Universidade de Brasília

Faculdade de Economia, Administração e Contabilidade

Departamento de Administração

Curso de Graduação em Administração

WAGNER DA FONSECA NEIVA

UMA AVALIAÇÃO DOS MÉTODOS DE INTELIGÊNCIA ARTIFICIAL PARA A CLASSIFICAÇÃO DE EMPRESAS CORRUPTAS ATRAVÉS DA MODELAGEM DAS RELAÇÕES PÚBLICO-PRIVADAS

Brasília – DF

2018

WAGNER DA FONSECA NEIVA

UMA AVALIAÇÃO DOS MÉTODOS DE INTELIGÊNCIA ARTIFICIAL PARA A CLASSIFICAÇÃO DE EMPRESAS CORRUPTAS ATRAVÉS DA MODELAGEM DAS RELAÇÕES PÚBLICO-PRIVADAS

Monografia apresentada à Universidade de Brasília (UnB) como requisito parcial para obtenção do grau de Bacharel em Administração.

Professor Orientador: