (Bachelor Thesis) - Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena, 2019.

Artificial inteligence (AI) it's an interdisciplinary field that aims build mechanisms that allow computers mimics the human inteligence and machine learning (ML) it is on its core. In this work we show the random forest algorithm concept to build a predictive classification model and we will evaluate this model by different metrics, and hoping our model to get 50% of test values right and the values of all metrics to be coherent among them. The model was built in a personal computer using softwares R, Microsoft Excel and SQL Server. The data to train our model was get from police reports collected in Great São Paulo between 2007 and 2014 and the model will try to predict the kind of crime someone can suffer based on its color skin, sex, age range, type of local the person was, day of the week and period of the day and other variable that was not described in our data dictionary. The model predicted $69,74\%$ right from the test data, with $30,26\%$ of OOB error. The Kappa Fleiss of our model is $0,551$, which means that the model agrees moderately with the real values. The mean decrease Gini was greater to the variable that represents the local type where the crime happened, and this indicates that this is the most important variable to the model. However, we noticed by analysing the confusion matrix that our model was biased because some values of our target variable was much more frequent in our data than others. At the end we can conclude that the model can be improved if we deal with the imbalanced data by computational techniques such as oversampling the less frequent values of our target variable in training data.

Keywords: machine learning, artificial inteligence, predictive analysis, random forest.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

A inteligência artificial (*IA*) é uma constelação de tecnologias que envolve diversas áreas do conhecimento visando criar mecanismos de mimetização da inteligência humana por meio de programas de computador que se ajustam a um conjunto de dados ou instruções. No núcleo da *IA* está o que chamamos de aprendizagem de máquina, que é um conjunto de conhecimento e tecnologias que permitem algoritmos aprenderem com os dados sem serem programados explicitamente. Esse tipo de tecnologia começou a dar seus primeiros passos já nos anos de 1950, mas só foi ganhar bastante popularidade nos últimos 20 anos devido ao *big data*. Desde então, diversos algoritmos de *IA* vêm sendo usados como ferramenta de negócios no mercado ou de suporte em diversas áreas da ciência, seja ela exata ou não (KANIOURA; EITEL-PORTER, 2020; O'LEARY, 2013).

O primeiro caso notável de aplicação de ciência da computação foi a construção das máquinas do tipo *Enigma*, cuja primeira fora patenteada pelo engenheiro eletrotécnico alemão Arthur Scherbius por volta de 1918 (University of Klagenfurt, 2020). As máquinas desse tipo faziam uso de um criativo circuito eletromecânico para encriptar mensagens trocadas entre diversos tipos de organização, inclusive as forças armadas de um país. Talvez as máquinas mais emblemáticas desta família tenham sido as *Bombe*, do grupo *Hut 8*, formado pelo governo inglês e inicialmente liderado pelo matemático briitânico Alan Mathison Turing. Na ocasião, essas máquinas foram usadas para decifrar as mensagens criptografadas trocadas entre as forças armadas da Alemanha Nazista na década de 1950 (MACHETTI, 2016). Mais tarde, Turing viera publicar o seu artigo *Computating Machinary and Intelligence*, onde os primeiros conceitos, definições e limitações da IA (sem ter esse nome ainda) começaram a se formar. Alguns anos depois teríamos o primeiro programa de *IA*.

Entre 1955 e 1956, Allen Newell, Cliff Shaw e Herbert Simon desenvolveram um programa de computador chamado *Logic Theorist*, considerado o primeiro programa "especialmente engenheirado para mimetizar as habilidades humanas de resolução de problemas". Esse programa foi apresentado a primeira vez na conferência *Dartmout Summer Research Project on Artificial Inteligence*, organizada por pessoas renomadas no campo da ciência da computação, tais como John McCarthy e Marvy Minsky. O objetivo da conferência era juntar estudiosos de diversos campos a fim de formalizar conceitos sobre o que era IA (o termo fora cunhado a partir daquela conferência), contudo, essas expectativas foram frustradas (ANYOHA, 2017), mas serviu para impulsionar a comunidade científica discutir sobre o assunto.

Com a melhoria dos computadores nos entre as décadas de 1960 e 1974, pesquisadores como Herbert Simon e Allen Newell desenvolveram trabalhos teóricos a fim de criar programas como o *General Problem Solver* e até mesmo sistemas de reconhecimento

de texto, como o programa *ELIZA* de Joseph Weizenbaum. Tais programas mostravam à comunidade o potencial da nova tecnologia e também serviam como argumentos para convencer outros estudiosos e agências de governo a investir no seu desenvolvimento, apesar de muito o que se falava na época se tratar apenas de provas de conceito estarem um tanto longe de desenvolvimento de aplicações para resolução de problemas reais (ANYOHA, 2017). O problema prático foi solucionado décadas depois, na era do *big data*.

O termo *big data* refere-se ao conhecimento e tecnologias utilizados para coletar, processar e analisar da forma mais rápida possível dados volumosos gerados nas mais variadas formas (Amazon, 2020). Os dados podem apresentar de forma estruturada como tabelas, de forma semiestruturada como o código *HTML* de uma página na *internet*, ou mesmo de uma forma não-estruturada, como um arquivo de áudio, imagem ou postagem em uma rede social. A captação e processamento desses tipos de dados, quando apresentado em grandes volumes, só tornou-se possível por volta de 2006, com o desenvolvimento do ecossistema *Hadoop*, que é um programa de código aberto utilizado para automatizar o armazenamento e processamento dos grandes e variados volumes de dados característicos do *big data*. Em termos práticos, o *Hadoop* permite que uma rede de computadores (ou *cluster*) atue como se fosse um computador só, distribuindo o esforço computacional entre as máquinas físicas, bem como a criação de diversas máquinas virtuais nas mais variadas configurações, desde que estas não superem a capacidade física coletiva dos computadores desta rede. Hoje existem muitos serviços de computação em nuvem que fazem o uso dessa tecnologia de diferentes formas, seja para armazenamento de arquivos de vídeo para *streaming* ou para o modelamento de algoritmos de aprendizagem de máquina, como faz a *Netflix* (GIORDANO, 2019; Oracle, 2020).

A *IA* ainda pode ser utilizada para criar ferramentas de classificação de microestrutura de aços (AZIMI et al., 2018), para compreender melhor a opinião de um grupo de indivíduos em um contexto social (MASON; VAUGHAN; WALLACH, 2013) ou para prever a tendência de espalhamento de um vírus durante uma epidemia (LI et al., 2020). A aplicação de inteligência artificial neste trabalho se encaixaria mais próximo do uso da ferramenta num contexto de ciências sociais.

O objetivo deste trabalho é avaliar um modelo de aprendizagem de máquina que classifica o tipo de crime que um indivíduo pode sofrer em função de sua idade, sexo, o tipo de local onde estava no momento do crime e o respectivo período do dia, a sua faixa etária, profissão, grau de instrução e uma outra variável denominada *CONT_PESSOA*, cuja descrição não estava presente no dicionário de dados da fonte.

O modelo, sujeito à limitações de um computador pessoal, será construído a partir de dados de boletins de ocorrência registrados entre os anos de 2007 e 2014 sobre crimes ocorridos na Grande São Paulo. Para atingir esse objetivo, também utilizaremos

recursos computacionais como *Microsoft Excel*, sistema de gerenciamento de banco de dados *Microsoft SQL Server*, o programa *R* com as bibliotecas *randomForest, irr, RODBC, dplyr* e *caret*.

É esperado que o modelo criado seja capaz de prever pelo menos 50% das ocorrências e que a interpretação dos resultados das métricas utilizadas nesta avaliação sejam coerentes entre si. As métricas utilizadas serão precisão, erro *out of bag*. índice Kappa de Fleiss e erro de classificação da matriz de confusão.

2 REVISÃO BIBLIOGRÁFICA

2.1 Aprendizagem de máquina ou *machine learning*

Aprendizagem de máquina ou *machine learning* é uma subárea de um campo de estudos mais abrangentes, a inteligência artificial (IA). O conceito de IA contém a ideia de que um sistema inteligente é capaz de reconhecer um aspecto específico] através da manipulação de símbolos e classificá-los (GARBADE, 2018) e pode ser realizado através de programação específica, como uma sequência de instruções "*if-then-else*". Já em um algoritmo de aprendizagem de máquina, ao invés de termos uma sequência "*if-then-else*", há uma função matemática de aprendizagem que é definida de acordo com os dados apresentados ao modelo (BROWNLEE, 2016a; KANIOURA; EITEL-PORTER, 2020):

$$Y_i = f(x_i) \tag{2.1}$$

No exemplo acima, Y_i representa o valor de uma variável resposta para uma instância i, ou seja, o que desejamos descobrir, dado o valor de uma variável preditora x_i correspondente. Em sua essência, um algoritmo de aprendizagem de máquina supervisionado busca uma função f dentro de uma família de funções F que minimiza uma função de custo, ou função de perda, L, de acordo com os as variáveis preditoras e respostas fornecidas. A função de custo L, por sua vez, muda de acordo com o algoritmo (WEINBERGER, 2017; BROWNLEE, 2016a) .

A maneira como o algoritmo faz esse mapeamento define se ele é paramétrico ou não-paramétrico. Algoritmos paramétricos resultam em modelos mais simples, pois assume-se que a forma da função já é conhecida e só precisamos descobrir os coeficientes da função otimizando a função de custo. Algoritmos não-paramétricos, por sua vez, não fazem suposição nenhuma sobre a forma da função (BROWNLEE, 2016a) .

Saber se o modelo é paramétrico ou não define se há a necessidade de ocorrer normalização dos dados. Quando fazemos alguma suposição sobre a forma da função resposta, deve-se atentar com a forma da distribuição dos dados inseridos no algoritmo. Já quando o algoritmo é não-paramétrico, ou estatístico, não há essa necessidade (RUSSELL, 2010). Dos modelos paramétricos, podemos citar a regressão logística e as redes neurais simplificadas. Já os não paramétricos podem ser citados a árvore de decisão e os k-vizinhos mais próximos (kNN) (BROWNLEE, 2016a) .

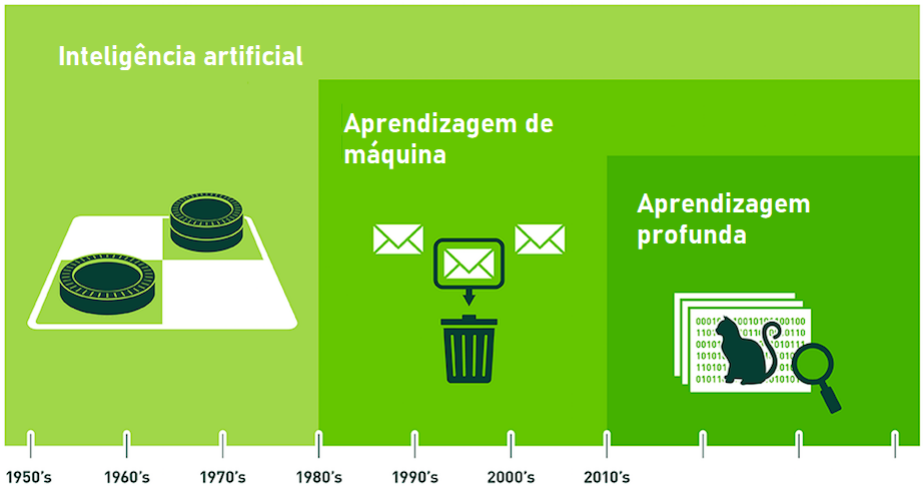
Outra forma de classificar um algoritmo de aprendizagem de máquina é quanto ao significado que se dá às variáveis quando inseridas no modelo. Nesse aspecto podemos dizer que um algoritmo é supervisionado ou não-supervisionado. Quando o algoritmo é supervisionado, inicia-se a modelagem já assumindo quais são as variáveis-alvo do modelo

e quais são as variáveis explicativas. Por outro lado, num modelo não-supervisionado, nada é assumido e o próprio algoritmo se encarrega de descobrir os padrões sozinho (BRUCE; BRUCE, 2019; BROWNLEE, 2016b). Exemplos de algoritmos de aprendizagem supervisionada são os de regressão logística, árvore de decisão e floresta aleatória. Já um exemplo de algoritmo não-supervisionado é o kNN.

Ainda podemos dizer se um algoritmo é de regressão ou de classificação de acordo com o tipo de variável alvo. Se a variável alvo apresenta valores contínuos dizemos que o algoritmo é de regressão, caso os valores sejam discretos ou categórico-nominais, dizemos que o algoritmo é de classificação. Contudo, ainda é possível que um algoritmo seja capaz de prever os dois tipos de variável alvo, como é o caso da floresta aleatória (BRUCE; BRUCE, 2019).

Existem outras maneiras de classificar os modelos de aprendizagem de máquina, como pela forma que o algoritmo aprende (instância ou modelo) e pelo tipo de processamento (aprendizagem por incremento ou lote) (GÉRON, 2019) e também pela forma de estruturação dos dados de treino. Nesse último caso, quando o algoritmo pode receber dados não-estruturados, ou seja, que não possuem formato de tabela, dizemos que o algorítmo é de aprendizagem profunda, uma sub-área de aprendizagem de máquina (KAPOOR, 2019).

Figura 1 – Um panorama geral da relação entre inteligência artificial, aprendizagem de máquina e aprendizagem profunda.



Fonte: adaptado de (Data Science Brigade, 2016).

Neste trabalho, usaremos um algoritmo de aprendizagem por modelo supervisionado e não-paramétrico chamado de floresta aleatória ou *random forest* para um problema de classificação.

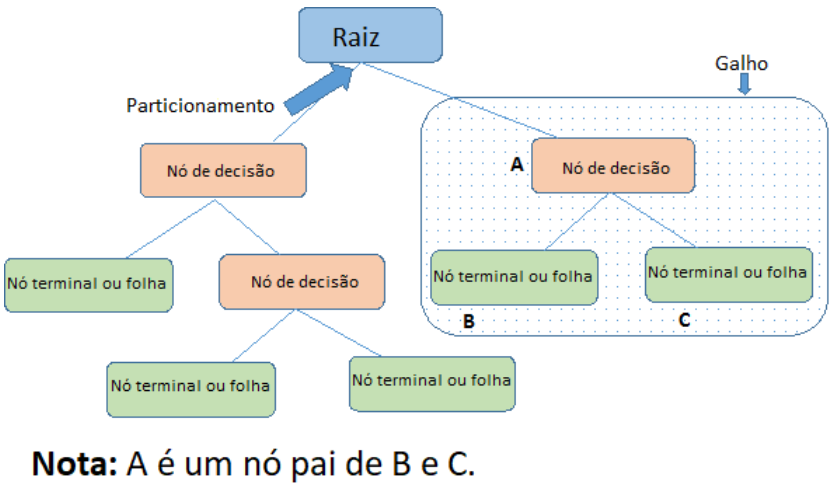
2.2 Floresta aleatória ou *random forest*

A floresta aleatória é um algoritmo de aprendizagem estatístico, onde temos um grupo de árvores de decisão que "votam"o valor para a nossa variável-alvo (BRUCE; BRUCE, 2019). Portanto, para compreender melhor o modelo de floresta aleatória, convém entender primeiramente o modelo de árvore de decisão.

2.2.1 Árvores de decisão em problemas de classificação

O algoritmo de árvore de decisão é um processo que auxilia na classificação de um dado elemento por meio de um particionamento recursivo de um ente, denominado *nó*, repetidamente, onde o primeiro nó é denominado *raiz* e contém todos os dados usados para treinamento. Quando o nó não é mais passível de particionamento, ele recebe o nome de *nó terminal* ou *folha*. Quando temos nós apresentando relação de parentesco entre si, o conjunto formado por eles pode ser chamado de *galho* (AWAD; KHANNA, 2015; LE, 2018). A Figura 2 abaixo ilustra uma árvore de decisão genérica.

Figura 2 – Ilustração de uma árvore de decisão e sua estrutura.



Fonte: adaptado de (LE, 2018)

Observando a Figura 2, algumas perguntas precisam ser respondidas para entender o funcionamento do algoritmo de árvore de decisão:

1. Como ocorre a divisão dos nós?

2. Quando uma uma árvore é considerada "boa"?

Tabela 1 – Observações realizadas para definir se um indivíduo é um vampiro ou não.

Presença de sombra	**Ingestão de alho**	**Compleição da pele**	**Qualidade do sotaque**	**É vampiro?**
Não sabemos.	Sim.	Pálida.	Sem sotaque.	Não.
Sim.	Sim.	Corada.	Sem sotaque.	Não.
Não sabemos.	Não.	Corada.	Sem sotaque.	Sim.
Não.	Não.	Normal.	Forte	Sim.
Não sabemos.	Não.	Normal.	Leve.	Sim.
Sim.	Não.	Pálida.	Forte.	Não.
Sim.	Não.	Normal.	Forte.	Não.
Não sabemos	Sim.	Corada.	Leve.	Não.

Fonte: adaptado de (WINSTON, 2010).

O problema ilustrado na próxima seção, baseado na aula sobre árvores de identificação ou de decisão (WINSTON, 2010), terá como objetivo responder a essas duas perguntas.

2.2.2 Funcionamento do algoritmo de árvore de decisão

O problema exemplo adotado pelo Prof. Winston em sua aula (WINSTON, 2010) é o de identificar vampiros imigrados do leste europeu nos Estados Unidos. Nessa situação, ele avalia quatro critérios (variáveis preditoras) para determinar se um indivíduo é vampiro ou não (variável alvo). Os critérios são: presença de sombra, ingestão de alho, compleição da pele e qualidade do sotaque. A tabela 1 apresenta os dados das observações realizadas.

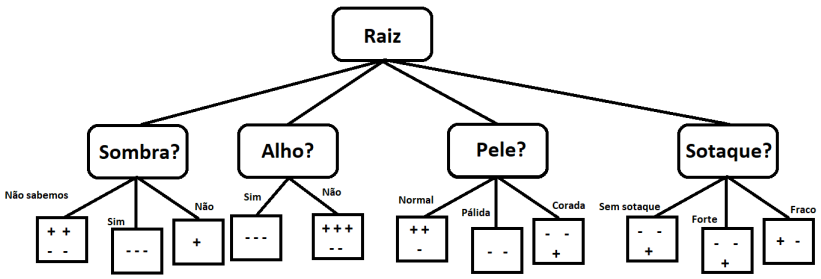
Nesse contexto e fazendo referência com a Figura (2), a Tabela (1) seria a raiz da árvore de decisão. Para dividir a raiz em outros nós, devemos realizar testes sobre esses dados e avaliar o que o teste faz com os dados. Os testes são basicamente perguntas que a tabela pode responder como, por exemplo, *o indivíduo possui sombra?*. Contudo, a pergunta deve ser não-óbvia, como a pergunta *eu posso ver o indivíduo?*[1](LE, 2018; WINSTON, 2010). "Perguntas óbvias"nos levarão a fazer mais perguntas a fim de filtrar novas características do indivíduo, o que faz a árvore de decisão ser maior. Isso nos leva a concluir que a melhor árvore de decisão é a menor possível para um dado conjunto de dados de treino(AWAD; KHANNA, 2015; LE, 2018; WINSTON, 2010).

Para continuar com o processo, seguiremos com os testes de presença de sombra, ingestão de alho, compleição da pele e qualidade do sotaque. Cada teste gerará um nó que será dividido no número de classes distintas existentes na coluna que responde a pergunta do teste realizado. Quando o indivíduo for um vampiro, ele será registrado com + na saída do novo nó, e com − caso contrário.

Os resultados dos testes são conjuntos de indivíduos classificados como vampiros (+) ou pessoas comuns (−) e, com isso, podemos chegar a algumas conclusões quando o

[1]Obviamente, se a aparência do indivíduo foi registrada, então é por que ele foi observado.

Figura 3 – Ilustração de uma árvore de decisão para o problema de classificar um indivíduo como pessoa comum ou um vampiro.



Fonte: adaptado de (WINSTON, 2010)

conjunto é homogêneo. Por exemplo: se o indivíduo possui sombra, o indivíduo é uma pessoa comum. Contudo, se não sabemos se o indivíduo possui ou não sombra por que só nos encontramos com ele em lugares escuros, não podemos concluir nada e devemos realizar um novo teste. Essa situação nos permite responder à pergunta 2 realizada na subseção anterior. Dividimos os nós até eles se apresentarem da forma mais homogênea possível (WINSTON, 2010). Então precisamos de uma métrica que avalie o grau de homogeneidade ou ordem desses grupos.

Uma das métricas que faz essa avaliação é o índice de impureza de Gini, dado por (GÉRON, 2019):

$$G_i = 1 - \sum_{k=1}^{m} \left(\frac{n_{i,k}}{N_i} \right)^2 \tag{2.2}$$

Onde $n_{i,k}$ é o número de ocorrência de uma dada instância k em um nó i e N_i é o número total de ocorrências em um nó i. Quanto maior esse índice, mais impuro é o nó e, então, maior é a necessidade de realizar uma nova divisão em outros nós, o que crescerá a árvore de decisão e torna-a pior.

Para o teste de sombra, temos os seguintes valores:

$$G_{sombra} = 1 - \left(\frac{n_{nãoSabemosVampiro}}{8} \right)^2 - \left(\frac{n_{nãoSabemosNormal}}{8} \right)^2 - \left(\frac{n_{simNormal}}{8} \right)^2 - \left(\frac{g_{nãoVampiro}}{8} \right)^2$$

$$G_{sombra} = 1 - \left(\frac{2}{8} \right)^2 - \left(\frac{2}{8} \right)^2 - \left(\frac{3}{8} \right)^2 - \left(\frac{1}{8} \right)^2$$

$$\boxed{G_{sombra} = 0,71875} \tag{2.3}$$

Se considerarmos agora nó filho "Não sabemos", teremos:

$$G_{sombraNãoSabemos} = 1 - \left(\frac{n_{normal}}{4} \right)^2 - \left(\frac{n_{vampiro}}{4} \right)^2$$

Tabela 2 – Subtabela de contendo todos os registros de indivíduos que não sabemos se possuem sombra ou não.

Presença de sombra	Ingestão de alho	Compleição da pele	Qualidade do sotaque	É vampiro?
Não sabemos.	Sim.	Pálida.	Sem sotaque.	Não.
Não sabemos.	Não.	Corada.	Sem sotaque.	Sim.
Não sabemos.	Não.	Normal.	Leve.	Sim.
Não sabemos	Sim.	Corada.	Leve.	Não.

Fonte: adaptado de (WINSTON, 2010).

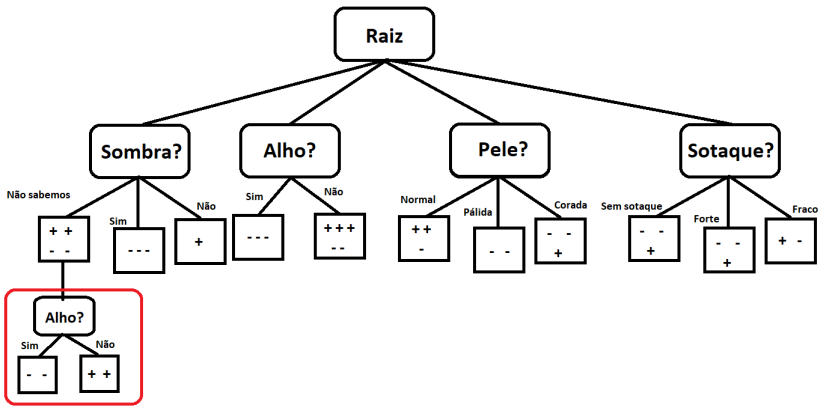
$$G_{sombraNãoSabemos} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$

$$\boxed{G_{sombraNãoSabemos} = 0,5} \tag{2.4}$$

Se analisarmos a equação (2.2), podemos concluir facilmente que $G_{sombraSim} = G_{sombraNão} = 0$. Então, um novo teste deve ser aplicado ao nó "Não sabemos"a fim de minimizar a impureza da árvore como um todo. Os dados que serão submetidos ao novo teste estão apresentados na Tabela 2.

Vamos fazer o teste do alho no conjunto da Tabela 2. Assim, nossa árvore de decisão passa a ter a configuração apresentada na Figura 4:

Figura 4 – Nova configuração da árvore de decisão após a aplicar o teste do alho no nó "Não sabemos"gerado no teste da sombra.



Fonte: adaptado de (WINSTON, 2010)

Então, podemos deduzir da Figura 4 e da equação (2.2) que $G_{sombraNSalhoS} = G_{sombraNSalhoNao} = 0$, indicando que não há mais necessidade de aplicar um novo teste, fazendo com que o galho da árvore gerado no teste de sombra pare de crescer.

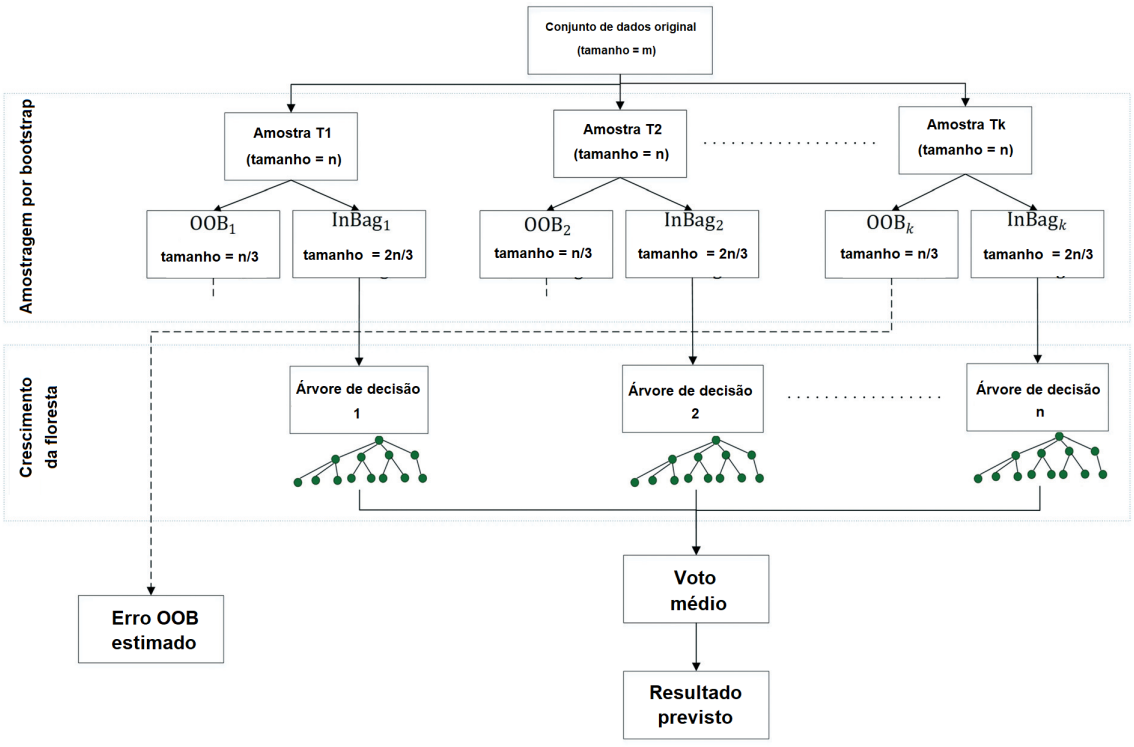
Nesse exemplo, contudo, como é sugerido por (WINSTON, 2010) e podemos observar se analisarmos a Tabela 1 cuidadosamente, o teste da sombra poderia ter sido aplicado apenas após o teste do alho, no nó "Não"logo abaixo da raiz, poderíamos ter chegado no mesmo resultado, porém a árvore teria um galho a menos.

Terminada essa breve digressão do algoritmo de árvore de decisão, voltemos para a revisão sobre o algoritmo de floresta aleatória.

2.2.3 O algoritmo de floresta aleatória

O algoritmo de floresta aleatória é simplesmente um conjunto de árvores de decisão crescidas a partir de amostras coletadas por *bootstrap* (HARTSHORN, 2016). Nesse processo, uma amostra aleatória é coletada da população e depois são realizadas outras amostragens desse subconjunto, coletando cerca de 2/3 para crescer uma árvore de decisão e depois repõem-se esses dados no conjunto. O valor aproximado de 1/3 não usado para crescer uma dada árvore de decisão é, então, usado para testá-la, comparando a sua previsão com o valor de teste (HARTSHORN, 2016). Essa é uma métrica importante para avaliação de modelos de floresta chamada erro *out of bag* (OOB) e será discutida mais adiante. A Figura 5 ilustra o processo de crescimento de uma floresta aleatória.

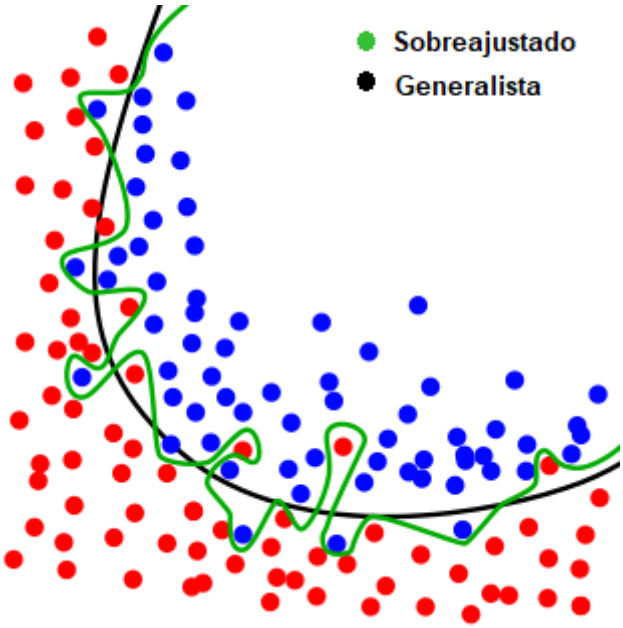
Figura 5 – Ilustração do processo de crescimento de uma floresta aleatória.



Fonte: adaptado de (SAADATKHAH et al., 2018).

Outro fato importante desse algoritmo é que, durante essa amostragem para os modelos de classificação, ele seleciona aleatoriamente $\lfloor \sqrt{p} \rfloor$ variáveis, sendo p o número total de variáveis distintas possíveis (WEINBERGER, 2018; HASTLE; TIBSHIRANI; FRIEDMAN, 2009) . Fazendo isso, o modelo de floresta aleatória torna-se mais generalista e menos dependente do conjunto de dados de treinamento do que a árvore de decisão. Modelos muito dependentes dos dados de treino apresentam alta variância ou sobreajuste e alta precisão quando testado por uma simples comparação, mas falham com a aplicação de um novo conjunto de dados (LIBERMAN, 2017; BRUCE; BRUCE, 2019). A Figura 6 ilustra as curvas geradas por um modelo sobreajustado (em verde) e generalista (em preto).

Figura 6 – Exemplo de sobreajuste.



Fonte: adaptado de (Elite Data Science, 2020).

2.2.4 Vantagens e desvantagens do algoritmo de floresta aleatória

Uma das maiores vantagens do algoritmo de floresta aleatória é que praticamente não há necessidade de realizar tratamento nos dados de treino do modelo para realizar as previsões, já que o modelos não-paramétricos não faz suposição nenhuma sobre a forma da função resposta (BROWNLEE, 2016a). Outra vantagem do algoritmo de floresta aleatória é que não há a necessidade de particionar o conjunto de dados em conjunto de treinamento e conjunto de teste por que o próprio algoritmo, ao crescer as árvores, utiliza os dados *OOB* de uma árvore para avaliá-la e estimar o erro *OOB* para o algoritmo de floresta aleatória (HARTSHORN, 2016).

Já como desvantagens, há maior necessidade de recurso computacional, o que pode inviabilizar o uso do modelo em aplicações em tempo real. Também há a o fato da floresta aleatória ser um sistema do tipo caixa preta, onde não há uma clara descrição das relações entre as variáveis no meio do processo (JANSEN, 2018), como numa equação em que podemos examinar o peso do coeficiente de uma variável sobre o total.

2.2.5 Decrescimento médio da impureza de Gini e a importância das variáveis.

Como apresentado anteriormente, o índice de impureza de Gini é uma quantidade que avalia o grau de impureza de um conjunto de dados e, no algoritmo de árvore de decisão, é uma métrica que pode definir quando um nó é dividido ou não pode ser dividido. No algoritmo de árvore de decisão, onde tem-se a intenção de criar os grupos mais homogêneos possíveis, percebemos que cada divisão gera nós filhos com menor índice de impureza do que o nó pai.

Também queremos que a árvore de decisão seja a menor possível para não termos problemas de sobreajuste, então precisaríamos pensar em uma forma de encontrar quais variáveis que, quando divididas, produzem os nós filhos mais homogêneos em relação aos nós pais (BRUCE; BRUCE, 2019). Uma forma de avaliar isso é calculando o índice de impureza de Gini, dada pela equação (2.2), no nó pai e subtraindo pelos índices de impureza de Gini de cada nó filho, multiplicado por um coeficiente de contribuição para aquele nó (LAURETTO; BASTOS; NASCIMENTO, 2014):

$$\Delta G_i = G_i - \left(\sum_{j=1}^{J} \frac{n_j}{n_i} G_j \right) \tag{2.5}$$

onde G_i é o índice de impureza de Gini para um nó pai i pertencente a uma determinada variável, n_i é o número de instâncias no nó pai, n_j é o número de instâncias no nó filho e G_j é o índice de impureza de Gini para um nó filho j. Em uma floresta aleatória, contudo, poderíamos ter vários nós-pais em uma mesma altura da árvore, então teríamos o decrescimento médio da impureza de Gini (LAURETTO; BASTOS; NASCIMENTO, 2014):

$$\Delta \overline{G}_m = \frac{1}{K} \cdot \sum_{k=1}^{K} \sum_{i \exists m=1}^{N} \Delta G_{k,i} \tag{2.6}$$

onde k é o índice da árvore que contém um nó i pertencente ao ramo de uma determinada variável m que contém um total de N nós, K é o número total de árvores de decisão na floresta aleatória e $\Delta G_{k,i}$ é o índice de impureza de Gini da equação (2.5) de um nó i em uma árvore k, que seria a equação aplicado a um nó dessa árvore.

Dessa forma, ao escolhermos os nós associados a uma determinada variável, podemos classificar a sua importância comparando os valores de decrescimento médio do índice de impureza de Gini, dado pela equação (2.6), onde a variável que apresentasse um maior valor de $\Delta\overline{G}_m$ do que outra, seria considerada mais importante.

2.3 Métricas de avaliação de um modelo gerado pelo algoritmo de floresta aleatória.

2.3.1 Matriz de confusão

A matriz de confusão é uma tabela que apresenta os as previsões corretas e incorretas de um modelo de classificação (BRUCE; BRUCE, 2019) a partir da partição de teste do conjunto de dados utilizados na construção do modelo. A estrutura de uma matriz de confusão para um modelo de classificação binário é apresentada na Figura 7:

Figura 7 – Contagem de classes distintas do campo *CONDUTA* e sua frequencia relativa acumulada.

		Valores reais	
		Positivo	**Negativo**
Valores previstos	**Positivo**	Verdadeiro positivo (TP)	Falso positivo (FP)
	Negativo	Falso negativo (FN)	Verdadeiro negativo (TN)

Fonte: adaptado de (BRUCE; BRUCE, 2019).

Na Figura 7, a região "Verdadeiro positivo"conterá a frequência com que um valor foi previsto pelo modelo como *Positivo* e o valor era realmente *Positivo*. A região "Verdadeiro negativo"conterá a frequencia com que um valor foi previsto pelo modelo como negativo e o valor era realmente *negativo*. Já na região "Falso positivo", temos a frequência de quando o modelo prevê o valor *Positivo* quando, na realidade, ele era *Negativo* e, por fim, na região "Falso negativo", teremos a frequência em que o modelo previu o valor *Positivo* quando, na verdade, ele era *Negativo*.

Existem métricas que podem ser construídas com base nos valores das diferentes regiões da matriz de confusão que podem facilitar a análise do modelo em casos de variáveis categóricas com mais de dois valores. Uma dessas métricas é a precisão (P), que

informa a relação entre o número de valores verdadeiros positivos $\sum X_{verdadeiroPositivo}$ e o total $\sum X_{verdadeiroPositivo} + \sum X_{falsoPositivo}$ de valores que o modelo previu como positivo (BRUCE; BRUCE, 2019):

$$P = \frac{\sum X_{verdadeiroPositivo}}{\sum X_{verdadeiroPositivo} + \sum X_{falsoPositivo}} \tag{2.7}$$

Outra métrica que pode ser obtida a partir da matriz de confusão é a revocação (R) ou sensibilidade. Essa medida dá a força com que o modelo é capaz de prever o resultado positivo relacionando o número de valores positivos $\sum X_{verdadeiroPositivo}$ com o total $\sum X_{falsoNegativo} + \sum X_{falsoPositivo}$ de valores *Positivo* que existe no conjunto de dados (BRUCE; BRUCE, 2019):

$$R = \frac{\sum X_{verdadeiroPositivo}}{\sum X_{verdadeiroPositivo} + \sum X_{falsoNegativo}} \tag{2.8}$$

Também há uma métrica usada para prever a capacidade de um modelo prever valores negativos, que é chamada de especifidade (E). É uma "precisão para valores negativos", obtida através da relação entre a frequência de valores verdadeiros negativos $\sum X_{verdadeiroNegativo}$ com o total $\sum X_{verdadeiroNegativo} + \sum X_{falsoNegativo}$ de valores que o modelo previu como negativo (BRUCE; BRUCE, 2019):

$$E = \frac{\sum X_{verdadeiroNegativo}}{\sum X_{verdadeiroNegativo} + \sum X_{falsoNegativo}} \tag{2.9}$$

No pacote *randomForest* do *R*, é apresentada a medida *class.error*, que pode ser dado por:

$$\overline{P} = 1 - P \tag{2.10}$$

Em que P é o valor da precisão dada pela equação (2.7).

2.3.2 Erro *out of bag*.

O erro *out of bag* é uma métrica de avaliação do modelo de floresta aleatória. Essa métrica, como mencionado anteriormente, é gerada ao testar uma árvore de decisão, que foi crescida com aproximadamente 2/3 de um conjunto de dados selecionado no processo de amostragem por *bootstrap*, e em que os dados de teste são os 1/3 restante desses dados não utilizados para crescer essa árvore.

Esses valores de 1/3 para teste e 2/3 para crescimento vêm do fato de que a probabilidade de um dado, no processo de amostragem aleatória com reposição, **não** ser

selecionado em um conjunto de dados de tamanho x é dada por (SWENSSON; SäRNDAL; WRETMAN, 1991):

$$\tilde{P} = \left(1 - \frac{1}{x}\right)^x \tag{2.11}$$

Então, para um conjunto de dados muito grande, onde $x \to \infty$, teremos:

$$\lim_{x\to\infty} \left(1 - \frac{1}{x}\right)^x \tag{2.12}$$

Contudo, ao considerarmos que:

$$\left(1 - \frac{1}{x}\right)^x = e^{\ln\left(1-\frac{1}{x}\right)^x}$$

Portanto:

$$\lim_{x\to\infty} \left(1 - \frac{1}{x}\right)^x = \lim_{x\to\infty} e^{\ln\left(1-\frac{1}{x}\right)^x} \tag{2.13}$$

Então aplicaremos o limite apenas no argumento da exponencial:

$$\lim_{x\to\infty} \ln\left(1 - \frac{1}{x}\right)^x = \lim_{x\to\infty} \left[x \cdot \ln\left(\frac{x-1}{x}\right)\right] = \lim_{x\to\infty} \left[\frac{\ln\left(\frac{x-1}{x}\right)}{\frac{1}{x}}\right] \tag{2.14}$$

O limite (2.14) tem a forma de uma função do tipo 0/0, o que nos permite aplicar a ténica de L'hospital. Então, teremos:

$$\lim_{x\to\infty} \frac{\frac{d}{dx}\left[\ln\left(\frac{x-1}{x}\right)\right]}{\frac{d}{dx}\left(\frac{1}{x}\right)} = \lim_{x\to\infty} \frac{x \cdot (x-1)}{\frac{-1}{x}} = \lim_{x\to\infty} \frac{x}{x-1}$$

Então, teremos:

$$\boxed{\lim_{x\to\infty} \frac{x}{x-1} = -1} \tag{2.15}$$

Ao substituirmos o resultado do limite (2.15) em (2.13), teremos:

$$\lim_{x\to\infty} \left(1 - \frac{1}{x}\right)^x = \lim_{x\to\infty} e^{\ln\left(1-\frac{1}{x}\right)^x} = \frac{1}{e} = 0,36788 \approx \boxed{\frac{1}{3}} \tag{2.16}$$

Agora que compreendemos a origem das parcelas de 1/3 dos dados *out of bag* utilizados para a avaliação do modelo e os 2/3 utilizados para crescer cada árvore, vamos prosseguir com um exemplo presente para ilustrar como o cálculo do erro *OOB* é realizado, assim como é descrito por (HARTSHORN, 2016).

Imagine que temos 10 dados que queremos classificar como pertencentes a uma entre três categorias denominadas por A, B ou C, utilizando um algoritmo de floresta aleatória que contém 15 árvores de decisão. Então, nesse caso, podemos ter 4 pontos OOB.

O primeiro dado OOB pertence, por exemplo, à categoria A e é um dado que não foi usado para crescer 5 árvores. Então esse dado é usado para avaliar cada uma dessas 5 árvores, onde 3 árvores o classificaram como pertencente à categoria A, uma o classificou como categoria B e outro como categoria C. Nesse caso, o conjunto de 5 árvores para aquele dado o classificaria como pertencente à categoria A, o que está correto.

O segundo dado é OOB para um conjunto de 7 árvores e pertence à categoria A. Do conjunto de 7 árvores, 3 o classificam como categoria A, 2 como B e as outras 2 restantes como C. A categoria que recebeu mais votos foi a A, então para esse conjunto de árvores a categoria desse dado é A.

Para um terceiro dado com categoria B, que é OOB para 4 árvores, duas árvores o classificam como C e uma só árvore como B. Nesse caso, o conjunto decidiria que esse dado é da categoria C, o que está errado.

Para o quarto dado, pertencente à categoria C e OOB para 4 árvores, temos que 2 árvores de decisão o classificam como B e outras duas como C. Em casos de empate a decisão da classificação depende do *software* implementado, que poderia usar como critério de avaliação a primeira categoria que atendeu 50% das árvores do conjunto. Nesse caso, a categoria votada por esse conjunto seria **B**, o que seria, também, um erro

Sendo assim, de 4 pontos, 2 foram classificados corretamente e outros 2 incorretamente, resultando num erro OOB de $0,5$. Caso o quarto dado tivesse sido classificado corretamente, o erro OOB seria de $0,25$.

O erro OOB é uma boa métrica para avaliar a qualidade do modelo como um todo, mas ainda podemos fazer uso de outra medida estatística para medir a qualidade da previsão de uma determinada classe. Isso é especialmente útil quando trabalhamos com mais de duas categorias na nossa variável alvo.

2.3.3 Concordância entre avaliadores e o índice Kappa de Fleiss.

O índice Kappa de Fleiss (κ_{Fleiss}) é uma medida estatística que permite avaliar o grau de concordância entre dois ou mais avaliadores (MINITAB, 2020). Esse índice é definido como (FLEISS, 1971):

$$\kappa_{Fleiss} = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \tag{2.17}$$

Tabela 3 – Interpretação do índice κ_{Fleiss}

κ_{Fleiss}	**Força de concordância**
até 0,00	Sem concordância
0,01 até 0,20	Concordância leve.
0,21 até 0,40	Concordância razoável.
0,41 até 0,60	Concordância moderada.
0,61 até 0,80	Concordância considerável.
0,81 até 1	Concordância quase absoluta (ou absoluta, no caso de 1)

Fonte: adaptado de (LANDIS; KOCH, 1977).

Onde o termo $1 - \overline{P_e}$ avalia o grau de concordância possível de se atingir além do que seria concordado por acaso. Já o termo $\overline{P} - \overline{P_e}$ avalia o grau de concordância realmente alcançado além do que seria concordado por acaso. Então podemos entender da equação (2.17) que valores onde $\kappa_{Fleiss} \to 1$ ocorrem quando há maiores graus de concordância entre os avaliadores e para $\kappa_{Fleiss} \to 0$, não há concordância. A interpretação para o κ_{Fleiss} é resumida na tabela (tab:interpretacao-kappa-fleiss) (LANDIS; KOCH, 1977):

Vamos agora desenvolver os termos da equação apresentada em (2.17). Nessa equação, o termo $\overline{P}$ representa a média da concordância total entre os avaliadores e é definida por (FLEISS, 1971):

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)} \cdot \left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn \right) \tag{2.18}$$

Onde:

$$P_i = \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right)$$

O termo P_i em (2.18) representa a proporção média da concordância entre n avaliadores a respeito da classificação de um i-ésimo indivíduo entre N indivíduos. O índice j, por sua vez, representa a classe em que se é possivel categorizar o indivíduo. Então o termo n_{ij} representa o número avaliadores, dentre um grupo de n, que atribuiu uma categoria j para um indivíduo i (FLEISS, 1971).

Já o termo $\overline{P}_e$ da equação (2.17) representa a proporção média em que os avaliadores atribuíram uma classe de forma totalmente aleatória e é dado por (FLEISS, 1971)?

$$\overline{P}_e = \sum_{j=1}^{k} p_j^2 \tag{2.19}$$

Tabela 4 – Frequência de atribuição de uma categoría j, dentre 5 possíveis, para indivíduo i, dentre 10, realizada por $n = 14$ avaliadores.

$\mathbf{n_{ij}}$	**j = 1**	**j=2**	**j=3**	**j=4**	**j=5**	$\mathbf{P_i}$
i=1	0	0	0	0	14	**1,000**
i=2	0	2	6	4	2	**0,253**
i=3	0	0	3	5	6	**0,308**
i=4	0	3	9	2	0	**0,440**
i=5	2	2	8	1	1	**0,330**
i=6	7	7	0	0	0	**0,462**
i=7	3	2	6	3	0	**0,242**
i=8	2	5	3	2	2	**0,176**
i=9	6	5	2	1	0	**0,286**
i=10	0	2	2	3	7	**0,286**
Total	**20**	**28**	**39**	**21**	**32**	**3,783**
$\mathbf{p_j}$	**0,143**	**0,200**	**0,279**	**0,150**	**0,229**	

Fonte: adaptado de (FLEISS, 1971; WIKIPÉDIA, 2020).

Na expressão (2.19), o termo p_j é a proporção em que uma categoria j é atribuida dentre todas as Nn atribuições realizadas. Por sua vez, p_j é definido por (FLEISS, 1971):

$$p_j = \frac{1}{Nn} \sum_{i=1}^{k} n_{ij} \tag{2.20}$$

Onde, como já definido anteriormente, N é o número total de indivíduos, n é o número de avaliadores e n_{ij} é o número de avaliadores que atribuíram uma categoria j para um indivíduo i.

Para efeito de ilustração, tomemos os exemplos trabalhados em (FLEISS, 1971; WIKIPÉDIA, 2020).

Para a Tabela 3, temos $N = 10$, $n = 14$, $Nn = 140$, $k = 5$. Para a categoria $j = 2$, usando a equação (2.20), teremos:

$$p_2 = \frac{0 + 2 + 0 + 3 + 2 + 7 + 2 + 5 + 5 + 2}{140} = \frac{28}{140} = 0,200$$

Para o indivíduo $i = 5$, de acordo com as equações (2.19) e (2.20), teremos:

$$P_5 = \frac{1}{14(14 - 1)} \cdot (2^2 + 2^2 + 8^2 + 1^2 + 1^2 - 14) = 0,330$$

O termo P, segundo a equação (2.18), será:

$$\overline{P} = \frac{1}{10} \cdot 3,780 = 0,3780$$

Já o termo $\overline{P_e}$, de acordo com (2.19), será:

$$\overline{P}_e = 0,143^2 + 0,200^2 + 0,279^2 + 0,150^2 + 0,229^2 = 0,213$$

Então, substituindo os termos acima em (2.17), teremos:

$$\kappa_{Fleiss} = \frac{0,378 - 0,213}{1 - 0,213} = 0,210$$

O κ_{Fleiss} se mostra importante principalmente quando um conjunto de classes é mais frequente do que outros em uma população. Ainda é possível converter o valor de κ_{Fleiss} para valores z_{score}, amplamente utilizados em estatística. Fazendo isso, podemos avaliar o o quão melhor é a concordância entre eles caso ela acontecesse por mero acaso. Em "linguagem estatística", podemos fazer um teste de hipóteses, onde a hipótese H_0 significa que houve concordância por acaso e a hipótese alternativa H_1 significa que não houve acaso. O z_{score} é definido por (FLEISS, 1971):

$$z_{score}(\kappa_{Fleiss}) = \frac{\kappa_{Fleiss}}{\sqrt{\sigma^2(\kappa_{Fleiss})}} \tag{2.21}$$

Onde $\sigma^2(\kappa_{Fleiss})$ é a variância do índice κ_{Fleiss}s, definido por (FLEISS, 1971) como:

$$\sigma^2(\kappa_{Fleiss}) = \frac{2}{Nn \cdot (n-1)} \cdot \frac{\sum_{j=1}^{k} p_j^2 - (2n-3) \cdot \left(\sum_{j=1}^{k} p_j^2\right)^2 + 2 \cdot (n-2) \cdot \sum_{j=1}^{k} p_j^2}{\left(1 - \sum_{j=1}^{k} p_j^2\right)^2} \tag{2.22}$$

Então, caso o índice $z_{score}(\kappa_{Fleiss})$ corresponda a um p_{valor} menor que um dado nível de significância, podemos rejeitar a hipótese nula H_0.

2.3.4 Teste de hipóteses.

Testes de hipóteses, ou testes de significância, tem como objetivo verificar o quão melhor é um resultado observado em comparação a um resultado aleatório. A ideia desses testes é verificar se um evento que ocorreu realmente é um resultado diferente dos anteriormente observados ou se ele ocorreu por mero acaso (BRUCE; BRUCE, 2019). Em livros de estatística (MAGALHÃES; LIMA, 2004), esse assunto costuma ser abordado em capítulos de inferência estatística onde, dado um nível de significância, deseja-se inferir

se os efeitos observados em um espaço amostral de eventos são iguais ao observado na população ou em outro gripo de eventos.

O método para realizar esses testes consiste em escolher uma medida estatística de interesse, definir a hipótese a ser testada (hipótese nula ou H_0) e a hipótese alternativa (H_1), calcular os valores que desejamos comparar fazendo uso das medidas estatísticas apropriadas, delimitar o intervalo de rejeição segundo o nivel de significância adotado, comparar os valores obtidos e fazer a conclusão. Assim como fizemos com o algoritmo de floresta aleatória, vamos ilustrar as ideias gerais de um teste de hipóteses com um exemplo.

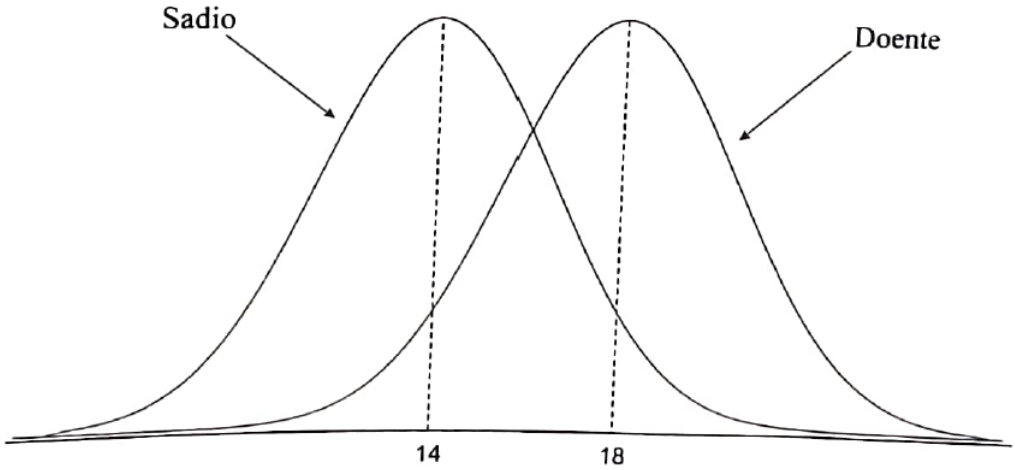
2.3.5 Teste de hipóteses para média populacional com variância conhecida: breve revisão.

Para ilustrar a aplicação de um teste de hipóteses, utilizaremos o exemplo 8.1 da referência bibliográfica (MAGALHÃES; LIMA, 2004). Assim como o título dessa subseção indica, a medida de interesse que queremos avaliar é a média e, para o teste que realizaremos, a variância ou desvio padrão populacional são conhecidos. Nesse teste, avaliamos se podemos afirmar que a média amostral de uma dada variável contínua é igual à média populacional , assumindo que a variável em questão tem uma distribuição normal. No exemplo aqui ilustrado, deseja-se saber se a concentração de uma determinada substância no sangue pode ser usada como indicador de uma doença específica. Sabe-se que a média da concentração dessa substância em uma população sadia é $\overline{x} = 14$ unidades por mL, mas média obtida em laboratório para a concentração da substância no sangue de pessoas com a doença foi de 18 unidades por mL. A distribuição de probabilidade das concentrações da substância no sangue de ambos grupos é ilustrada na Figura 8:

Como podemos observar na Figura 8, as distribuições se cruzam a partir determinado ponto e é nessa região onde os grupos se sobrepõem que podemos tomar conclusões erradas em nossa avaliação, no caso, se a concentração da substância no sangue pode ser considerada como um indicador da doença. Existe uma probabilidade não-nula de classificarmos o teste sanguíneo como adequado para indicar a doença e uma probabilidade não-nula de considerarmos o teste como ineficaz para a avaliação quando, na verdade, ele é adequado. Também poderíamos pensar em classificar uma pessoa como doente quando na verdade não é ou o contrário, classificar uma pessoa como sadia quando na verdade ela não é.

Em estatística, esses erros são denominados de *Erro tipo I* ou *Erro* α e *Erro tipo II* ou *Erro* β. O erro α consiste em rejeitar a hipótese nula (H_0) quando ela é verdadeira, já o erro β significa que não rejeitamos a hipótese nula quando ela se mostrou como falsa. A probabilidade de ocorrer o erro α, ou nível de significancia, é representada pela região

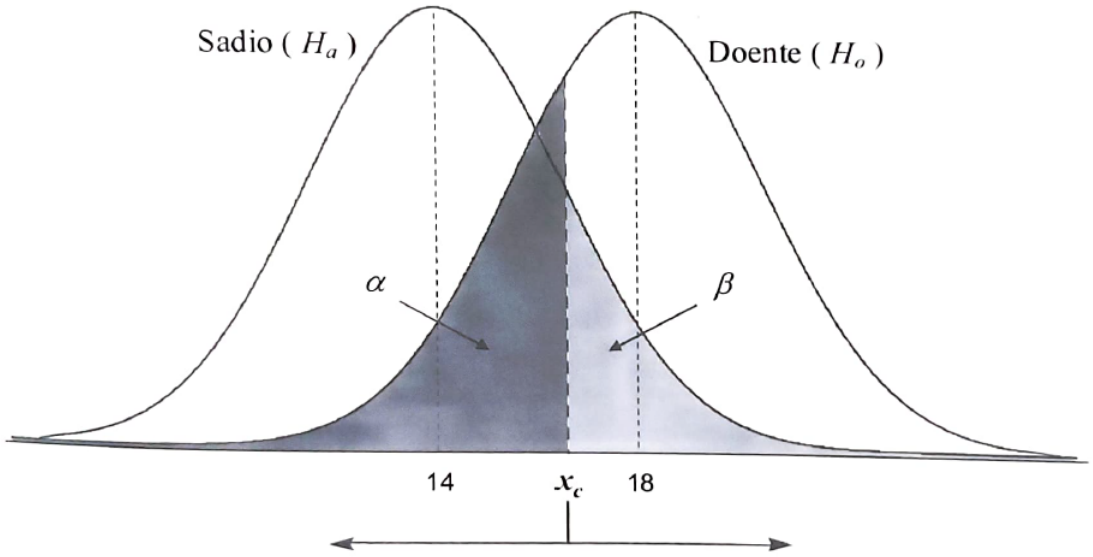
Figura 8 – Distribuição de probabilidade da concentração de uma determinada substância em um grupo de pessoas sadias e doentes.



Fonte: (MAGALHÃES; LIMA, 2004).

α e a probabilidade de ocorrer o tipo β é de ocorrer o erro β é representada pela região β. Ambas as regiões são delimitadas por um valor crítico x_c, como é ilustrado na Figura 9:

Figura 9 – Erros α e β representados no gráfico de distribuição de probabilidades.



Fonte: (MAGALHÃES; LIMA, 2004).

Na Figura 9, atribui-se como hipótese nula H_0 a ideia de que a distribuição de probabilidade de concentração da substancia ser igual em ambos grupos, saudável e doente. A hipótese alternativa (H_a ou H_1), por sua vez, admite que os grupos possuem médias

diferentes e, portanto, a distribuição não é igual. Adota-se como hipótese nula a construção lógica em que nada aconteceu e qualquer diferença entre observações se deram por mero acaso, já na ideia de hipótese alternativa, aceitamos que as diferenças entre as observações são, de fato, significativas (BRUCE; BRUCE, 2019). Então, caso a média do conjunto de observações adentre a região crítica α, podemos rejeitar a hipótese dos dois grupos serem iguais, caso o contrário, não teremos evidências de que a diferença entre os dois conjuntos de observações é significativo. Na Figura 9, por exemplo, dado um nível de significância α, temos indícios de que há uma diferença significativa na concentração da substância no sangue de pessoas doentes e sadias e o nível de confiança desse resultado é de $1 - \alpha$.

Contudo, é comum padronizar a medida estatística que estamos avaliando para consultar valores tabelados e realizar as devidas comparações. Nessa padronização, calculamos uma medida chamada z_{score} e comparamos com o valor $z_{crítico}$. Para os testes de média, o z_{score} é calculado da seguinte forma:

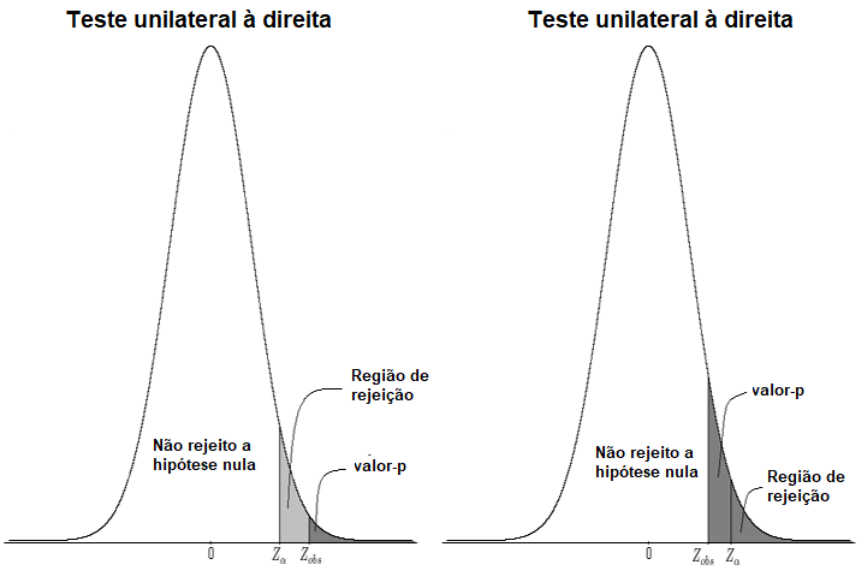
$$z_{score} = \frac{\overline{x}_{observado} - \overline{X}_{populacional}}{\frac{\sigma_{populacional}}{\sqrt{n}}} \tag{2.23}$$

Onde $\overline{x}_{observado}$ é a média dos valores observados, $\overline{X}_{populacional}$ é a média populacional, $\sigma_{populacional}$ é o desvio médio padrão da população e n é o número de observações aleatórias realizadas. Para obter o valor de $z_{crítico}$, basta substituir o valor de $\overline{x}_{observado}$ por $\overline{x}_{critico}$ e o n pelo número correspondente de observações utilizados para obter $\overline{x}_{critico}$. Nessa situação, a área que contém a região do erro α passa a ser delimitada pelo valor $z_{crítico}$ e qualquer valor de z_{score} calculado que adentre essa região, permitirá a rejeição de H_0.

Outra forma de se avaliar o teste de hipótese é considerando a probabilidade de se observar valores mais extremos que a estatística de teste. Essa medida, denominada como *valor-p* ou nível descritivo (FERREIRA; PATINO, 2015; MAGALHÃES; LIMA, 2004), seria a área delimitada pela curva da distribuição de probabilidade e o valor da estatística z_{score}. Caso o valor-p seja menor do que o nível de significância α, significando que z_{score} é um valor mais extremo do que $z_{critico}$, poderíamos rejeitar a hipótese nula (MAGALHÃES; LIMA, 2004; PORTAL ACTION, 2020). A Figura 10 ilustra a interpretação gráfica do valor-p em uma distribuição de probabilidades de uma variável qualquer.

Outro conceito importante em testes de hipóteses é a direcionalidade do teste. Quando realizamos um teste de hipótese unicaudal, podemos avaliar se o erro a estatística de teste é menor ou maior do que a de um determinado grupo. Podemos realizar um teste unicaudal à esquerda para avaliar se o valor médio de uma distribuição está dentro da região extrema inferior de outra distribuição, como ilustrado na Figura 9, onde o valor médio da concentração da substância no sangue de pessoas sadias é um valor extremo

Figura 10 – Valor-p no teste de hipótese unilateral, considerando duas variáveis de distribuição normal.



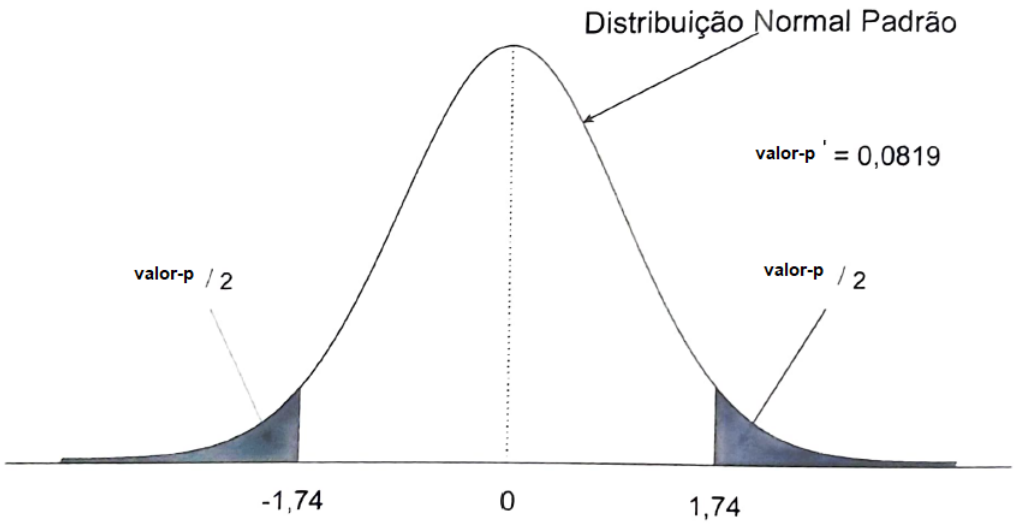
Fonte: adaptado de (PORTAL ACTION, 2020).

inferior em relação à dirtribuição de concentração da substância no sangue de pessoas doentes. Também poderíamos realizar um teste unicaudal à direita para avaliar o quão significativamente maior é o valor médio da concentração da substância no sangue nas pessoas doentes em relação às pessoas sadias. Neste último caso, a hipótese nula seria a distribuição de amostras sadias e a alternativa as amostras doentes. Também há a possibilidade de biredicionalidade e, nesse caso, só importa se os valores das medidas estatísticas são diferente. Nesse caso, comparamos os valores de α e do nível descritivo distribuído nos limites da curva de distribuição (MAGALHÃES; LIMA, 2004; BRUCE; BRUCE, 2019) Em muitas funções do *software R*, como é o caso da função $kappa.mfleiss$ da biblioteca IRR (GAMER, 2019), os testes realizados são bicaudais. A Figura 11 ilustra um teste bicaudal.

2.3.6 Teste de hipóteses para o kappa de Fleiss.

O teste de hipóteses para o κ_{Fleiss} funciona com a mesma ideia de um teste de hipóteses para média. Poderíamos testar hipótese de dois valores de κ_{Fleiss} serem diferentes a partir de uma generalização da equação (2.21), dada por (PAAS SAMPLE

Figura 11 – Valor-p no teste de hipótese bilateral de uma variável com distribuição normal.



Fonte: adaptado de (MAGALHÃES; LIMA, 2004).

SIZE SOFTWARE, 2020):

$$z_{score}(\kappa_{Fleiss}) = \frac{\kappa_{Fleiss} - \kappa_0}{\sqrt{\sigma^2(\kappa_{Fleiss})}} \tag{2.24}$$

onde o termo κ_0 é o valor da estatística atribuída à hipótese nula. Então, caso se considerarmos $\kappa_0 = 0$, que seria o valor crítico nesse cenário, temos como hipótese nula que a concordância entre dois avaliadores ocorre meramente ao acaso. Dessa forma, um dado nível de significância α, poderíamos fazer uso do teste de hipóteses para avaliar se a concordância entre avaliadores é aleatória ou não.

3 MATERIAIS E MÉTODOS

Para construir um modelo de classificação por meio do algoritmo de floresta aleatória, usamos os seguintes recursos:

- 1 computador pessoal com processador Intel i5 8^a geração com 8GB de memória RAM
- 1 conjunto de dados com (descrever a base de dados) registros de boletins de ocorrência coletados entre 2007 e 2014, disponível em (ZANCHI, 2019).
- Sistema gerenciador de banco de dados *Microsoft SQL Server* para pré-processamento e contenção dos dadoss
- *Software* R para construção e avaliação do modelo.
- *Software* Microsoft Excel para a construção de tabelas e gráficos.

3.1 Exploração do conjunto de dados

O conjunto de dados que vamos analisar contém registros de boletins de ocorrência de crimes ocorridos na Grande São Paulo entre 01/01/2007 e 31/12/2014. Há, em média, 822313 registros por ano, com desvio de ±89723. Desses registros, 8060086 foram realizados pelas vítimas da ocorrência, sendo que a média de registros por ano foi de 503755, com desvio de 63590 registros. O número de colunas em cada tabela é 30 segundo o dicionário de dados, contudo algumas tabelas possuem colunas extras em branco ou com valores predominantemente nulos ou são redundantes. A Tabela 5 resume o conjunto de dados que foi utilizado para construir o modelo e a Tabela 6 é o dicionário de dados que acompanha o conjunto de dados.

As colunas 17, 20 e 21 não estavam presentes. Já a coluna *CONT_PESSOA*, estava presente no conjunto de dados mas não apareceu dicionário de dados. O campo *COR_CUTIS* também não estava presente no conjunto de dados, contudo, considerou-se que o campo *COR* seria o equivalente.

Das colunas apresentadas na Tabela (6, foram mantidas apenas *CONDUTA*, *HORA_OCORRENCIA_BO*, *DATA_OCORRENCIA_BO*, *SEXO_PESSOA*, *IDADE_PESSOA*, *DESCR_PROFISSAO*, *DESCR_GRAU_INSTRUCAO* e *RUBRICA*. A coluna *CONDUTA* foi mantida por, apesar da descrição dos dados, seus valores se tratarem de uma descrição do tipo de local. Desta forma, a coluna *CONDUTA*, *HORA_OCORRENCIA_BO* e *DATA_OCORRENCIA_BO*, podemos considerar em qual dia da semana e horário a vítima da ocorrência estava num determinado tipo de local onde

Tabela 5 – Descrição do conjunto de dados utilizados no modelamento

Nome da tabela	Número de registros	Casos registrados por vítimas	Número de colunas
BO_2007_1	725958	420979	30
BO_2007_2	712646	414295	30
BO_2008_1	746214	441379	30
BO_2008_2	743788	471749	30
BO_2009_1	801790	492327	33
BO_2009_2	777962	482068	32
BO_2010_1	756312	470924	30
BO_2010_2	770961	477452	30
BO_2011_1	797120	501882	52
BO_2011_2	841561	527781	30
BO_2012_1	882195	558072	33
BO_2012_2	845897	534523	30
BO_2013_1	866691	552523	30
BO_2013_2	1043729	669113	30
BO_2014_1	954112	565005	30
BO_2014_2	890075	480014	30
Total:	**13157011**	**8060086**	**N.A.**
Média:	**822313,19**	**503755,40**	**N.A.**
Desvio médio padrão (DMP):	**89723,31**	**63589,99**	**N.A.**
Coeficiente de variação:	**0,11**	**0,13**	**N.A.**

Fonte: autoria própria.

Tabela 6 – Dicionário de dados

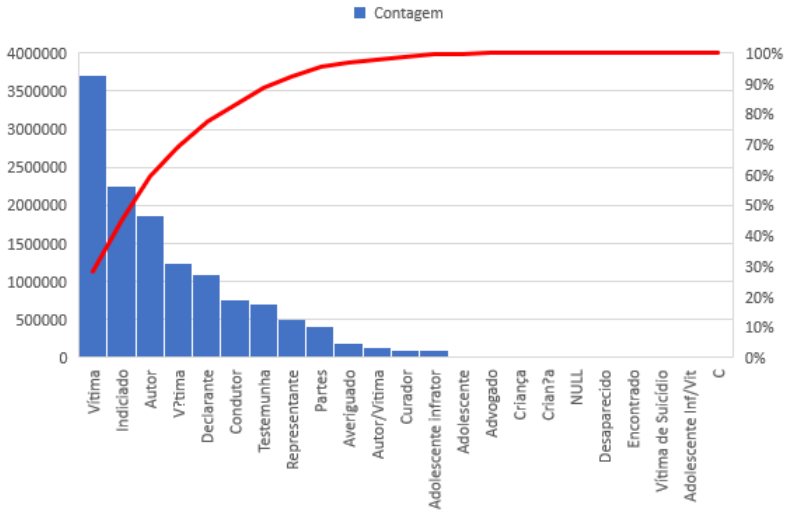
Campos	Descrição	Número da Coluna
ID_DELEGACIA	Código da delegacia responsável pelo registro da ocorrencia	1
NOME_DEPARTAMENTO	Departamento responsável pelo registro	2
NOME_SECCIONAL	Delegacia Seccional responsável pelo registro	3
NOME_DELEGACIA	Delegacia responsávelpelo registro	4
CIDADE	Cidade de Registro	5
ANO_BO	Ano da ocorrencia	6
NUM_BO	Nmero do BO	7
NOME_DEPARTAMENTO_CIRC	Departamento de Circunscrição	8
NOME_SECCIONAL_CIRC	Seccional de Circunscrição	9
NOME_DELEGACIA_CIRC	Delegacia de Circunscrição	10
NOME_MUNICIPIO_CIRC	Municpio de Circunscrição	11
DESCR_TIPO_BO	Tipo de Documento	12
DATA_OCORRENCIA_BO	Data da Ocorrência	13
HORA_OCORRENCIA_BO	Hora da Ocorrência	14
DATAHORA_COMUNICACAO_BO	Data Hora da Comunicação da Ocorrência	15
FLAG_STATUS	Status da Ocorrência	16
RUBRICA	Natureza jurídica da ocorrência	17
DESCR_CONDUTA	Conduta na Ocorrência	18
DESDOBRAMENTO	Desdobramento na Ocorrência	19
DESCR_TIPOLOCAL	Tipo de Local	20
DESCR_SUBTIPOLOCAL	Descrição do subtipo de local	21
LOGRADOURO	Logradouro dos fatos	22
NUMERO_LOGRADOURO	Numero do Logradouro dos fatos	23
LATITUDE	Latitude da Ocorrência	24
LONGITUDE	Longitude da Ocorrência	25
DESCR_TIPO_PESSOA	Qualificação do envolvido na ocorrência	26
FLAG_VITIMA_FATAL	Concondição do Autor / Vítma na ocorrência	27
SEXO_PESSOA	Sexo	28
IDADE_PESSOA	Idade	29
COR_CUTIS	Cor da Pele	30

Fonte: adaptado de (ZANCHI, 2019).

houve a ocorrência. Os campos *SEXO_PESSOA*, *IDADE_PESSOA* e *COR* descrevem características físicas da vítima, já *DESCR_PROFISSAO*, *DESCR_GRAU_INSTRUCAO* assumiu-se que poderiam descrever a posição social da vítima em questão.

Quanto ao campo *DESC_TIPO_PESSOA*, o conjunto possui registros de 25 classes diferentes, sendo que analisaremos apenas os registros da classe "Vítima", que considera vítimas de suicídio, Autor/Vítima e "V?tima"como um campo só. A Figura 12 apresenta a contagem de ocorrência de cada classe no conjunto de dados no campo *DESC_TIPO_PESSOA*.

Figura 12 – Contagem de classes distintas do campo *DESC_TIPO_PESSOA* e sua frequencia relativa acumulada.



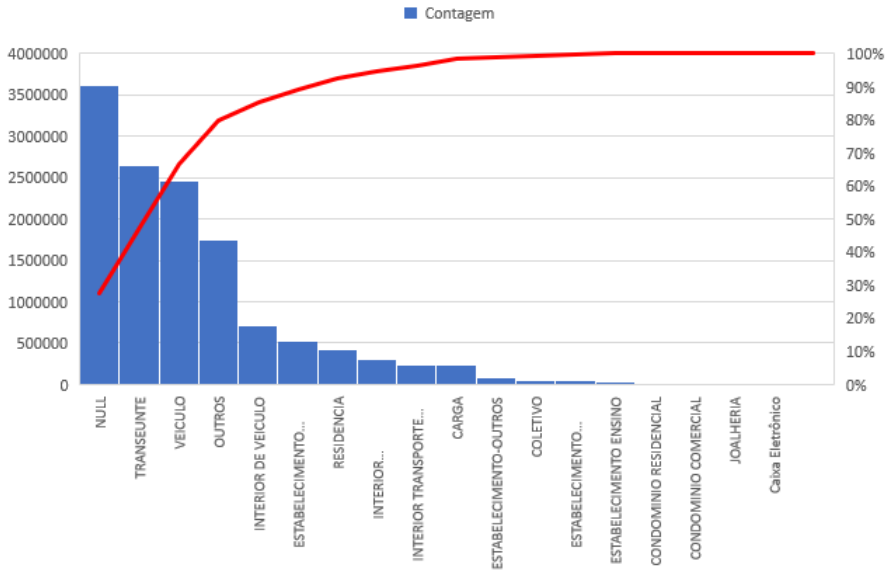
Fonte: autoria própria.

Para o campo *CONDUTA*, foram encontradas 19 classes diferentes. Não houve necessidade de realizar agrupamento devido a erros de digitação ou por considerações de similaridade de classes. A Figura 13 apresenta a contagem de ocorrência de cada classe no campo *CONDUTA*.

No campo *DESC_GRAU_INSTRUCAO*, o conjunto possui registros de 16 classes diferentes, sendo que algumas categorias descrevem aspectos físicos. Esses valores deverão passar por alguma transformação para não comprometer a análise. A Figura 14 apresenta a contagem de ocorrência de cada classe no conjunto de dados no campo *DESC_GRAU_INSTRUCAO*.

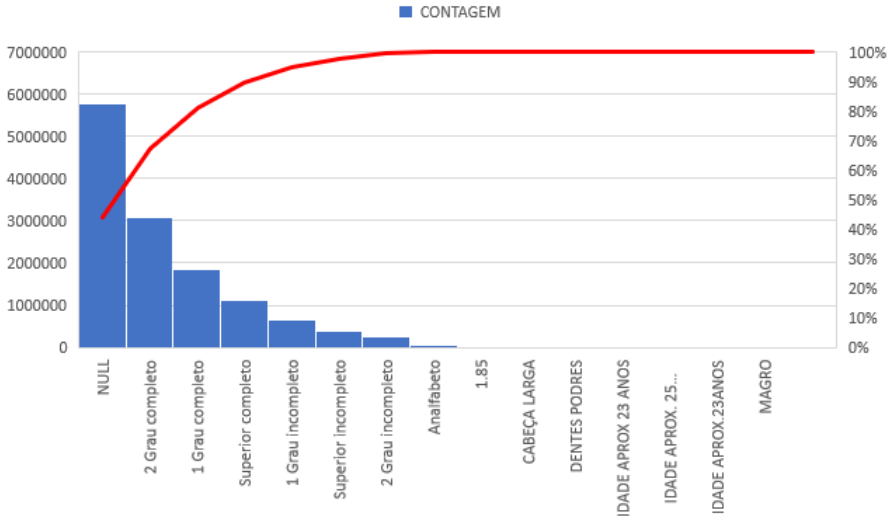
O campo *DESC_PROFISSAO*, por sua vez, apresentou 916 classes distintas. Contudo, a biblioteca que utilizaremos só permite variáveis categóricas de até 53 classes.

Figura 13 – Contagem de classes distintas do campo *CONDUTA* e sua frequencia relativa acumulada.



Fonte: autoria própria.

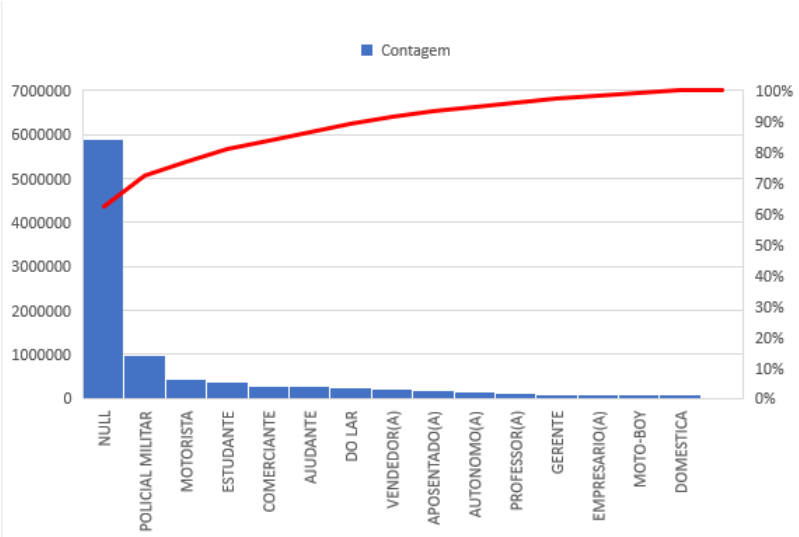
Figura 14 – Contagem de classes distintas do campo *DESC_GRAU_INSTRUCAO* e sua frequencia relativa acumulada.



Fonte: autoria própria.

A Figura 15 apresenta a contagem de ocorrência das 15 classes mais frequentes no campo *DESC_PROFISSAO*.

Figura 15 – As 15 classes mais frequentes no campo *DESC_PROFISSAO* e sua frequencia relativa acumulada.



Fonte: autoria própria.

Quanto ao campo *RUBRICA*, são 66 classes diferentes, contudo muitas delas são divergem uma da outra devido apenas a erros de digitação, como por exemplo nos valores "Les?o corporal seguida de morte (art. 129, ?3o.)"e "Lesão corporal seguida de morte (art. 129, §3o.)", evidenciando assim a necessidade de tratamento para esse campo. A Figura 16 apresenta a contagem de ocorrência das 66 do campo *RUBRICA*.

O campo *SEXO_PESSOA* apresentou 6 valores distintos, sendo 2 passíveis de união por serem duas formas diferentes de se atribuir o valor (*NULL*) ao campo em questão. A Figura 17 apresenta a contagem de ocorrência das 6 do campo *SEXO_PESSOA*.

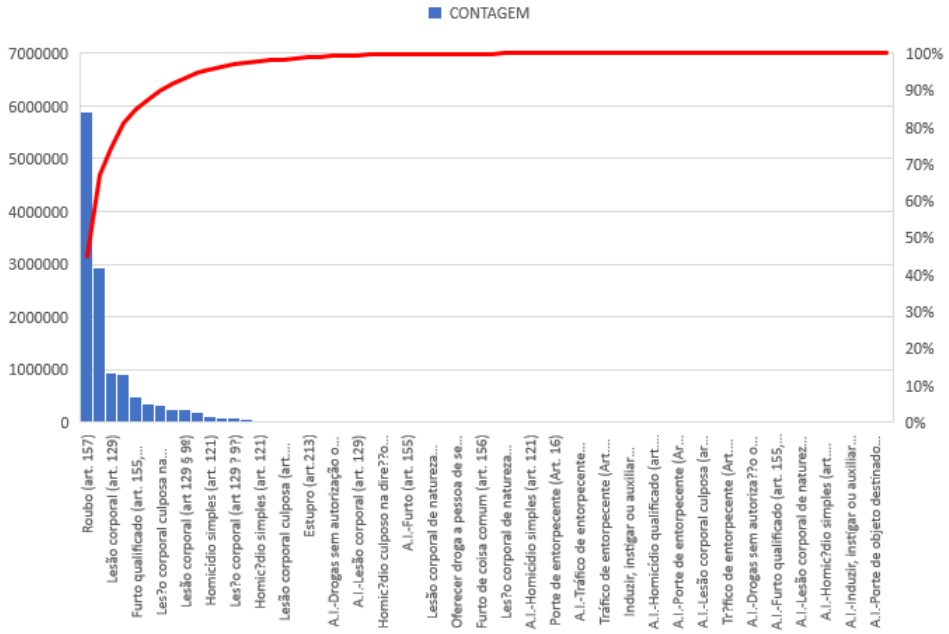
O campo *DATA_OCORRENCIA_BO* possui 2923 valores distintos, sendo 64646 valores *NULL*, e o campo *HORA_OCORRENCIA_BO* possui 1441, sendo todos os minutos do de um período de 24 horas mais o valor *NULL*[1]. Assumiu-se, então, que houveram registros de ocorrências a todo minuto no período avaliado.

O campo *CONT_PESSOA*, cuja interpretação não temos no dicionário de dados, apresentou 133 valores distintos. Os 20 valores mais frequentes são apresentados na Figura 18.

O campo *IDADE_PESSOA* apresentou 250 valores distintos, incluindo idades negativas, sendo -40 anos a menor, valores positivos muito elevados, como 182, 258 e 216 anos. O campo também apresenta texto, como valores "APARENTAVA MENOR DE

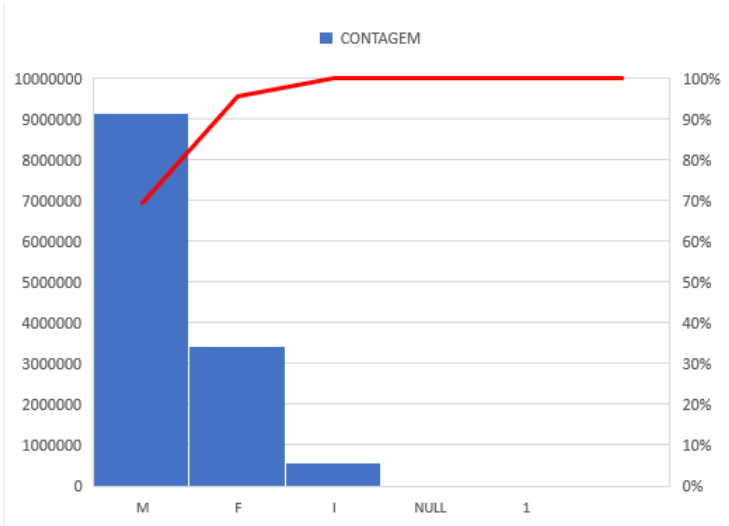
[1]Conta rápida: em um ano, temos aproximadamente 3700000 minutos e a base de dados que estamos avaliando possui 13157011 registros. É uma média de $3,55$ registros de boletins de ocorrência por minuto.

Figura 16 – Contagem de classes do campo *RUBRICA* e sua frequência relativa acumulada.



Fonte: autoria própria.

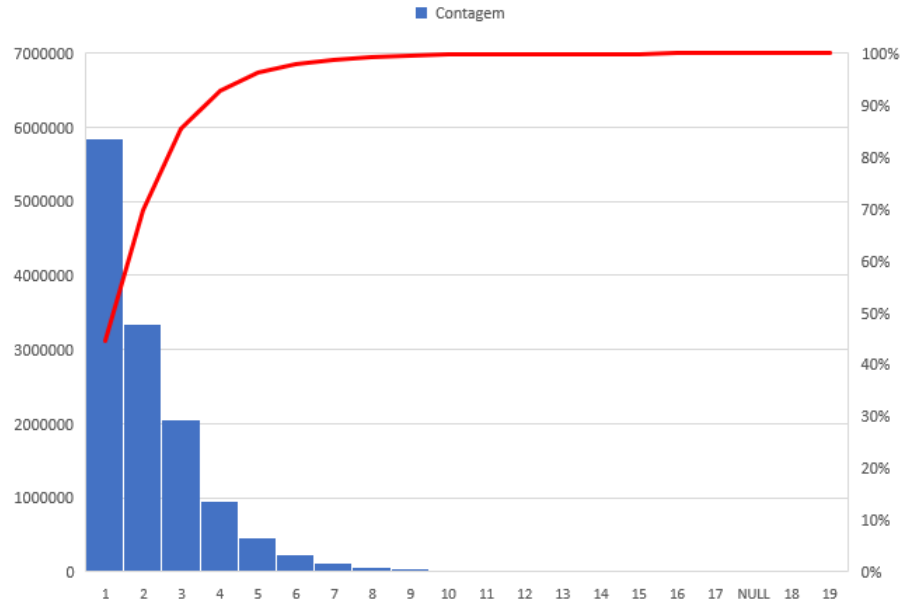
Figura 17 – Contagem de valores do campo *SEXO_PESSOA* e sua frequência relativa acumulada.



Fonte: autoria própria.

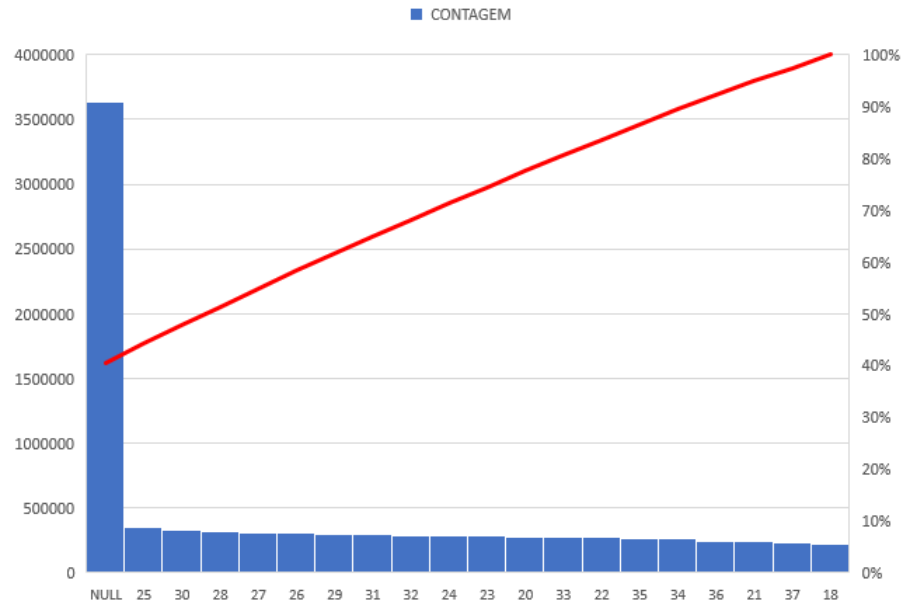
IDADE", e "IDADE APROX.30A", além de valores "*NULL*", que apareceram 3633863 vezes. Pelo fato do campo não ser numérico, não é possível analisar a distribuição das idades por um diagrama do tipo *box-plot*. A figura 19 apresenta a contagem para os 20 valores mais frequentes do campo *IDADE_PESSOA*.

Figura 18 – Contagem de valores do campo *CONT_PESSOA* e sua frequência relativa acumulada.



Fonte: autoria própria.

Figura 19 – Contagem de valores do campo *IDADE_PESSOA* e sua frequência relativa acumulada.



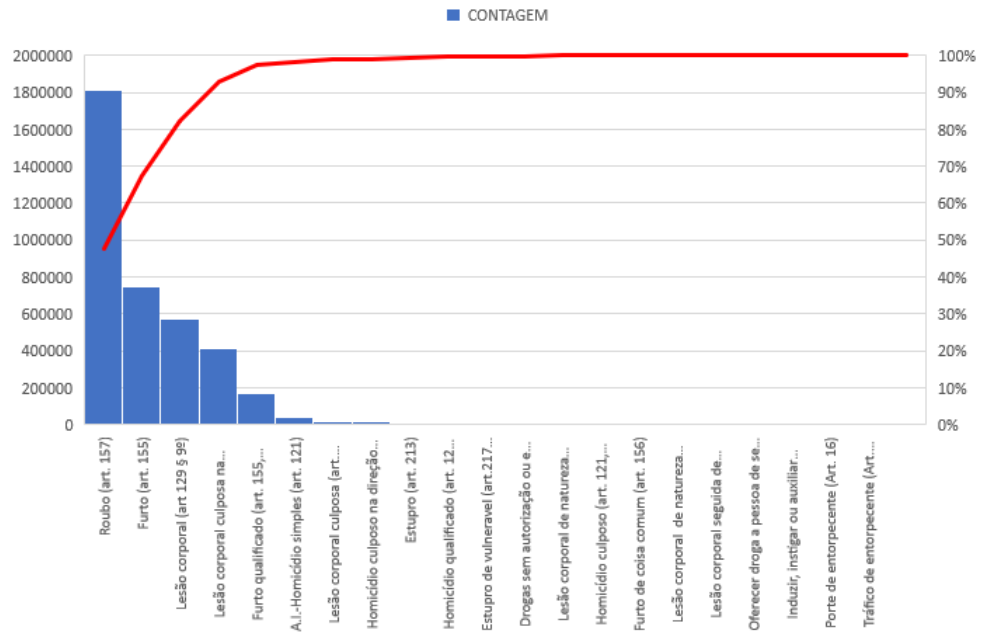
Fonte: autoria própria.

3.2 Extração e transformação dos dados

Para criar o modelo, extraímos os dados do conjunto descrito acima contando com as seguintes considerações:

- Só serão considerados registros onde o valor do campo *DESC_TIPO_PESSOA* contém a sequência de caracteres "tima". Com isso, estaremos considerando todos os tipos de registros de vítimas possíveis.
- No campo *RUBRICA*, corrigiram-se os erros de digitação. A Figura 20 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

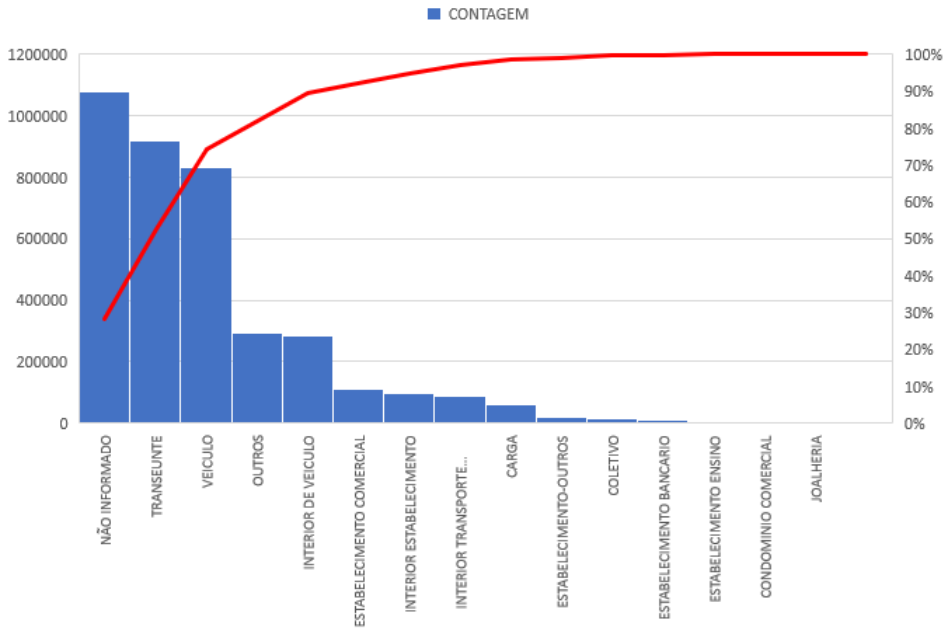
Figura 20 – Contagem de valores do campo *RUBRICA* e sua frequência relativa acumulada após a transformação do campo.



Fonte: autoria própria.

- No campo *CONDUTA*, valores *NULL* foram transformados em *NÃO INFORMADO*. Os valores *RESIDENCIA* e *Caixa Eletrônico* não apareceram no caso em que o campo *DESC_TIPO_PESSOA* apresentou o valor de "Vítima"após o tratamento. A Figura 21 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

Figura 21 – Contagem de valores do campo *CONDUTA* e sua frequência relativa acumulada após a transformação do campo.



Fonte: autoria própria.

- O campo *CONT_PESSOA* foi definido em 5 valores diferentes: *1*, *2 A 3*,*4 A 5*, *MAIS QUE 5*, *NÃO INFORMADO*. A Figura 22 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

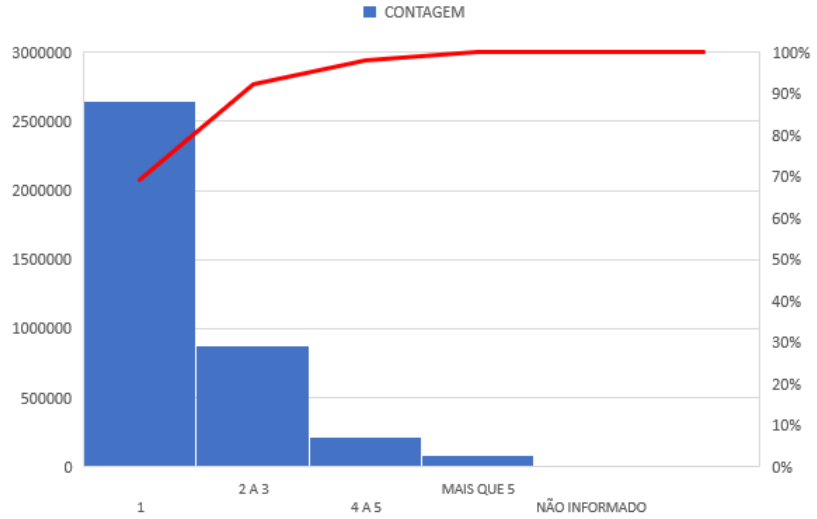
- O campo *DATA_OCORRENCIA_BO* foi transformado em *DIA_DA_SEMANA*, tendo como valores os sete dias da semana e os valores *NULL* foram descartados. A Figura 23 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

- O campo *HORA_OCORRENCIA_BO* foi transformado em *PERIODO_OCORRENCIA*.

  Os valores permitidos para *PERIODO_OCORRENCIA* são *MADRUGADA 1*, *MADRUGADA 2*, *MADRUGADA 3*, *MANHÃ 1*, *MANHÃ 2*, *MANHÃ 3*, *TARDE 1*, *TARDE 2*, *TARDE 3*, *NOITE 1*, *NOITE 2* e *NOITE 3* e *NÃO INFORMADO*.

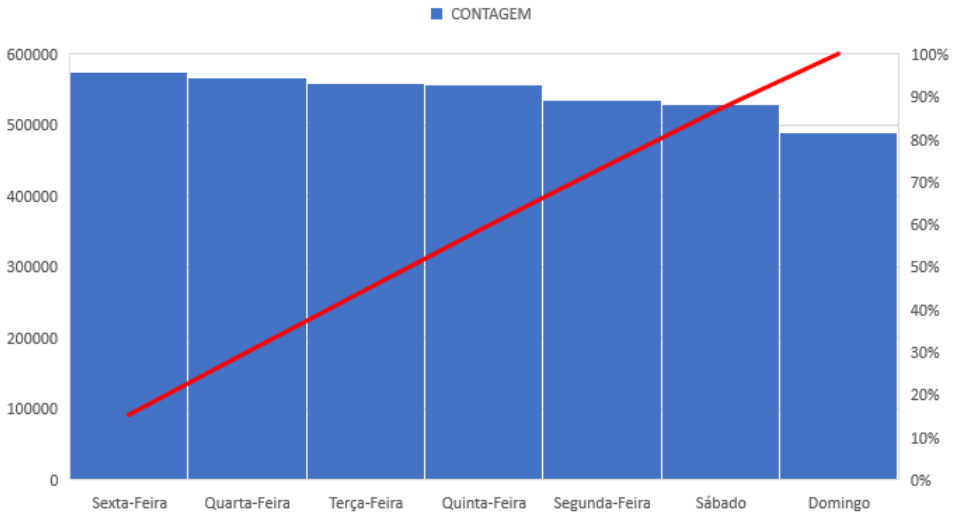
  Cada valor do campo *PERIODO_OCORRENCIA* representa um intervalo de 01 : 59 : 59, começando por *MADRUGADA 1* (00 : 00 : 00 à 01 : 59 : 59, inclusive), seguido por *MADRUGADA 1* (02:00:00 à 03:59:59, inclusive), e assim sucessivamente. O valor *NÃO INFORMADO* contempla outros valores digitados errados ee valores

Figura 22 – Contagem de valores do campo *CONT_PESSOA* e sua frequência relativa acumulada após a transformação do campo.



Fonte: autoria própria.

Figura 23 – Contagem de valores do campo *DIA_DA_SEMANA* e sua frequência relativa acumulada após a transformação do campo.

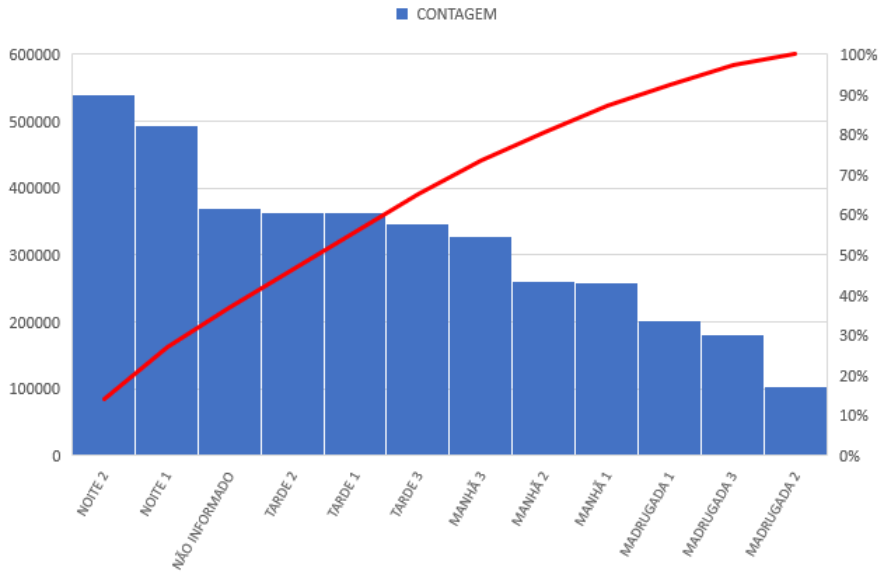


Fonte: autoria própria.

NULL. A Figura 24 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

- O campo *IDADE_PESSOA* transformado no campo *FAIXA_ETARIA*.

Figura 24 – Contagem de valores do campo *PERIODO_OCORRENCIA* e sua frequência relativa acumulada após a transformação do campo.



Fonte: autoria própria.

Os valores possíveis são *0 A 10 ANOS*, *11 À 20 ANOS*, *11 A 20 ANOS*, *21 A 30 ANOS*, *31 A 40 ANOS*, *41 A 50 ANOS*, *51 A 60 ANOS*, *61 A 70 ANOS*, *71 A 80 ANOS*, *MAIOR QUE 80 ANOS* e demais valores fora dessas faixas etárias foram atribuidos à *NÃO INFORMADA*. A Figura 25 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

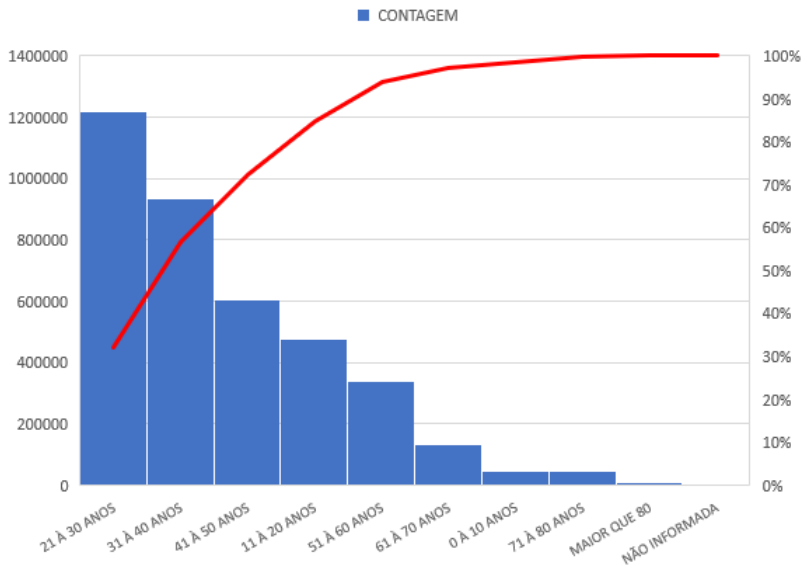
- *DESCR_PROFISSAO*: foram selecionados os 46 valores mais frequentes.

  Os demais 870 valores distintos foram alocados em *OUTRO COM 2 GRAU COMPLETO*, *OUTRO - NÃO INFORMADO*, *OUTRO COM 1 GRAU COMPLETO*, *OUTRO COM SUPERIOR COMPLETO*, *OUTRO COM 1 GRAU INCOMPLETO* e *OUTRO - ANALFABETO*. Esses valores levam em conta o campo *DESC_GRAU_INSTRUCAO* e os valores de *DESCR_PROFISSAO* que não estão entre os primeiros 46 mais frequentes. A Figura 26 apresenta a contagem dos valores e sua respectiva frequência relativa acumulada após ta transformação do campo.

3.3 Construção do modelo usando o software *R*

Primeiro, fazemos o carregamento dos pacotes. O pacote *RODBC* é responsável por fazer a conexão do *software R* com o banco de dados no *SQL Server*. O pacote

Figura 25 – Contagem de valores do campo *FAIXA_ETARIA* e sua frequência relativa acumulada após a transformação do campo.



Fonte: autoria própria.

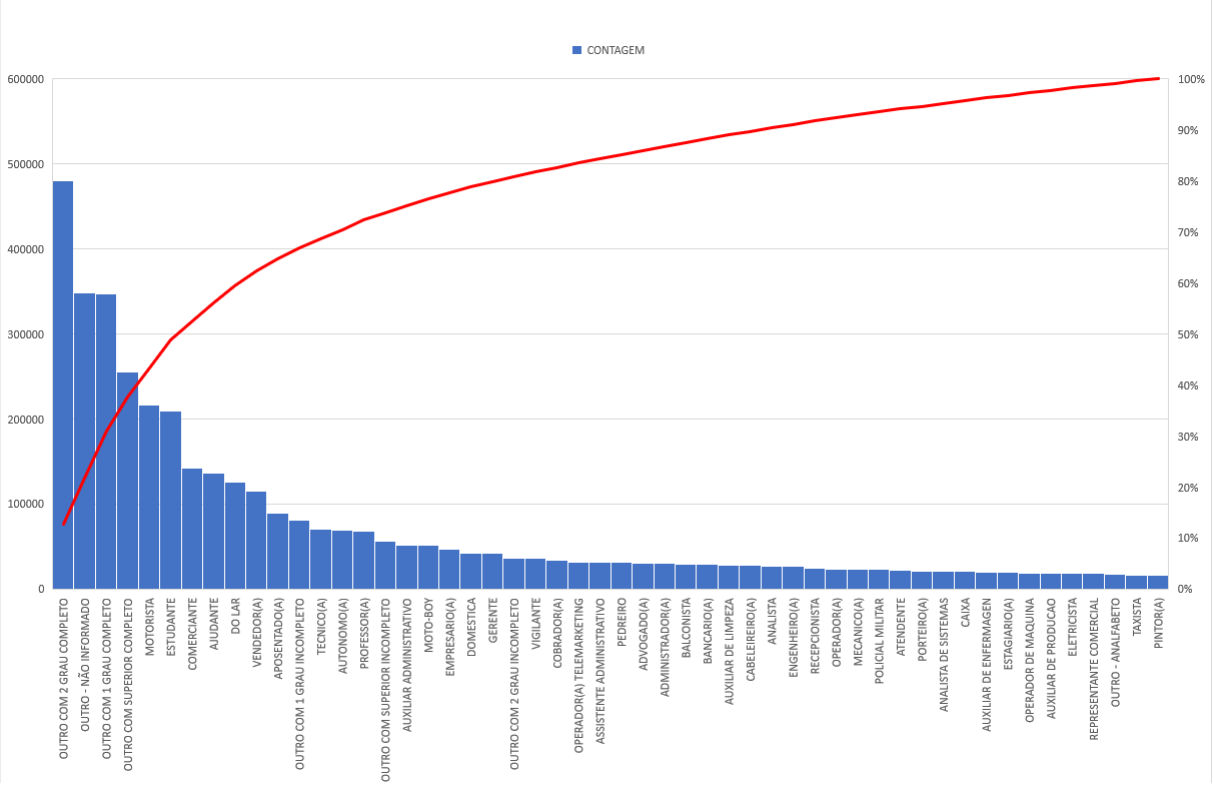
dplyr, por sua vez, serve para realizar operações em tabelas como agrupamentos ou fazer encadeamento de operações que agirão sobre uma tabela. O pacote *randomForest* contém o algoritmo que criará o modelo que desejamos avaliar, bem como nos fornece o valor do decrescimento médio de Gini para as variáveis e o erro *OOB* do nosso modelo. Já os pacotes *caret* e *irr* serão usados para, respectivamente, criar o gráfico para avaliação de importância de variável e avaliação do coeficiente Kappa de Fleiss.

```
library(RODBC)
library(dplyr) # Pacote para manusear os dados.
library(randomForest) # Pacote para criar o modelo
library(caret)
library(irr)
```

Em seguida, criamos um objeto do tipo conexão com banco de dados *SQL*, que aqui foi denominado de *cn*. O objeto *tccDF* logo em seguida é a tabela que consta todos os dados tratados que apresentamos na seção anterior.

```
cn <- odbcDriverConnect(connection="Driver={SQLServer};
server=DESKTOP-FL71F1C;
database=tccDB;
trusted_connection=yes;")
```

Figura 26 – Contagem de valores do campo *DESCR_PROFISSAO* e sua frequência relativa acumulada após a transformação do campo.



Fonte: autoria própria.

```
tccDF = sqlQuery(cn, 'SELECT * FROM [tccDB].[dbo].[vw_TCC]')
```

Devido a limitação computacional, não conseguimos utilizar toda a tabela armazenada no objeto *tccDF*. Para contornar esse problema, fazemos uma amostragem de 5%, o que significa 190525 registros da base. Essa quantidade de registros foi limitada à capacidade computacional do equipamento utilizado para a construção do modelo. A amostragem foi feita de forma aleatória e com reposição a fim de conservar a probabilidade de se extrair um determinado registro.

```
sampleTCCDF <- tccDF %>%
  group_by(RUBRICA) %>%
  sample_frac(0.05, replace = TRUE) %>%
  ungroup
View(sampleTCCDF)
```

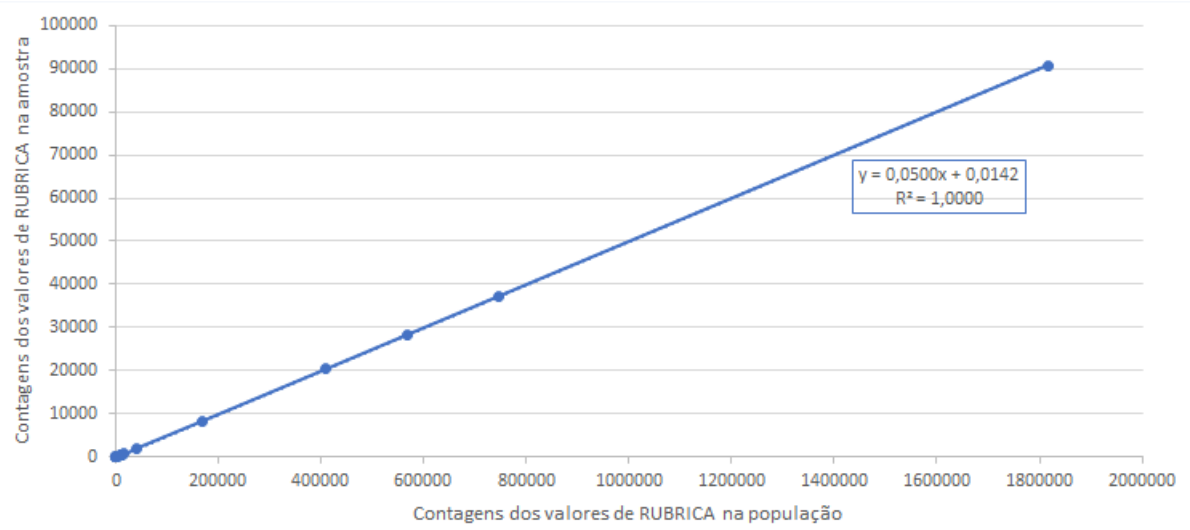
A amostragem não alterou de forma significativa a proporção entre as contagem dos valores da variável *RUBRICA* em relação à população. Isso pode ser observado na Tabela 7 e através da relação linear entre as duas quantidades exibidas na Figura 27.

Tabela 7 – Relação entre os dados da população e os dados presentes na amostra

Valor de RUBRICA	Contagem população	Contagem amostra	Razão amostra/população
Roubo (art. 157)	1815787	90789	0,050
Furto (art. 155)	746430	37322	0,050
Lesão corporal (art 129 § 9º)	568660	28433	0,050
Lesão corporal culposa na direção de veículo automotor (Art. 303)	408814	20441	0,050
Furto qualificado (art. 155, §4o.)	168544	8427	0,050
A.I.-Homicídio simples (art. 121)	41223	2061	0,050
Lesão corporal culposa (art. 129. §6o.)	15531	777	0,050
Homicídio culposo na direção de veículo automotor (Art. 302)	13202	660	0,050
Estupro (art. 213)	10749	537	0,050
Homicídio qualificado (art. 121, §2o.)	8195	410	0,050
Estupro de vulneravel (art.217-A)	8154	408	0,050
Drogas sem autorização ou em desacordo (Art.33, caput)	1898	95	0,050
Lesão corporal de natureza GRAVE (art. 129, §1o.)	1662	83	0,050
Homicídio culposo (art. 121, §3o.)	883	44	0,050
Furto de coisa comum (art. 156)	309	15	0,049
Lesão corporal de natureza GRAVÍSSIMA (art. 129, §2o.)	190	10	0,053
Lesão corporal seguida de morte (art. 129, §3o.)	152	8	0,053
Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)	65	3	0,046
Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)	26	1	0,038
Porte de entorpecente (Art. 16)	16	1	0,063
Tráfico de entorpecente (Art. 12)	8	0	0,000

Fonte: autoria própria.

Figura 27 – Relação entre as contagens dos valores de *RUBRICA* na amostra e na população.



Fonte: autoria própria.

Contudo, nesse processo é possível que alguns valores da variável *RUBRICA* perdemos o valor *Tráfico de entorpecente (Art. 12)*, então foi preciso redefinir os fatores associados aos valores dessa variável em questão, o que é feito com a sequência de comandos a seguir:

```
sampleTCCDF$RUBRICA = as.character(sampleTCCDF$RUBRICA)
sampleTCCDF$RUBRICA = as.factor(sampleTCCDF$RUBRICA)
```

Em seguida, definimos a partição de treino e de teste do modelo. Apesar de não ser necessário, devido à natureza do modelo de floresta aleatória, faremos essa separação e usar 80% dos registros do objeto *sampleTCCDF* e os 20% restantes serão utilizados para a avaliação do modelo com o índice Kappa de Fleiss.

```
set.seed(24165)
trainIndexSample = createDataPartition(sampleTCCDF$RUBRICA,
p = 0.80,
list = FALSE,
times = 1)

trainSample = sampleTCCDF[trainIndexSample,]
testSample = sampleTCCDF[-trainIndexSample,]
```

Então, criamos o modelo de floresta aleatória, onde a variável que desejamos ter como alvo é *RUBRICA* e todas as demais variáveis da tabela *tranSample* serão utilizadas para prevê-la. O número de árvores de decisão foi de 50:

```
modelRF = randomForest(trainSample$RUBRICA ~.,
data=trainSample,
ntrees = 500,
importance = TRUE)
```

Para obter avaliar a importância das variáveis, obtemos uma tabela com o decrescimento médio de Gini de cada variável a partir do seguinte comando:

```
importance(modelRF, type = 2)
```

Para visualizar o mesmo resultado de forma gráfica, utilizamos:

```
varImpPlot(modelRF, type = 2)
```

Em seguida, utilizamos o modelo e os dados de teste para fazer as predições da variável *RUBRICA*.

```
predict = predict(modelRF,testSample)

result = data.frame(predicted =  as.character(predict),
observed = as.character(testSample$RUBRICA))
```

O objeto *predict* armazena o resultado da predição realizada pelo modelo. O objeto *result* é um objeto que armazena o resultado da da predição na coluna *predicted* e o valor da variável *RUBRICA* real da partição de teste na coluna *observed*.

Em seguida, iniciamos a avaliação do modelo. Primeiro, criamos o objeto *predictRight*, que armazena todos os valores previstos pelo modelo (*predicted*) e que são iguais ao valor da partição de teste (*observed*). Em seguida, criamos o objeto *accuracy*, que é a razão entre o número de elementos previstos corretamente pelo modelo pelo número total de elementos na partição de teste.

```
predictRight = filter(result,
as.character(predicted) == as.character(observed))

accuracy = nrow(predictRight)/nrow(testSample)
print(accuracy)
```

Já para obter o erro *OOB*, executamos o seguinte comando:

```
print(modelRF)
```

E, por fim, para obter o índice Kappa de Fleiss, utilizamos o objeto *result* e a função *kappam.fleiss* do pacote *irr*:

```
kappam.fleiss(result, detail = TRUE)
```

4 RESULTADOS E DISCUSSÃO

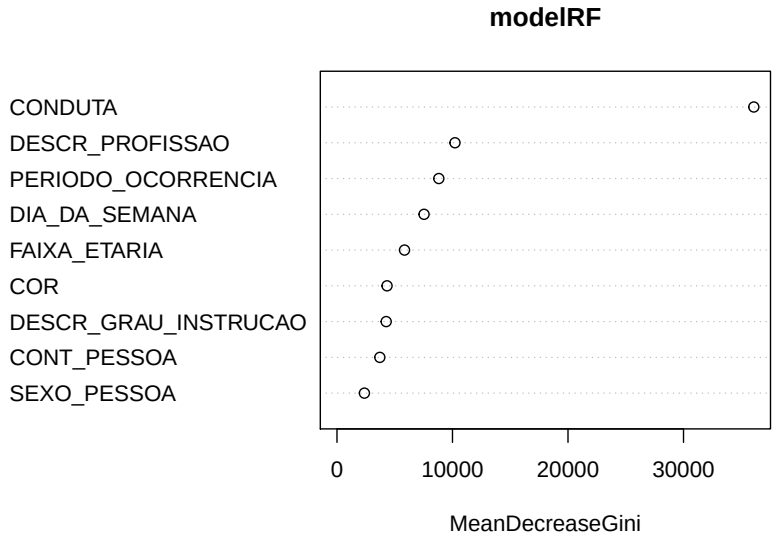
A Tabela 8, obtida através do comando *importance(modelRF, type = 2)* apresenta os valores de decrescimento médio de Gini para as variáveis utilizadas no modelo. A Figura 28, obtido através do comando *varImpPlot(modelRF, type = 2)*, apresenta os mesmos valores, mas de forma gráfica.

Tabela 8 – Decrescimento médio de Gini das variáveis utilizadas no modelo.

Variável	**Decrescimento médio de Gini**
CONDUTA	36101,411
DESCR_PROFISSAO	10189,819
PERIODO_OCORRENCIA	8747,719
DIA_DA_SEMANA	7562,898
FAIXA_ETARIA	5877,945
COR	4367,054
DESCR_GRAU_INSTRUCAO	4222,093
CONT_PESSOA	3737,556
SEXO_PESSOA	2454,288

Fonte: autoria própria.

Figura 28 – Contagem de classes distintas do campo *CONDUTA* e sua frequencia relativa acumulada.



Fonte: autoria própria.

O resultado sugere que o campo *CONDUTA* é a variável mais importante para a construção do modelo e a menos importante é a variável *SEXO_PESSOA*. Segundo esse resultado, o modelo sugere que o tipo do local em que uma pessoa está é o que influencia

mais no tipo de crime que ela irá sofrer. A importância da variável também pode mudar de acordo com o valor de *RUBRICA*, como pode ser visto no anexo. Por exemplo, no caso em que *RUBRICA = Estupro (art. 213)*, a variável *CONDUTA* permaneceu como a mais importante, com índice igual a $46,124116$ e a variável *SEXO_PESSOA* foi a segunda mais importante, com índice igual a $38,985391$.

O erro *OOB* obtido foi de $30,26\%$. Isso significa que o modelo erra $30,26\%$ das tentativas e acerta as outras $69,74\%$, concordando com o valor da razão entre o número de acertos e o número total de tentativas de predição *accuracy*. Mas a variável que desejamos prever possui 21 valores diferentes e com frequências diferentes. Então, fazemos uso de uma matriz de confusão para calcular a taxa de erro de previsão para cada valor da variável *RUBRICA*. O resultado da matriz de confusão é apresentado na Tabela 9:

Tabela 9 – Erros de predição segundo a matriz de confusão.

Valor de RUBRICA	**Erro**
Roubo (art. 157)	0.1193000
Lesão corporal (art 129 § 9º)	0.2326461
Lesão corporal culposa na direção de veículo automotor (Art. 303)	0.3672109
Furto (art. 155)	0.5682564
Furto qualificado (art. 155, §4o.)	0.7770691
Drogas sem autorização ou em desacordo (Art.33, caput)	0.9210526
A.I.-Homicídio simples (art. 121)	0.9484536
Estupro de vulneravel (art.217-A)	0.9602446
Homicídio culposo na direção de veículo automotor (Art. 302)	0.9753788
Lesão corporal culposa (art. 129. §6o.)	0.9823151
Homicídio qualificado (art. 121, §2o.)	0.9878049
Estupro (art. 213)	0.9976744
Furto de coisa comum (art. 156)	1.0000000
Homicídio culposo (art. 121, §3o.)	1.0000000
Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)	1.0000000
Lesão corporal de natureza GRAVÍSSIMA (art. 129, §2o.)	1.0000000
Lesão corporal de natureza GRAVE (art. 129, §1o.)	1.0000000
Lesão corporal seguida de morte (art. 129, §3o.)	1.0000000
Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)	1.0000000
Porte de entorpecente (Art. 16)	1.0000000

Fonte: autoria própria.

A Tabela 9 diz que o modelo apresenta melhores predições para os casos em que o valor de *RUBRICA* apresenta os valores de *Roubo (art. 157)* (erro de $11,93\%$), *Lesão corporal (art 129 § 9o)* (erro de $23,26\%$) e *Lesão corporal culposa na direção de veículo automotor (Art. 303)* (erro de $36,72\%$). Para os valores *Furto (art. 155)* e *Furto qualificado (art. 155, §4o.)*, com erros respectivos de $56,82\%$ e $77,70\%$, o modelo acerta poucas vezes. Para os demais valores possíveis de *RUBRICA*, o modelo praticamente não

acerta a previsão.

Era esperado que o valor de *Roubo (art. 157)* apresentasse maior taxa de acerto devido ao maior número de ocorrências em relação aos outros valores, como podemos verificar pela Tabela 7. Contudo, os valores de *Lesão corporal (art 129 § 9o)* e *Lesão corporal culposa na direção de veículo automotor (Art. 303)*, apesar de menos frequentes, foram previstos com maior taxa de acerto do que o valor *Furto (art. 155)*, mesmo esses dois últimos valores serem menos frequentes na amostra. Ao calcularmos a impurezas de Gini para cada variável preditora, através da equação (eq:impureza-gini) e extrairmos a média desse valor, veremos que quando *RUBRICA* apresenta o valor *Furto (art. 155)*, teremos grupos mais heterogêneos do que quando *RUBRICA* assume o valor de *Lesão corporal (art 129 § 9o)* e *Lesão corporal culposa na direção de veículo automotor (Art. 303)*. Os valores obtidos foram $0,6893$ para *Furto (art. 155)*, $0,6483$ para *Lesão corporal (art 129 § 9o)* e $0,6499$ para *Lesão corporal culposa na direção de veículo automotor (Art. 303)*. As tabelas e os valores utilizados para chegar nesses resultados estão nas Tabelas 11,12 e 13 anexas. Ainda se considerarmos o Índice de concordância Kappa de Fleiss da Tabela 10, veremos que a atribuição para o valor *Lesão corporal (art 129 § 9o)* é atribuido à variável *RUBRICA* de forma menos aleatória que o valor *Lesão corporal culposa na direção de veículo automotor (Art. 303)*.

Tabela 10 – Valores do índice Kappa de Fleiss e respectivos z_{scores} e valores-p para *RUBRICA*.

Valor de RUBRICA	**Kappa**	**z-score**	**Valor-p**
Lesão corporal (art 129 § 9º)	0,676	131,912	0,000
Roubo (art. 157)	0,638	124,485	0,000
Lesão corporal culposa na direção de veículo automotor (Art. 303)	0,567	110,574	0,000
Lesão corporal de natureza GRAVÍSSIMA (art. 129, §2o.)	0,500	97,588	0,000
Furto (art. 155)	0,392	76,450	0,000
Furto qualificado (art. 155, §4o.)	0,244	47,687	0,000
Drogas sem autorização ou em desacordo (Art.33, caput)	0,182	35,442	0,000
A.I.-Homicídio simples (art. 121)	0,075	14,704	0,000
Estupro de vulneravel (art.217-A)	0,059	11,470	0,000
Lesão corporal culposa (art. 129. §6o.)	0,047	9,121	0,000
Homicídio culposo na direção de veículo automotor (Art. 302)	0,040	7,898	0,000
Estupro (art. 213)	0,017	3,272	0,001
Lesão corporal de natureza GRAVE (art. 129, §1o.)	0,000	-0,041	0,967
Homicídio culposo (art. 121, §3o.)	0,000	-0,023	0,982
Furto de coisa comum (art. 156)	0,000	-0,008	0,994
Lesão corporal seguida de morte (art. 129, §3o.)	0,000	-0,003	0,998
Homicídio qualificado (art. 121, §2o.)	-0,001	-0,221	0,825

Fonte: autoria própria.

Para medir a concordância entre os valores previstos pelo modelo e os valores reais, considerando a possibilidade da concordância ter ocorrido ao acaso, calculamos o índice de concordância Kappa de Fleiss. O resultado obtido está registrado na Tabela 10 ,

em conjunto com seus respectivos z_{scores} e valores-p. Os valores de *RUBRICA Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)*, *Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)* e *Porte de entorpecente (Art. 16)* não aparecem na Tabela 10 pelo fato do modelo não ter previsto esses valores nenhuma vez.

5 CONCLUSÃO

O modelo apresentou uma taxa de acerto de $69,74\%$ durante o teste e as métricas utilizadas se mostraram coerentes. Contudo, a taxa de acerto se mostrou elevada principalmente por que o modelo apresentou taxas de acerto elevadas para os valores mais frequentes da variável alvo, sendo isso um indício de que o modelo está enviesado.

O modelo criado neste trabalho conseguiu acertar mais da metade das vezes as previsões da natureza jurídica do crime apenas para os casos de *Roubo (art. 157)*, *Lesão corporal (art 129 § 9º)* e *Lesão corporal culposa na direção de veículo automotor (Art. 303)*, onde o erro apresentado na matriz de confusão foram de $0,119$, $0,233$ e $0,367$, respectivamente. Os casos de *Roubo (art. 157)* e *Lesão corporal (art 129 § 9º)* também apresentaram bons resultados ao calcular o índice Kappa de Fleiss, $0,638$ e $0,676$ resectivamente, onde o modelo apresentou concordância considerável com os valores reais.

Outros valores como *Furto (art. 155)* e *Furto qualificado (art. 155, §4o.)* apresentaram desempenho de concordância razoável entre o modelo e os dados reais, com erro de $0,568$ e $0,777$ segundo a matriz de confusão e Kappa de Fleiss iguais a $0,392$ e $0,244$, respectivamente, havendo desempenho de concordância razoável. Para os casos em que o erro da matriz de confusão foi maior que o do valor de *Drogas sem autorização ou em desacordo (Art.33, caput)*, $0,921$ e Kappa de Fleiss $0,182$ (concordância leve), o modelo não desempenha previsão.

Para a importância das variáveis nesse modelo constatou-se que o local onde o crime ocorre (*CONDUTA*) é mais importante para determinar o tipo de crime que a pessoa sofrerá antes de registrar a ocorrencia, com o decrescimento médio de Gini calculado pelo *R* igual a $36101,411$ e a variável de menor importância foi a *SEXO_PESSOA*, com valor de *2454,288*. Isso significa que a divisão dos nós relacionados à variável *CONDUTA* leva a grupos com mais homogeneidade de forma mais rápida que as demais variáveis, isso não significa que no dia-a-dia o tipo do local influencia mais para que uma pessoa sofra qualquer tipo de crime.

É importante ressaltar que o aqui modelo foi construído com dados de pessoas que sofreram algum tipo de crime e registraram o boletim de ocorrência. No que se refere aos dados, o modelo poderia ser melhorado caso considerássemos alguma métrica de não-ocorrência de crime, pois o modelo aqui elaborado sempre compreenderá que uma registro estará sujeito a algum tipo de crime presente na variável *RUBRICA*, e não considera a possibilidade dela não ser vítima de crime algum. Quanto ao viés do modelo em favor da previsão dos valores mais frequentes da variável alvo, poderia-se fazer a amostragem de forma melhor balanceada ou fazer uso de outras técnicas computacionais, como sobreamostragem dos valores menos favorecidos.

REFERÊNCIAS

Amazon. *O que é big data?* 2020. Disponível em: <https://aws.amazon.com/pt/big-data/what-is-big-data/>. Citado na página 18.

ANYOHA, R. 2017. Disponível em: <http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>. Citado 2 vezes nas páginas 17 e 18.

AWAD, M.; KHANNA, R. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. 2015. E-Book. Disponível em: <https://www.amazon.com.br/Efficient-Learning-Machines-Applications-Engineers-ebook/dp/B01J9UUO7A/ref=sr_1_1?___mk_pt_BR=ÅMÅŽ~O~N&keywords=EfficientLearningMachines:Theories,Concepts,andApplicationsforEngineersandSystemDesigners&qid=1584316742&s=digital-text&sr=1-1>. Citado 2 vezes nas páginas 23 e 24.

AZIMI, S. M. et al. Advanced steel microstructural classification by deep learning methods. *Scientific Reports*, Springer Science and Business Media LLC, v. 8, n. 1, feb 2018. Citado na página 18.

BROWNLEE, J. *Parametric and Nonparametric Machine Learning Algorithms*. 2016. Disponível em: <https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/>. Citado 2 vezes nas páginas 21 e 28.

BROWNLEE, J. *Supervised and Unsupervised Machine Learning Algorithms*. 2016. Disponível em: <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>. Citado na página 22.

BRUCE, P.; BRUCE, A. *Estatística prática para cientistas de dados*. [S.l.]: Alta Books, 2019. Citado 9 vezes nas páginas 22, 23, 28, 29, 30, 31, 36, 39 e 40.

Data Science Brigade. *A Diferença Entre Inteligência Artificial, Machine Learning e Deep Learning*. 2016. Disponível em: <https://medium.com/data-science-brigade/a-diferença-entre-inteligência-artificial-machine-learning-e-deep-learning-930b5cc2aa42>. Citado na página 22.

Elite Data Science. 2020. Disponível em: <https://elitedatascience.com/overfitting-in-machine-learning>. Citado na página 28.

FERREIRA, J. C.; PATINO, C. M. What does the p value really mean? *Jornal Brasileiro de Pneumologia*, FapUNIFESP (SciELO), v. 41, n. 5, p. 485–485, oct 2015. Citado na página 39.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, American Psychological Association (APA), v. 76, n. 5, p. 378–382, 1971. Citado 4 vezes nas páginas 33, 34, 35 e 36.

GAMER, M. *IRR package*. 2019. Disponível em: <https://www.rdocumentation.org/packages/irr/versions/0.84.1/source>. Citado na página 40.

GARBADE, M. J. *Clearing the Confusion: AI vs Machine Learning vs Deep Learning Differences*. 2018. Disponível em: <https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb>. Citado na página 21.

GÉRON, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly UK Ltd., 2019. ISBN 1492032646. Disponível em: <https://www.ebook.de/de/product/33315532/aurelien_geron_hands_on_machine_learning_with_scikit_learn_keras_and_tensorflow.html>. Citado 2 vezes nas páginas 22 e 25.

GIORDANO, T. *Netflix, Hadoop and big data*. 2019. Disponível em: <https://www.ibm.com/blogs/services/2019/05/22/netflix-hadoop-and-big-data/>. Citado na página 18.

HARTSHORN, S. *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*. 2016. Disponível em: <https://www.amazon.com.br/Machine-Learning-Random-Forests-Decision-ebook/dp/B01JBL8YVK/ref=sr_1_1?___mk_pt_BR=ÅMÅŽ~O~N&keywords=randomforest&qid=1583371378&sr=8-1>. Citado 3 vezes nas páginas 27, 28 e 32.

HASTLE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *he Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. [s.n.], 2009. Disponível em: <https://www.springer.com/gp/book/9780387848570>. Citado na página 28.

JANSEN, S. *Hands-on machine learning for algorithmic trading : design and implement investment strategies based on smart algorithms that learn from data using Python*. Birmingham, UK: Packt Publishing, 2018. ISBN 9781789346411. Citado na página 29.

KANIOURA, A.; EITEL-PORTER, R. *What is AI exacly?* 2020. Disponível em: <https://www.accenture.com/hk-en/insights/artificial-intelligence/what-ai-exactly>. Citado 2 vezes nas páginas 17 e 21.

KAPOOR, A. *Deep Learning vs. Machine Learning: A Simple Explanation*. 2019. Disponível em: <https://hackernoon.com/deep-learning-vs-machine-learning-a-simple-explanation-47405b3eef08>. Citado na página 22.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, JSTOR, v. 33, n. 1, p. 159, mar 1977. Citado na página 34.

LAURETTO, M. de S.; BASTOS, D. G.; NASCIMENTO, P. S. *Análise Empírica de Desempenho de Quatro Métodos de Seleção de Características para Random Forests*. 2014. Revista Brasileira de Sistemas de Informação. Revista Brasileira de Sistemas de Informação. Disponível em: <http://www.seer.unirio.br/index.php/isys/article/view/3309>. Citado na página 29.

LE, J. *Decision Trees in R*. 2018. Disponível em: <https://www.datacamp.com/community/tutorials/decision-trees-R>. Citado 2 vezes nas páginas 23 e 24.

LI, M. et al. Predicting the epidemic trend of COVID-19 in china and across the world using the machine learning approach. Cold Spring Harbor Laboratory, mar 2020. Citado na página 18.

LIBERMAN, N. 2017. Disponível em: <https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991>. Citado na página 28.

MACHETTI, K. *Alan Turing e a Enigma*. 2016. Disponível em: <http://horizontes.sbc.org.br/index.php/2016/11/22/alan-turing-e-a-enigma/>. Citado na página 17.

MAGALHÃES, M. N.; LIMA, A. C. P. de. *Noções de probabilidade e estatística*. São Paulo: EDUSP, 2004. ISBN 9788531406775. Citado 6 vezes nas páginas 36, 37, 38, 39, 40 e 41.

MASON, W.; VAUGHAN, J. W.; WALLACH, H. Computational social science and social computing. *Machine Learning*, Springer Science and Business Media LLC, v. 95, n. 3, p. 257–260, nov 2013. Citado na página 18.

MINITAB. *Estatísticas kappa para análise de concordância por atributos*. 2020. Disponível em: <https://support.minitab.com/pt-br/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/how-to/attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/kappa-statistics/>. Citado na página 33.

O'LEARY, D. E. Artificial inteligence and big data. *A.I. Innovation in Industry*, abr. 2013. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6547979>. Citado na página 17.

Oracle. *O que é big data?* 2020. Disponível em: <https://www.oracle.com/br/big-data/guide/what-is-big-data.html>. Citado na página 18.

PAAS SAMPLE SIZE SOFTWARE. *Kappa Test for Agreement between two raters*. 2020. Disponível em: <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Kappa_Test_for_Agreement_Between_Two_Raters.pdf>. Citado na página 41.

PORTAL ACTION. *5.1.2 - Cálculo e interpretação do p-valor*. 2020. Disponível em: <http://www.portalaction.com.br/inferencia/512-calculo-e-interpretacao-do-p-valor>. Citado 2 vezes nas páginas 39 e 40.

RUSSELL, S. *Artificial intelligence : a modern approach*. 3. ed. Upper Saddle River, New Jersey: Prentice Hall, 2010. ISBN 0136042597. Disponível em: <https://www.amazon.com/dp/0136042597?tag=inspiredalgor-20>. Citado na página 21.

SAADATKHAH, N. et al. Flame-assisted spray pyrolysis to size-controlled LiyAlxMn2-xO4: a supervised machine learning approach. *CrystEngComm*, Royal Society of Chemistry (RSC), v. 20, n. 46, p. 7590–7601, 2018. Citado na página 27.

SWENSSON, B.; SäRNDAL, C.-E.; WRETMAN, J. *Model Assisted Survey Sampling*. Springer-Verlag GmbH, 1991. ISBN 0387406204. Disponível em: <https://www.amazon.com.br/Assisted-Survey-Sampling-Carl-Erik-Sarndal/dp/3540975284>. Citado na página 32.

University of Klagenfurt. *Scherbiu's Enigma*. 2020. Disponível em: <http://cs-exhibitions.uni-klu.ac.at/index.php?id=282>. Citado na página 17.

WEINBERGER, K. Supervised learning. 2017. Disponível em: <www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote01_MLsetup.html>. Citado na página 21.

WEINBERGER, K. *Machine Learning Lecture 31 "Random Forests / Bagging-Cornell CS4780 SP17*. 2018. Disponível em: <https://www.youtube.com/watch?v=4EOCQJgqAOY>. Citado na página 28.

WIKIPÉDIA. *Fleiss' kappa*. 2020. Disponível em: <https://en.wikipedia.org/wiki/Fleiss'_kappa>. Citado na página 35.

WINSTON, P. *6.034 Artificial Intelligence. Fall 2010. Lecture 11: Learning: Identification Trees, Disorder*. MIT, 2010. 1 video (49 min.). Disponível em: <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/lecture-videos/lecture-11-learning-identification-trees-disorder/>. Citado 4 vezes nas páginas 24, 25, 26 e 27.

ZANCHI, M. *Crime Data in Brazil*. 2019. Disponível em: <https://www.kaggle.com/inquisitivecrow/crime-data-in-brazil>. Citado 2 vezes nas páginas 43 e 44.

Anexos

Média do índice de impureza de Gini para as variáveis preditoras quando *RUBRICA* assume o valor *Furto (art. 155).*

Tabela 11 – Média do índice de impureza de Gini para as variáveis preditoras quando *RUBRICA* assume o valor *Furto (art. 155).*

DESCR_PROFISSAO	FREQUÊNCIA	(FREQ./TOTAL)^2	DESCR_GRAU_INSTRUCAO	FREQUÊNCIA	(FREQ./TOTAL)^2
OUTRO COM 2 GRAU COMPLETO	4707	0,015905912	2 Grau completo	12746	0,116632077
OUTRO COM 1 GRAU COMPLETO	3125	0,007010843	1 Grau completo	8440	0,051139335
OUTRO COM SUPERIOR COMPLETO	2993	0,006431076	Superior completo	6701	0,032236614
OUTRO - NÃO INFORMADO	2632	0,004973269	1 Grau incompleto	3058	0,006713441
MOTORISTA	1911	0,002621752	NÃO INFORMADO	2632	0,004973269
COMERCIANTE	1606	0,00185166	Superior incompleto	1630	0,001907416
ESTUDANTE	1531	0,001682754	2 Grau incompleto	1252	0,001125327
APOSENTADO(A)	1454	0,001517746	NULL	619	0,000275075
VENDEDOR(A)	1267	0,001152454	Analfabeto	244	4,27415E-05
DO LAR	1175	0,000991165	**TOTAL**	**37322**	**0,215045296**
AJUDANTE	951	0,000649279	**Índice de impureza de Gini**	**0,784954704**	
TECNICO(A)	911	0,000595809	**DIA_DA_SEMANA**	**FREQUÊNCIA**	**(FREQ./TOTAL)^2**
PROFESSOR(A)	903	0,000585391	Quarta-Feira	5794	0,024100562
AUTONOMO(A)	842	0,000508973	Terça-Feira	5777	0,023959344
OUTRO COM 1 GRAU INCOMPLETO	828	0,000492188	Quinta-Feira	5717	0,023464244
OUTRO COM SUPERIOR INCOMPLETO	619	0,000275075	Sexta-Feira	5689	0,023234966
EMPRESARIO(A)	506	0,000183811	Segunda-Feira	5259	0,019855303
DOMESTICA	454	0,000147973	Sábado	5082	0,018541271
PEDREIRO	453	0,000147322	Domingo	4004	0,011509549
AUXILIAR ADMINISTRATIVO	453	0,000147322	**TOTAL**	**37322**	**0,14466524**
GERENTE	414	0,000123047	**Índice de impureza de Gini**	**0,85533476**	
MOTO-BOY	406	0,000118337	**PERIODO_OCORRENCIA**	**FREQUÊNCIA**	**(FREQ/TOTAL)^2**
ENGENHEIRO(A)	362	9,40778E-05	TARDE 1	4942	0,017533785
OPERADOR(A) TELEMARKETING	349	8,74422E-05	TARDE 2	4664	0,015616628
ADMINISTRADOR(A)	342	8,39697E-05	MANHÃ 3	4615	0,015290215
ADVOGADO(A)	336	8,10492E-05	TARDE 3	4301	0,013280336
VIGILANTE	328	7,72357E-05	NOITE 1	4294	0,013237143
MECANICO(A)	327	7,67654E-05	NOITE 2	3561	0,00910362
OUTRO COM 2 GRAU INCOMPLETO	301	6,50434E-05	MANHÃ 2	3252	0,007592263
ASSISTENTE ADMINISTRATIVO	299	6,41819E-05	NOITE 3	2732	0,005358356
AUXILIAR DE LIMPEZA	281	5,66869E-05	MANHÃ 1	2112	0,00320227
BANCARIO(A)	273	5,35051E-05	MADRUGADA 1	998	0,000715042
BALCONISTA	262	4,92802E-05	MADRUGADA 3	983	0,000693709
CABELEIREIRO(A)	257	4,74173E-05	MADRUGADA 2	868	0,000540891
REPRESENTANTE COMERCIAL	252	4,55902E-05	**TOTAL**	**37322**	**0,102164258**
ANALISTA	249	4,45112E-05	**Índice de impureza de Gini**	**0,897835742**	
ELETRICISTA	248	4,41544E-05	**CONDUTA**	**FREQUÊNCIA**	**(FREQ./TOTAL)^2**
OPERADOR(A)	236	3,99847E-05	VEICULO	12674	0,115318129
AUXILIAR DE ENFERMAGEN	235	3,96466E-05	TRANSEUNTE	8216	0,048460852
POLICIAL MILITAR	215	3,31854E-05	OUTROS	6350	0,028947938
RECEPCIONISTA	213	3,25709E-05	INTERIOR DE VEICULO	3943	0,01116153
COBRADOR(A)	190	2,59166E-05	INTERIOR TRANSPORTE COLETIVO	2512	0,004530117
OPERADOR DE MAQUINA	189	2,56445E-05	INTERIOR ESTABELECIMENTO	2391	0,004104208
PORTEIRO(A)	189	2,56445E-05	ESTABELECIMENTO COMERCIAL	560	0,000225137
ESTAGIARIO(A)	186	2,48368E-05	ESTABELECIMENTO-OUTROS	255	4,66821E-05
AUXILIAR DE PRODUCAO	185	2,45705E-05	ESTABELECIMENTO BANCARIO	131	1,23201E-05
PINTOR(A)	180	2,32603E-05	ESTABELECIMENTO ENSINO	110	8,68671E-06
ATENDENTE	166	1,97827E-05	COLETIVO	108	8,37371E-06
ANALISTA DE SISTEMAS	154	1,7026E-05	CARGA	66	3,12722E-06
OUTRO - ANALFABETO	137	1,34745E-05	CONDOMINIO COMERCIAL	3	6,46119E-09
TAXISTA	128	1,17622E-05	NÃO INFORMADO	3	6,46119E-09
CAIXA	112	9,00547E-06	-	-	-
TOTAL	**37322**	**0,049450376**	**TOTAL**	**37322**	**0,212827114**
Índice de impureza de Gini	**0,950549624**		**Índice de impureza de Gini**	**0,787172886**	
CONT_PESSOA	**FREQUÊNCIA**	**(FREQ./TOTAL)^2**	**SEXO_PESSOA**	**FREQUÊNCIA**	**(FREQ./TOTAL)^2**
1	34235	0,841416157	M	22175	0,353018479
2 A 3	2756	0,005452914	F	14930	0,160025723
4 A 5	254	4,63167E-05	I	215	3,31854E-05
MAIS QUE 5	77	4,25649E-06	NÃO INFORMADO	2	2,87164E-09
TOTAL	**37322**	**0,846919644**	**TOTAL**	**37322**	**0,513077391**
Índice de impureza de Gini	**0,153080356**		**Índice de impureza de Gini**	**0,486922609**	
FAIXA_ETARIA	**FREQUÊNCIA**	**(FREQ./TOTAL)^2**	**COR**	**FREQUÊNCIA**	**(FREQ./TOTAL)^2**
21 A 30 ANOS	10485	0,07892363	Branca	25626	0,47144586
31 A 40 ANOS	9434	0,063894272	Parda	6891	0,034090603
41 A 50 ANOS	6682	0,032054066	NÃO DECLARADA	2720	0,005311388
51 A 60 ANOS	4392	0,013848248	Preta	1345	0,001298718
11 A 20 ANOS	3161	0,007173303	Outros	378	0,000102578
61 A 70 ANOS	2102	0,003172018	Amarela	355	9,04746E-05
71 A 80 ANOS	740	0,000393128	Vermelha	7	3,51776E-08
0 A 10 ANOS	169	2,05042E-05	**TOTAL**	**37322**	**0,512339656**
MAIOR QUE 80	157	1,76958E-05	**Índice de impureza de Gini**	**0,487660344**	
TOTAL	**37322**	**0,199496865**	**Soma dos indices de impureza de Gini:**	**6,20401416**	
Índice de impureza de Gini	**0,800503135**		**Média do índice de impureza de Gini:**	**0,689334907**	

Fonte: autoria própria.

Média do índice de impureza de Gini para as variáveis preditoras quando *RUBRICA* assume o valor *Lesão corporal (art 129 § 9º).*

Tabela 12 – Média do índice de impureza de Gini para as variáveis preditoras quando *RUBRICA* assume o valor *Lesão corporal (art 129 § 9º).*

DESCR_PROFISSAO	FREQUÊNCIA	(FREQ/TOTAL)^2	DESCR_GRAU_INSTRUCAO	FREQUÊNCIA	(FREQ/TOTAL)^2
OUTRO COM 2 GRAU COMPLETO	3466	0,014859758	2 Grau completo	8794	0,095659378
OUTRO COM 1 GRAU COMPLETO	3135	0,012157092	1 Grau completo	8261	0,084415048
DO LAR	2705	0,009050846	1 Grau incompleto	3870	0,018525782
OUTRO - NÃO INFORMADO	2357	0,006871852	NÃO INFORMADO	2357	0,006871852
ESTUDANTE	2295	0,006515084	Superior completo	2144	0,005685965
AJUDANTE	1023	0,001294511	2 Grau incompleto	1559	0,003006401
OUTRO COM 1 GRAU INCOMPLETO	973	0,001171063	Superior incompleto	752	0,000699504
OUTRO COM SUPERIOR COMPLETO	934	0,001079067	Analfabeto	369	0,000168425
VENDEDOR(A)	883	0,000964442	NULL	327	0,000132267
DOMESTICA	857	0,000908482	**TOTAL**	**28433**	**0,215164623**
COMERCIANTE	680	0,000571969	**Índice de impureza de Gini**	**0,784835377**	
AUTONOMO(A)	598	0,000442341	DIA_DA_SEMANA	FREQUÊNCIA	(FREQ/TOTAL)^2
APOSENTADO(A)	581	0,000417548	Domingo	5805	0,041683011
MOTORISTA	496	0,000304311	Sábado	4900	0,029699339
AUXILIAR DE LIMPEZA	471	0,000274408	Terça-Feira	3636	0,016353186
PROFESSOR(A)	406	0,000203895	Sexta-Feira	3576	0,01581793
OUTRO COM 2 GRAU INCOMPLETO	375	0,000173947	Segunda-Feira	3553	0,01561511
BALCONISTA	340	0,000142992	Quarta-Feira	3552	0,015606321
CABELEIREIRO(A)	339	0,000142152	Quinta-Feira	3411	0,014391898
AUXILIAR ADMINISTRATIVO	337	0,00014048	**TOTAL**	**28433**	**0,149166795**
OUTRO COM SUPERIOR INCOMPLETO	327	0,000132267	**Índice de impureza de Gini**	**0,850833205**	
OPERADOR(A) TELEMARKETING	304	0,000114315	PERIODO_OCORRENCIA	FREQUÊNCIA	(FREQ/TOTAL)^2
TECNICO(A)	287	0,000101887	NOITE 2	4162	0,021426872
VIGILANTE	268	8,88432E-05	NOITE 1	3833	0,018173236
PEDREIRO	258	8,23368E-05	NOITE 3	3796	0,017824076
RECEPCIONISTA	242	7,24412E-05	TARDE 3	3031	0,011363876
ATENDENTE	235	6,8311E-05	TARDE 2	2513	0,007811593
POLICIAL MILITAR	231	6,60053E-05	TARDE 1	2491	0,007675419
AUXILIAR DE ENFERMAGEN	209	5,40315E-05	MANHÃ 3	2234	0,006173351
EMPRESARIO(A)	206	5,24915E-05	MADRUGADA 1	1660	0,003408559
AUXILIAR DE PRODUCAO	188	4,3719E-05	MANHÃ 2	1560	0,003010259
OUTRO - ANALFABETO	181	4,0524E-05	MANHÃ 1	1155	0,001650132
GERENTE	181	4,0524E-05	MADRUGADA 2	1121	0,001554411
OPERADOR(A)	171	3,61699E-05	MADRUGADA 3	877	0,00095138
PORTEIRO(A)	164	3,32692E-05	**TOTAL**	**28433**	**0,101023164**
ASSISTENTE ADMINISTRATIVO	163	3,28647E-05	**Índice de impureza de Gini**	**0,898976836**	
CAIXA	147	2,67294E-05	CONDUTA	FREQUÊNCIA	(FREQ/TOTAL)^2
PINTOR(A)	137	2,32164E-05	NÃO INFORMADO	28433	1
MECANICO(A)	133	2,18805E-05	-	-	-
ADVOGADO(A)	121	1,81103E-05	-	-	-
COBRADOR(A)	117	1,69327E-05	-	-	-
OPERADOR DE MAQUINA	109	1,46963E-05	-	-	-
MOTO-BOY	109	1,46963E-05	-	-	-
ANALISTA	105	1,36375E-05	-	-	-
ELETRICISTA	91	1,02432E-05	-	-	-
BANCARIO(A)	89	9,79794E-06	-	-	-
ESTAGIARIO(A)	89	9,79794E-06	-	-	-
ADMINISTRADOR(A)	83	8,5214E-06	-	-	-
TAXISTA	66	5,38819E-06	-	-	-
REPRESENTANTE COMERCIAL	62	4,75486E-06	-	-	-
ENGENHEIRO(A)	56	3,8791E-06	-	-	-
ANALISTA DE SISTEMAS	53	3,47461E-06	-	-	-
TOTAL	**28433**	**0,058951998**	**TOTAL**	**28433**	**1**
Índice de impureza de Gini	**0,941048002**		**Índice de impureza de Gini**	**0**	
CONT_PESSOA	FREQUÊNCIA	(FREQ/TOTAL)^2	SEXO_PESSOA	FREQUÊNCIA	(FREQ/TOTAL)^2
1	16906	0,353538210	F	17256	0,368328135
2 A 3	9099	0,102409904	M	10940	0,148043476
4 A 5	1918	0,004550424	I	233	6,71532E-05
MAIS QUE 5	510	0,000321733	NÃO INFORMADO	4	1,97913E-08
TOTAL	**28433**	**0,460820271**	**TOTAL**	**28433**	**0,516438784**
Índice de impureza de Gini	**0,539179729**		**Índice de impureza de Gini**	**0,483561216**	
FAIXA_ETARIA	FREQUÊNCIA	(FREQ/TOTAL)^2	COR	FREQUÊNCIA	(FREQ/TOTAL)^2
21 A 30 ANOS	9085	0,102095005	Branca	17211	0,366409596
31 A 40 ANOS	7163	0,063466498	Parda	7637	0,072143999
11 A 20 ANOS	4818	0,028713638	NÃO DECLARADA	1775	0,003897188
41 A 50 ANOS	3871	0,018535358	Preta	1333	0,002197935
51 A 60 ANOS	1893	0,004432573	Outros	377	0,000175807
61 A 70 ANOS	709	0,000621795	Amarela	94	1,09298E-05
0 A 10 ANOS	592	0,000433509	Vermelha	6	4,45305E-08
71 A 80 ANOS	236	6,88936E-05	**TOTAL**	**28433**	**0,4448355**
MAIOR QUE 80	66	5,38819E-06	**Índice de impureza de Gini**	**0,5551645**	
TOTAL	**28433**	**0,218372656**	**Soma dos índices de impureza de Gini:**	**5,835226209**	
Índice de impureza de Gini	**0,781627344**		**Média do índice de impureza de Gini:**	**0,648358468**	

Fonte: autoria própria.

Média do índice de impureza de Gini para as variáveis preditoras quando *RUBRICA* assume o valor *Lesão corporal culposa na direção de veículo automotor (Art. 303)).*

Tabela 13 – Média do índice de impureza de Gini para as variáveis preditoras quando *RUBRICA* assume o valor *Lesão corporal culposa na direção de veículo automotor (Art. 303)).*

DESCR_PROFISSAO	FREQUÊNCIA	(FREQ/TOTAL)^2	DESCR_GRAU_INSTRUCAO	FREQUÊNCIA	(FREQ/TOTAL)^2
OUTRO - NÃO INFORMADO	4381	0,045934842	1 Grau completo	6326	0,095775427
OUTRO COM 1 GRAU COMPLETO	3086	0,022792269	2 Grau completo	6070	0,088180611
OUTRO COM 2 GRAU COMPLETO	2758	0,01820473	NÃO INFORMADO	4381	0,045934842
ESTUDANTE	1102	0,002906424	1 Grau incompleto	1337	0,004278175
AJUDANTE	944	0,002132749	Superior completo	1162	0,003231528
MOTO-BOY	912	0,001990606	2 Grau incompleto	471	0,00053093
MOTORISTA	543	0,00070566	Superior incompleto	301	0,000216835
OUTRO COM SUPERIOR COMPLETO	533	0,000679908	Analfabeto	273	0,00017837
DO LAR	476	0,000542263	NULL	120	3,44634E-05
APOSENTADO(A)	392	0,000367763	**TOTAL**	**20441**	**0,238361183**
OUTRO COM 1 GRAU INCOMPLETO	361	0,000311896	**Índice de impureza de Gini**	**0,761638817**	
VENDEDOR(A)	349	0,000291505	DIA_DA_SEMANA	FREQUÊNCIA	(FREQ/TOTAL)^2
AUTONOMO(A)	311	0,000231482	Sábado	3514	0,029552841
TECNICO(A)	300	0,000215396	Sexta-Feira	3223	0,024860871
COMERCIANTE	250	0,000149581	Domingo	3187	0,024308595
POLICIAL MILITAR	234	0,000131047	Quarta-Feira	2671	0,017074323
VIGILANTE	222	0,000117951	Quinta-Feira	2663	0,016972197
AUXILIAR ADMINISTRATIVO	206	0,000101562	Segunda-Feira	2622	0,016453606
PROFESSOR(A)	200	9,57317E-05	Terça-Feira	2561	0,015696936
OUTRO - ANALFABETO	195	9,10049E-05	**TOTAL**	**20441**	**0,144919368**
PEDREIRO	179	7,66835E-05	**Índice de impureza de Gini**	**0,855080632**	
MECANICO(A)	153	5,60246E-05	PERIODO_OCORRENCIA	FREQUÊNCIA	(FREQ/TOTAL)^2
DOMESTICA	140	4,69085E-05	NOITE 1	2738	0,01794166
OUTRO COM 2 GRAU INCOMPLETO	133	4,23349E-05	TARDE 3	2351	0,013228207
OPERADOR DE MAQUINA	120	3,44634E-05	TARDE 2	2161	0,011176485
OUTRO COM SUPERIOR INCOMPLETO	120	3,44634E-05	TARDE 1	2048	0,010038195
AUXILIAR DE PRODUCAO	110	2,89588E-05	NOITE 2	2001	0,009582744
OPERADOR(A)	107	2,74008E-05	MANHÃ 1	1910	0,008730969
ELETRICISTA	107	2,74008E-05	NOITE 3	1715	0,007039211
PORTEIRO(A)	99	2,34567E-05	MANHÃ 2	1702	0,006932898
PINTOR(A)	96	2,20566E-05	MANHÃ 3	1573	0,005921792
BALCONISTA	95	2,15995E-05	MADRUGADA 1	873	0,001823997
GERENTE	92	2,02568E-05	MADRUGADA 3	800	0,001531707
AUXILIAR DE LIMPEZA	84	1,68871E-05	MADRUGADA 2	569	0,000774855
ASSISTENTE ADMINISTRATIVO	79	1,49365E-05	**TOTAL**	**20441**	**0,094722721**
COBRADOR(A)	79	1,49365E-05	**Índice de impureza de Gini**	**0,905277279**	
OPERADOR(A) TELEMARKETING	77	1,41898E-05	CONDUTA	FREQUÊNCIA	(FREQ/TOTAL)^2
ANALISTA	77	1,41898E-05	NÃO INFORMADO	20441	1
RECEPCIONISTA	76	1,38237E-05	-	-	-
CABELEIREIRO(A)	72	1,24068E-05	-	-	-
ATENDENTE	71	1,20646E-05	-	-	-
EMPRESARIO(A)	63	9,49898E-06	-	-	-
AUXILIAR DE ENFERMAGEN	63	9,49898E-06	-	-	-
ADVOGADO(A)	60	8,61585E-06	-	-	-
BANCARIO(A)	56	7,50536E-06	-	-	-
ESTAGIARIO(A)	51	6,22495E-06	-	-	-
ADMINISTRADOR(A)	48	5,51415E-06	-	-	-
ANALISTA DE SISTEMAS	48	5,51415E-06	-	-	-
CAIXA	41	4,02312E-06	-	-	-
ENGENHEIRO(A)	40	3,82927E-06	-	-	-
REPRESENTANTE COMERCIAL	25	1,49581E-06	-	-	-
TAXISTA	25	1,49581E-06	-	-	-
TOTAL	**20441**	**0,09863303**	**TOTAL**	**20441**	**1**
Índice de impureza de Gini	**0,90136697**		**Índice de impureza de Gini**	**0**	
CONT_PESSOA	FREQUÊNCIA	(FREQ/TOTAL)^2	SEXO_PESSOA	FREQUÊNCIA	(FREQ/TOTAL)^2
2 A 3	10429	0,26030413	M	14147	0,478987777
1	6648	0,105773712	F	6031	0,087051123
4 A 5	2612	0,016328341	I	263	0,000165542
MAIS QUE 5	752	0,001353416	-	-	-
TOTAL	**20441**	**0,383759599**	**TOTAL**	**20441**	**0,566204442**
Índice de impureza de Gini	**0,616240401**		**Índice de impureza de Gini**	**0,433795558**	
FAIXA_ETARIA	FREQUÊNCIA	(FREQ/TOTAL)^2	COR	FREQUÊNCIA	(FREQ/TOTAL)^2
21 A 30 ANOS	7265	0,1263185	Branca	11797	0,333072581
31 A 40 ANOS	4070	0,039644646	Parda	5269	0,066443442
11 A 20 ANOS	3505	0,029401654	NÃO DECLARADA	1529	0,005595137
41 A 50 ANOS	2289	0,012539705	Outros	961	0,002210256
51 A 60 ANOS	1299	0,004038444	Preta	783	0,001467301
0 A 10 ANOS	866	0,001794864	Amarela	102	2,48998E-05
61 A 70 ANOS	634	0,000961998	-	-	-
71 A 80 ANOS	378	0,000341963	**TOTAL**	**20441**	**0,408813616**
MAIOR QUE 80	135	4,36178E-05	**Índice de impureza de Gini**	**0,591186384**	
TOTAL	**20441**	**0,215085392**	**Soma dos índices de impureza de Gini:**	**5,84950065**	
Índice de impureza de Gini	**0,784914608**		**Média do índice de impureza de Gini:**	**0,649944517**	

Fonte: autoria própria.

Arquivo R-Markdown

Carregando pacotes:

```
library(RODBC) # Pacote para criar conexão com o banco de dados no
               # SQL Server
library(randomForest) # Pacote para criar o modelo
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(dplyr) # Pacote para manusear os dados.
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret) # Pacote para plotar o gráfico de importância de atributos.
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
library(irr) # Pacote para avaliar o resultado do modelo.
```

```
## Loading required package: lpSolve
```

Carregando a base dedados

Cria conexão com o banco de dados e carrega a tabela *[tccDB].[dbo].[vw_TCC]* para o *dataframe tccDF*. Por fim, exibe os dados carregados.

```
cn <- odbcDriverConnect(connection="Driver={SQLServer};
server=DESKTOP-FL71F1C;
database=tccDB;
trusted_connection=yes;")

tccDF = sqlQuery(cn, 'SELECT * FROM [tccDB].[dbo].[vw_TCC]')
```

Criando o modelo

Seleciona 5% da base de dados para criar o modelo e exibe os dados no final. A amostragem é simples, aleatória e com reposição.

```
sampleTCCDF <- tccDF %>%
  group_by(RUBRICA) %>%
  sample_frac(0.05, replace = TRUE) %>%
  ungroup
View(sampleTCCDF)
```

Alguns fatores podem ser omitidos da amostra devido à baixa frequência que aparecem na tabela original. Para contornar esse problema, convertemos a variável-alvo do tipo fator para o tipo caracter e depois reconvertemos para o tipo fator.

```
sampleTCCDF$RUBRICA = as.character(sampleTCCDF$RUBRICA)
sampleTCCDF$RUBRICA = as.factor(sampleTCCDF$RUBRICA)
```

Cria partição de treino e teste:

```
set.seed(24165) # Para gerar aleatoriedade
trainIndexSample = createDataPartition(sampleTCCDF$RUBRICA,
p = 0.80,
list = FALSE,
times = 1)
```

```
## Warning in createDataPartition(sampleTCCDF$RUBRICA, p = 0.8, list = FALSE, :
## Some classes have a single record ( Induzir, instigar ou auxiliar
## alguém ao uso indevido de droga(Art.33,§2º)
## , Porte de entorpecente (Art. 16) ) and these will
## be selected for the sample
```

```
trainSample = sampleTCCDF[trainIndexSample,]
testSample = sampleTCCDF[-trainIndexSample,]
```

Cria modelo de árvore aleatória:

```
modelRF = randomForest(trainSample$RUBRICA ~., data=trainSample,
ntrees = 500,
importance = TRUE)
```

```
importance(modelRF, type = 2)
```

```
##                       MeanDecreaseGini
## DIA_DA_SEMANA                 7562.898
## PERIODO_OCORRENCIA            8747.719
## CONDUTA                      36101.411
## CONT_PESSOA                   3737.556
## SEXO_PESSOA                   2454.208
## FAIXA_ETARIA                  5877.945
## COR                           4367.054
## DESCR_PROFISSAO              10189.819
## DESCR_GRAU_INSTRUCAO          4222.093
```

Erro OOB:

```r
print(modelRF)
```

```
##
## Call:
##  randomForest(formula = trainSample$RUBRICA ~ .,
 data = trainSample,      ntrees = 500, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 30.26%
## Confusion matrix:
##                                                    A.I.-Homicídio simples (art. 121)
## A.I.-Homicídio simples (art.121)                                                  85
## Drogas sem autorização ou em desacordo (Art.33,caput)                              0
## Estupro (art.213)                                                                  0
## Estupro de vulneravel(art.217-A)                                                   2
## Furto (art.155)                                                                    0
## Furto de coisa comum (art. 156)                                                    0
## Furto qualificado (art. 155,§4o.)                                                  0
## Homicídio culposo (art. 121,§3o.)                                                  0
## Homicídio culposo na direção de veículo automotor (Art.302)                        5
## Homicídio qualificado (art. 121,§2o.)                                              5
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)          0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129,§2o.)                             0
## Lesão corporal (art 129 §9º)                                                      48
## Lesão corporal culposa (art. 129. §6o.)                                            0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)                 53
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                                  0
## Lesão corporal seguida de morte (art. 129, §3o.)                                   0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                         0
## Porte de entorpecente (Art. 16)                                                    0
## Roubo (art. 157)                                                                   0
##                                Drogas sem autorização ou em desacordo (Art.33, caput)
## A.I.-Homicídio simples (art. 121)                                                  0
## Drogas sem autorização ou em desacordo (Art.33, caput)                             6
## Estupro (art. 213)                                                                 0
## Estupro de vulneravel (art.217-A)                                                  0
## Furto (art. 155)                                                                   0
## Furto de coisa comum (art. 156)                                                    0
## Furto qualificado (art. 155, §4o.)                                                 0
```

```
## Homicídio culposo (art. 121, §3o.)                                    0
## Homicídio culposo na direção de veículo automotor (Art. 302)          0
## Homicídio qualificado (art. 121, §2o.)                                0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)               0
## Lesão corporal (art 129 § 9º)                                         0
## Lesão corporal culposa (art. 129. §6o.)                               0
## Lesão corporal culposa na direção de veículo automotor (Art. 303      4
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                     0
## Lesão corporal seguida de morte (art. 129, §3o.)                      0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)            0
## Porte de entorpecente (Art. 16)                                       1
## Roubo (art. 157)                                                      0
##                                                        Estupro (art. 213)
## A.I.-Homicídio simples (art. 121)                                     0
## Drogas sem autorização ou em desacordo (Art.33, caput)                0
## Estupro (art. 213)                                                    1
## Estupro de vulneravel (art.217-A)                                     1
## Furto (art. 155)                                                      0
## Furto de coisa comum (art. 156)                                       0
## Furto qualificado (art. 155, §4o.)                                    0
## Homicídio culposo (art. 121, §3o.)                                    0
## Homicídio culposo na direção de veículo automotor (Art. 302)          1
## Homicídio qualificado (art. 121, §2o.)                                0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)               0
## Lesão corporal (art 129 § 9º)                                         4
## Lesão corporal culposa (art. 129. §6o.)                               0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)     4
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                     0
## Lesão corporal seguida de morte (art. 129, §3o.)                      0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)            0
## Porte de entorpecente (Art. 16)                                       0
## Roubo (art. 157)                                                      0
##                                          Estupro de vulneravel (art.217-A)
## A.I.-Homicídio simples (art. 121)                                     2
## Drogas sem autorização ou em desacordo (Art.33, caput)                0
## Estupro (art. 213)                                                    5
## Estupro de vulneravel (art.217-A)                                    13
## Furto (art. 155)                                                      0
## Furto de coisa comum (art. 156)                                       0
```

```
## Furto qualificado (art. 155, §4o.)                                    0
## Homicídio culposo (art. 121, §3o.)                                    0
## Homicídio culposo na direção de veículo automotor (Art. 302)          0
## Homicídio qualificado (art. 121, §2o.)                                0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)               0
## Lesão corporal (art 129 § 9º)                                        14
## Lesão corporal culposa (art. 129. §6o.)                               4
## Lesão corporal culposa na direção de veículo automotor (Art. 303)    20
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                     0
## Lesão corporal seguida de morte (art. 129, §3o.)                      0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)            0
## Porte de entorpecente (Art. 16)                                       0
## Roubo (art. 157)                                                      0
##                                                        Furto (art. 155)
## A.I.-Homicídio simples (art. 121)                                     0
## Drogas sem autorização ou em desacordo (Art.33, caput)                0
## Estupro (art. 213)                                                    0
## Estupro de vulneravel (art.217-A)                                     0
## Furto (art. 155)                                                  12891
## Furto de coisa comum (art. 156)                                       6
## Furto qualificado (art. 155, §4o.)                                 2161
## Homicídio culposo (art. 121, §3o.)                                    0
## Homicídio culposo na direção de veículo automotor (Art. 302)          0
## Homicídio qualificado (art. 121, §2o.)                                0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)               0
## Lesão corporal (art 129 § 9º)                                         0
## Lesão corporal culposa (art. 129. §6o.)                               0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)     0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                     0
## Lesão corporal seguida de morte (art. 129, §3o.)                      0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)            0
## Porte de entorpecente (Art. 16)                                       0
## Roubo (art. 157)                                                   7606
##                                             Furto de coisa comum (art. 156)
## A.I.-Homicídio simples (art. 121)                                     0
## Drogas sem autorização ou em desacordo (Art.33, caput)                0
## Estupro (art. 213)                                                    0
## Estupro de vulneravel (art.217-A)                                     0
## Furto (art. 155)                                                      1
```

```
## Furto de coisa comum (art. 156)                                          0
## Furto qualificado (art. 155, §4o.)                                       0
## Homicídio culposo (art. 121, §3o.)                                       0
## Homicídio culposo na direção de veículo automotor (Art. 302)             0
## Homicídio qualificado (art. 121, §2o.)                                   0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º) 0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                  0
## Lesão corporal (art 129 § 9º)                                            0
## Lesão corporal culposa (art. 129. §6o.)                                  0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)        0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                        0
## Lesão corporal seguida de morte (art. 129, §3o.)                         0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)               0
## Porte de entorpecente (Art. 16)                                          0
## Roubo (art. 157)                                                         0
##                                              Furto qualificado (art. 155, §4o.)
## A.I.-Homicídio simples (art. 121)                                        0
## Drogas sem autorização ou em desacordo (Art.33, caput)                   0
## Estupro (art. 213)                                                       0
## Estupro de vulneravel (art.217-A)                                        0
## Furto (art. 155)                                                       854
## Furto de coisa comum (art. 156)                                          2
## Furto qualificado (art. 155, §4o.)                                    1503
## Homicídio culposo (art. 121, §3o.)                                       0
## Homicídio culposo na direção de veículo automotor (Art. 302)             0
## Homicídio qualificado (art. 121, §2o.)                                   0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º) 0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                  0
## Lesão corporal (art 129 § 9º)                                            0
## Lesão corporal culposa (art. 129. §6o.)                                  0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)        0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                        0
## Lesão corporal seguida de morte (art. 129, §3o.)                         0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)               0
## Porte de entorpecente (Art. 16)                                          0
## Roubo (art. 157)                                                      1047
##                                              Homicídio culposo (art. 121, §3o.)
## A.I.-Homicídio simples (art. 121)                                        0
## Drogas sem autorização ou em desacordo (Art.33, caput)                   0
## Estupro (art. 213)                                                       0
## Estupro de vulneravel (art.217-A)                                        0
```

```
## Furto (art. 155)                                                    0
## Furto de coisa comum (art. 156)                                     0
## Furto qualificado (art. 155, §4o.)                                  0
## Homicídio culposo (art. 121, §3o.)                                  0
## Homicídio culposo na direção de veículo automotor (Art. 302)        0
## Homicídio qualificado (art. 121, §2o.)                              0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)             0
## Lesão corporal (art 129 § 9º)                                       0
## Lesão corporal culposa (art. 129. §6o.)                             0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)   0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                   0
## Lesão corporal seguida de morte (art. 129, §3o.)                    0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)          0
## Porte de entorpecente (Art. 16)                                     0
## Roubo (art. 157)                                                    0
##                       Homicídio culposo na direção de veículo automotor (Art. 302)
## A.I.-Homicídio simples (art. 121)                                   1
## Drogas sem autorização ou em desacordo (Art.33, caput)              0
## Estupro (art. 213)                                                  1
## Estupro de vulneravel (art.217-A)                                   0
## Furto (art. 155)                                                    0
## Furto de coisa comum (art. 156)                                     0
## Furto qualificado (art. 155, §4o.)                                  0
## Homicídio culposo (art. 121, §3o.)                                  0
## Homicídio culposo na direção de veículo automotor (Art. 302)       13
## Homicídio qualificado (art. 121, §2o.)                              1
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)             0
## Lesão corporal (art 129 § 9º)                                       7
## Lesão corporal culposa (art. 129. §6o.)                             1
## Lesão corporal culposa na direção de veículo automotor (Art. 303)  15
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                   0
## Lesão corporal seguida de morte (art. 129, §3o.)                    0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)          0
## Porte de entorpecente (Art. 16)                                     0
## Roubo (art. 157)                                                    0
##                                          Homicídio qualificado (art. 121, §2o.)
## A.I.-Homicídio simples (art. 121)                                   3
## Drogas sem autorização ou em desacordo (Art.33, caput)              0
## Estupro (art. 213)                                                  0
```

```
## Estupro de vulneravel (art.217-A)                                          0
## Furto (art. 155)                                                           0
## Furto de coisa comum (art. 156)                                            0
## Furto qualificado (art. 155, §4o.)                                         0
## Homicídio culposo (art. 121, §3o.)                                         0
## Homicídio culposo na direção de veículo automotor (Art. 302)               1
## Homicídio qualificado (art. 121, §2o.)                                     4
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                    0
## Lesão corporal (art 129 § 9º)                                              4
## Lesão corporal culposa (art. 129. §6o.)                                    0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)          8
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                          0
## Lesão corporal seguida de morte (art. 129, §3o.)                           0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                 0
## Porte de entorpecente (Art. 16)                                            0
## Roubo (art. 157)                                                           0
##         Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)
## A.I.-Homicídio simples (art. 121)                                          0
## Drogas sem autorização ou em desacordo (Art.33, caput)                     0
## Estupro (art. 213)                                                         0
## Estupro de vulneravel (art.217-A)                                          0
## Furto (art. 155)                                                           0
## Furto de coisa comum (art. 156)                                            0
## Furto qualificado (art. 155, §4o.)                                         0
## Homicídio culposo (art. 121, §3o.)                                         0
## Homicídio culposo na direção de veículo automotor (Art. 302)               0
## Homicídio qualificado (art. 121, §2o.)                                     0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                    0
## Lesão corporal (art 129 § 9º)                                              0
## Lesão corporal culposa (art. 129. §6o.)                                    0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)          0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                          0
## Lesão corporal seguida de morte (art. 129, §3o.)                           0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                 0
## Porte de entorpecente (Art. 16)                                            0
## Roubo (art. 157)                                                           0
##                           Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)
## A.I.-Homicídio simples (art. 121)                                          0
## Drogas sem autorização ou em desacordo (Art.33, caput)                     0
```

```
## Estupro (art. 213)                                                    0
## Estupro de vulneravel (art.217-A)                                     0
## Furto (art. 155)                                                      0
## Furto de coisa comum (art. 156)                                       0
## Furto qualificado (art. 155, §4o.)                                    0
## Homicídio culposo (art. 121, §3o.)                                    0
## Homicídio culposo na direção de veículo automotor (Art. 302)          0
## Homicídio qualificado (art. 121, §2o.)                                0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)               0
## Lesão corporal (art 129 § 9º)                                         0
## Lesão corporal culposa (art. 129. §6o.)                               0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)     0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                     0
## Lesão corporal seguida de morte (art. 129, §3o.)                      0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)            0
## Porte de entorpecente (Art. 16)                                       0
## Roubo (art. 157)                                                      0
##                                              Lesão corporal (art 129 § 9º)
## A.I.-Homicídio simples (art. 121)                                   691
## Drogas sem autorização ou em desacordo (Art.33, caput)               34
## Estupro (art. 213)                                                  368
## Estupro de vulneravel (art.217-A)                                   179
## Furto (art. 155)                                                      2
## Furto de coisa comum (art. 156)                                       0
## Furto qualificado (art. 155, §4o.)                                    1
## Homicídio culposo (art. 121, §3o.)                                   19
## Homicídio culposo na direção de veículo automotor (Art. 302)        171
## Homicídio qualificado (art. 121, §2o.)                              131
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)   1
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)               6
## Lesão corporal (art 129 § 9º)                                     17455
## Lesão corporal culposa (art. 129. §6o.)                             345
## Lesão corporal culposa na direção de veículo automotor (Art. 303)  5886
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                    30
## Lesão corporal seguida de morte (art. 129, §3o.)                      2
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)            1
## Porte de entorpecente (Art. 16)                                       0
## Roubo (art. 157)                                                      5
##                                         Lesão corporal culposa (art. 129. §6o.)
## A.I.-Homicídio simples (art. 121)                                     0
```

```
## Drogas sem autorização ou em desacordo (Art.33, caput)                    0
## Estupro (art. 213)                                                        0
## Estupro de vulneravel (art.217-A)                                         0
## Furto (art. 155)                                                          0
## Furto de coisa comum (art. 156)                                           0
## Furto qualificado (art. 155, §4o.)                                        0
## Homicídio culposo (art. 121, §3o.)                                        0
## Homicídio culposo na direção de veículo automotor (Art. 302)              0
## Homicídio qualificado (art. 121, §2o.)                                    1
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                   0
## Lesão corporal (art 129 § 9º)                                            12
## Lesão corporal culposa (art. 129. §6o.)                                  11
## Lesão corporal culposa na direção de veículo automotor (Art. 303)        12
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                         0
## Lesão corporal seguida de morte (art. 129, §3o.)                          0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                0
## Porte de entorpecente (Art. 16)                                           0
## Roubo (art. 157)                                                          0
##                 Lesão corporal culposa na direção de veículo automotor (Art. 303)
## A.I.-Homicídio simples (art. 121)                                       867
## Drogas sem autorização ou em desacordo (Art.33, caput)                   35
## Estupro (art. 213)                                                       55
## Estupro de vulneravel (art.217-A)                                       132
## Furto (art. 155)                                                          1
## Furto de coisa comum (art. 156)                                           0
## Furto qualificado (art. 155, §4o.)                                        0
## Homicídio culposo (art. 121, §3o.)                                       17
## Homicídio culposo na direção de veículo automotor (Art. 302)            336
## Homicídio qualificado (art. 121, §2o.)                                  187
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                   2
## Lesão corporal (art 129 § 9º)                                          5201
## Lesão corporal culposa (art. 129. §6o.)                                 261
## Lesão corporal culposa na direção de veículo automotor (Art. 303)     10348
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                        37
## Lesão corporal seguida de morte (art. 129, §3o.)                          5
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                2
## Porte de entorpecente (Art. 16)                                           0
## Roubo (art. 157)                                                          7
##                               Lesão corporal de natureza GRAVE (art. 129, §1o.)
```

```
## A.I.-Homicídio simples (art. 121)                                    0
## Drogas sem autorização ou em desacordo (Art.33, caput)               0
## Estupro (art. 213)                                                   0
## Estupro de vulneravel (art.217-A)                                    0
## Furto (art. 155)                                                     0
## Furto de coisa comum (art. 156)                                      0
## Furto qualificado (art. 155, §4o.)                                   0
## Homicídio culposo (art. 121, §3o.)                                   0
## Homicídio culposo na direção de veículo automotor (Art. 302)         0
## Homicídio qualificado (art. 121, §2o.)                               0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)              0
## Lesão corporal (art 129 § 9º)                                        1
## Lesão corporal culposa (art. 129. §6o.)                              0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)    3
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                    0
## Lesão corporal seguida de morte (art. 129, §3o.)                     0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)           0
## Porte de entorpecente (Art. 16)                                      0
## Roubo (art. 157)                                                     0
##                                 Lesão corporal seguida de morte (art. 129, §3o.)
## A.I.-Homicídio simples (art. 121)                                    0
## Drogas sem autorização ou em desacordo (Art.33, caput)               0
## Estupro (art. 213)                                                   0
## Estupro de vulneravel (art.217-A)                                    0
## Furto (art. 155)
## Furto de coisa comum (art. 156)                                      0
## Furto qualificado (art. 155, §4o.)                                   0
## Homicídio culposo (art. 121, §3o.)                                   0
## Homicídio culposo na direção de veículo automotor (Art. 302)         0
## Homicídio qualificado (art. 121, §2o.)                               0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)              0
## Lesão corporal (art 129 § 9º)                                        0
## Lesão corporal culposa (art. 129. §6o.)                              0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)    0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                    0
## Lesão corporal seguida de morte (art. 129, §3o.)                     0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)           0
## Porte de entorpecente (Art. 16)                                      0
## Roubo (art. 157)                                                     0
```

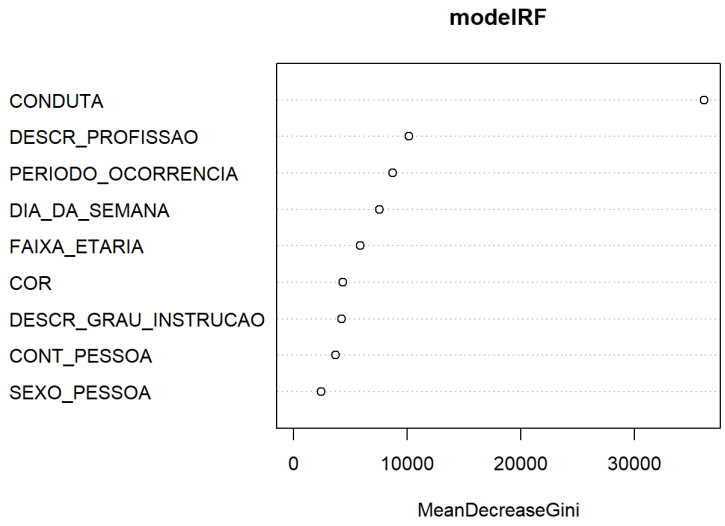
```
##                         Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)
## A.I.-Homicídio simples (art. 121)                                                0
## Drogas sem autorização ou em desacordo (Art.33, caput)                           0
## Estupro (art. 213)                                                               0
## Estupro de vulneravel (art.217-A)                                                0
## Furto (art. 155)                                                                 0
## Furto de coisa comum (art. 156)                                                  0
## Furto qualificado (art. 155, §4o.)                                               0
## Homicídio culposo (art. 121, §3o.)                                               0
## Homicídio culposo na direção de veículo automotor (Art. 302)                     0
## Homicídio qualificado (art. 121, §2o.)                                           0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)        0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                          0
## Lesão corporal (art 129 § 9º)                                                    0
## Lesão corporal culposa (art. 129. §6o.)                                          0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)                0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                                0
## Lesão corporal seguida de morte (art. 129, §3o.)                                 0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                       0
## Porte de entorpecente (Art. 16)                                                  0
## Roubo (art. 157)                                                                 0
##                                                  Porte de entorpecente (Art. 16)
## A.I.-Homicídio simples (art. 121)                                                0
## Drogas sem autorização ou em desacordo (Art.33, caput)                           1
## Estupro (art. 213)                                                               0
## Estupro de vulneravel (art.217-A)                                                0
## Furto (art. 155)                                                                 0
## Furto de coisa comum (art. 156)                                                  0
## Furto qualificado (art. 155, §4o.)                                               0
## Homicídio culposo (art. 121, §3o.)                                               0
## Homicídio culposo na direção de veículo automotor (Art. 302)                     0
## Homicídio qualificado (art. 121, §2o.)                                           0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)        0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)                          0
## Lesão corporal (art 129 § 9º)                                                    0
## Lesão corporal culposa (art. 129. §6o.)                                          0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)                0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                                0
## Lesão corporal seguida de morte (art. 129, §3o.)                                 0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)                       0
## Porte de entorpecente (Art. 16)                                                  0
```

```
## Roubo (art. 157)                                                    0
##                                                          Roubo (art. 157)
## A.I.-Homicídio simples (art. 121)                                   0
## Drogas sem autorização ou em desacordo (Art.33, caput)              0
## Estupro (art. 213)                                                  0
## Estupro de vulneravel (art.217-A)                                   0
## Furto (art. 155)                                                16109
## Furto de coisa comum (art. 156)                                     4
## Furto qualificado (art. 155, §4o.)                               3077
## Homicídio culposo (art. 121, §3o.)                                  0
## Homicídio culposo na direção de veículo automotor (Art. 302)        0
## Homicídio qualificado (art. 121, §2o.)                              0
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)  0
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)             0
## Lesão corporal (art 129 § 9º)                                       0
## Lesão corporal culposa (art. 129. §6o.)                             0
## Lesão corporal culposa na direção de veículo automotor (Art. 303)   0
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                   0
## Lesão corporal seguida de morte (art. 129, §3o.)                    0
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)          0
## Porte de entorpecente (Art. 16)                                     0
## Roubo (art. 157)                                                63967
##
                                                            class.error
## A.I.-Homicídio simples (art. 121)                        0.9484536
## Drogas sem autorização ou em desacordo (Art.33, caput)   0.9210526
## Estupro (art. 213)                                       0.9976744
## Estupro de vulneravel (art.217-A)                        0.9602446
## Furto (art. 155)                                         0.5682564
## Furto de coisa comum (art. 156)                          1.0000000
## Furto qualificado (art. 155, §4o.)                       0.7770691
## Homicídio culposo (art. 121, §3o.)                       1.0000000
## Homicídio culposo na direção de veículo automotor (Art. 302)  0.9753788
## Homicídio qualificado (art. 121, §2o.)                   0.9878049
## Induzir, instigar ou
## auxiliar alguém ao uso indevido de droga(Art.33,§2º)     1.0000000
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)  1.0000000
## Lesão corporal (art 129 § 9º)                            0.2326461
## Lesão corporal culposa (art. 129. §6o.)                  0.9823151
## Lesão corporal culposa na direção
## de veículo automotor (Art. 303)                          0.3672109
```

```
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                1.0000000
## Lesão corporal seguida de morte (art. 129, §3o.)                 1.0000000
## Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)       1.0000000
## Porte de entorpecente (Art. 16)                                  1.0000000
## Roubo (art. 157)                                                 0.1193000
```

Avalia importância das variáveis

```
varImpPlot(modelRF, type = 2)
```




```
varImp(modelRF)
```

```
##                                          A.I.-Homicídio simples (art. 121)
## DIA_DA_SEMANA                                            17.30524
## PERIODO_OCORRENCIA                                          50.92962
## CONDUTA                                                    133.52707
## CONT_PESSOA                                                 41.56555
## SEXO_PESSOA                                                 56.00556
## FAIXA_ETARIA                                                15.65886
## COR                                                         31.51424
## DESCR_PROFISSAO                                  28.83849
## DESCR_GRAU_INSTRUCAO                            20.96978
##                      Drogas sem autorização ou em desacordo (Art.33, caput)
## DIA_DA_SEMANA                                                16.31222
## PERIODO_OCORRENCIA                                           16.63368
## CONDUTA                                                      24.31911
```

```
## CONT_PESSOA                                          24.05431
## SEXO_PESSOA                                          14.45723
## FAIXA_ETARIA                                         11.70961
## COR                                                  15.80429
## DESCR_PROFISSAO                                      17.59569
## DESCR_GRAU_INSTRUCAO                                 13.79305
##                      Estupro (art. 213) Estupro de vulneravel (art.217-A)
## DIA_DA_SEMANA                 5.700959                          0.1415451
## PERIODO_OCORRENCIA           14.362566                          0.4417911
## CONDUTA                      46.124116                         60.9219705
## CONT_PESSOA                   9.732843                        -16.9360852
## SEXO_PESSOA                  38.985391                         23.1723847
## FAIXA_ETARIA                 18.356118                         51.0729060
## COR                           7.852942                          0.5956606
## DESCR_PROFISSAO               9.038622                         27.9566305
## DESCR_GRAU_INSTRUCAO          4.182383                         21.7221514
##                      Furto (art. 155) Furto de coisa comum (art. 156)
## DIA_DA_SEMANA               30.70123                          0.000000
## PERIODO_OCORRENCIA         141.02056                          0.000000
## CONDUTA                    564.32826                         -1.001002
## CONT_PESSOA                340.90255                          0.000000
## SEXO_PESSOA                 71.61716                          0.000000
## FAIXA_ETARIA                39.06686                          0.000000
## COR                         17.77579                          0.000000
## DESCR_PROFISSAO             26.54744                          0.000000
## DESCR_GRAU_INSTRUCAO         9.45652                          0.000000
##                      Furto qualificado (art. 155, §4o.)
## DIA_DA_SEMANA                                  21.41224
## PERIODO_OCORRENCIA                             40.62792
## CONDUTA                                       320.03721
## CONT_PESSOA                                    73.96437
## SEXO_PESSOA                                    20.10121
## FAIXA_ETARIA                                   36.00594
## COR                                            24.20444
## DESCR_PROFISSAO                                54.92042
## DESCR_GRAU_INSTRUCAO                           46.62450
##                      Homicídio culposo (art. 121, §3o.)
## DIA_DA_SEMANA                                 1.5893695
## PERIODO_OCORRENCIA                           -0.9592172
```

```
## CONDUTA                                              3.5077383
## CONT_PESSOA                                         -0.5206149
## SEXO_PESSOA                                          2.3499818
## FAIXA_ETARIA                                        -2.1519087
## COR                                                  1.9596658
## DESCR_PROFISSAO                                      2.7777118
## DESCR_GRAU_INSTRUCAO                                 2.2072664
##              Homicídio culposo na direção de veículo automotor (Art. 302)
## DIA_DA_SEMANA                                       24.46089
## PERIODO_OCORRENCIA                                  23.15080
## CONDUTA                                             52.82031
## CONT_PESSOA                                         26.12366
## SEXO_PESSOA                                         26.31861
## FAIXA_ETARIA                                        17.64190
## COR                                                 21.83273
## DESCR_PROFISSAO                                     19.89689
## DESCR_GRAU_INSTRUCAO                                19.32716
##                         Homicídio qualificado (art. 121, §2o.)
## DIA_DA_SEMANA                                        12.225992
## PERIODO_OCORRENCIA                                   11.532301
## CONDUTA                                              33.304851
## CONT_PESSOA                                          22.856356
## SEXO_PESSOA                                          11.253444
## FAIXA_ETARIA                                         10.750851
## COR                                                   7.464387
## DESCR_PROFISSAO                                       7.846303
## DESCR_GRAU_INSTRUCAO                                  8.919181
## Induzir, instigar ou auxiliar alguém ao uso indevido de droga(Art.33,§2º)
## DIA_DA_SEMANA                                                0
## PERIODO_OCORRENCIA                                           0
## CONDUTA                                                      0
## CONT_PESSOA                                                  0
## SEXO_PESSOA                                                  0
## FAIXA_ETARIA                                                 0
## COR                                          0
## DESCR_PROFISSAO                              0
## DESCR_GRAU_INSTRUCAO                         0
##        Lesão corporal de natureza GRAVÍSSIMA (art. 129, §2o.)
## DIA_DA_SEMANA                                                0
```

```
## PERIODO_OCORRENCIA                                    0
## CONDUTA                                               0
## CONT_PESSOA                                           0
## SEXO_PESSOA                                           0
## FAIXA_ETARIA                                          0
## COR                                                   0
## DESCR_PROFISSAO                                       0
## DESCR_GRAU_INSTRUCAO                                  0
##                                Lesão corporal (art 129 § 9º)
## DIA_DA_SEMANA                                     46.77539
## PERIODO_OCORRENCIA                                49.69659
## CONDUTA                                         1263.44292
## CONT_PESSOA                                      -33.98628
## SEXO_PESSOA                                      177.87787
## FAIXA_ETARIA                                      81.02916
## COR                                               31.55148
## DESCR_PROFISSAO                                   78.66203
## DESCR_GRAU_INSTRUCAO                              47.69056
##                       Lesão corporal culposa (art. 129. §6o.)
## DIA_DA_SEMANA                                     13.741381
## PERIODO_OCORRENCIA                                16.604055
## CONDUTA                                           49.008441
## CONT_PESSOA                                        6.051817
## SEXO_PESSOA                                        9.088175
## FAIXA_ETARIA                                      20.181862
## COR                                               13.931310
## DESCR_PROFISSAO                                   11.347840
## DESCR_GRAU_INSTRUCAO                               4.925487
##     Lesão corporal culposa na direção de veículo automotor (Art. 303)
## DIA_DA_SEMANA                                     34.32927
## PERIODO_OCORRENCIA                                85.58386
## CONDUTA                                          796.26154
## CONT_PESSOA                                      150.20382
## SEXO_PESSOA                                      150.59222
## FAIXA_ETARIA                                      71.02730
## COR                                               22.06241
## DESCR_PROFISSAO                                   83.76662
## DESCR_GRAU_INSTRUCAO                              62.65359
##                       Lesão corporal de natureza GRAVE (art. 129, §1o.)
```

```
## DIA_DA_SEMANA                                        3.702506
## PERIODO_OCORRENCIA                                   6.214142
## CONDUTA                                              7.629811
## CONT_PESSOA                                          1.816620
## SEXO_PESSOA                                          5.393231
## FAIXA_ETARIA                                         7.531878
## COR                                                  4.153679
## DESCR_PROFISSAO                                      4.865270
## DESCR_GRAU_INSTRUCAO                                 4.970878
##                      Lesão corporal seguida de morte (art. 129, §3o.)
## DIA_DA_SEMANA                                                   0
## PERIODO_OCORRENCIA                                              0
## CONDUTA                                                         0
## CONT_PESSOA                                                     0
## SEXO_PESSOA                                                     0
## FAIXA_ETARIA                                                    0
## COR                                                             0
## DESCR_PROFISSAO                                                 0
## DESCR_GRAU_INSTRUCAO                                            0
##              Oferecer droga a pessoa de seu relacionamento (Art.33,§3º)
## DIA_DA_SEMANA                                                   0
## PERIODO_OCORRENCIA                                              0
## CONDUTA                                                         0
## CONT_PESSOA                                                     0
## SEXO_PESSOA                                                     0
## FAIXA_ETARIA                                                    0
## COR                                                             0
## DESCR_PROFISSAO                                                 0
## DESCR_GRAU_INSTRUCAO                                            0
##                      Porte de entorpecente (Art. 16)      Roubo (art. 157)
## DIA_DA_SEMANA                                      0         59.15870
## PERIODO_OCORRENCIA                                 0        233.03103
## CONDUTA                                            0       1339.06856
## CONT_PESSOA                                        0         41.49901
## SEXO_PESSOA                                        0         69.26546
## FAIXA_ETARIA                                       0         91.45182
## COR                                                0         39.65161
## DESCR_PROFISSAO                                    0        154.12578
## DESCR_GRAU_INSTRUCAO                               0         85.85713
```

‘ Fazendo previsões com o modelo, agrupando o resultado previsto e observado em uma tabela.

```
predict = predict(modelRF,testSample)
result = data.frame(predicted =  as.character(predict),
observed = as.character(testSample$RUBRICA))
```

Avaliando o modelo

Criando dois vetores. Um conterá todos os valores acertados pelo modelo e o outro os errados. A razão entre os valores acertados sobre o total de valores dá a precisão do modelo.

```
predictRight = filter(result,
as.character(predicted) == as.character(observed))
predictWrong = filter(result,
as.character(predicted) != as.character(observed))

accuracy = nrow(predictRight)/nrow(testSample)
print(accuracy)
```

```
## [1] 0.696456
```

Plota o índice Kappa de Fleiss para o modelo e detalha o mesmo índice para cada classe da previsão:

```
kappam.fleiss(result, detail = TRUE)
```

```
##  Fleiss' Kappa for m Raters
##
##  Subjects = 38098
##    Raters = 2
##     Kappa = 0.551
##
##         z = 186
##   p-value = 0
##
##                                                          Kappa
## A.I.-Homicídio simples (art. 121)                        0.075
```

```
## Drogas sem autorização ou em desacordo (Art.33, caput)             0.182
## Estupro (art. 213)                                                 0.017
## Estupro de vulneravel (art.217-A)                                  0.059
## Furto (art. 155)                                                   0.392
## Furto de coisa comum (art. 156)                                    0.000
## Furto qualificado (art. 155, §4o.)                                 0.244
## Homicídio culposo (art. 121, §3o.)                                 0.000
## Homicídio culposo na direção de veículo automotor (Art. 302)       0.040
## Homicídio qualificado (art. 121, §2o.)                            -0.001
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)            0.500
## Lesão corporal (art 129 § 9º)                                      0.676
## Lesão corporal culposa (art. 129. §6o.)                            0.047
## Lesão corporal culposa na direção de veículo automotor (Art. 303)  0.567
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                  0.000
## Lesão corporal seguida de morte (art. 129, §3o.)                   0.000
## Roubo (art. 157)                                                   0.638
##                                                                      z
## A.I.-Homicídio simples (art. 121)                                 14.704
## Drogas sem autorização ou em desacordo (Art.33, caput)            35.442
## Estupro (art. 213)                                                 3.272
## Estupro de vulneravel (art.217-A)                                 11.470
## Furto (art. 155)                                                  76.450
## Furto de coisa comum (art. 156)                                   -0.008
## Furto qualificado (art. 155, §4o.)                                47.687
## Homicídio culposo (art. 121, §3o.)                                -0.023
## Homicídio culposo na direção de veículo automotor (Art. 302)       7.898
## Homicídio qualificado (art. 121, §2o.)                            -0.221
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)           97.588
## Lesão corporal (art 129 § 9º)                                    131.912
## Lesão corporal culposa (art. 129. §6o.)                            9.121
## Lesão corporal culposa na direção de veículo automotor (Art. 303) 110.574
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                 -0.041
## Lesão corporal seguida de morte (art. 129, §3o.)                  -0.003
## Roubo (art. 157)                                                 124.485
##                                                                  p.value
## A.I.-Homicídio simples (art. 121)                                  0.000
## Drogas sem autorização ou em desacordo (Art.33, caput)             0.000
## Estupro (art. 213)                                                 0.001
## Estupro de vulneravel (art.217-A)                                  0.000
```

```
## Furto (art. 155)                                                  0.000
## Furto de coisa comum (art. 156)                                   0.994
## Furto qualificado (art. 155, §4o.)                                0.000
## Homicídio culposo (art. 121, §3o.)                                0.982
## Homicídio culposo na direção de veículo automotor (Art. 302)      0.000
## Homicídio qualificado (art. 121, §2o.)                            0.825
## Lesão corporal  de natureza GRAVÍSSIMA (art. 129, §2o.)           0.000
## Lesão corporal (art 129 § 9º)                                     0.000
## Lesão corporal culposa (art. 129. §6o.)                           0.000
## Lesão corporal culposa na direção de veículo automotor (Art. 303) 0.000
## Lesão corporal de natureza GRAVE (art. 129, §1o.)                 0.967
## Lesão corporal seguida de morte (art. 129, §3o.)                  0.998
## Roubo (art. 157)                                                  0.000
```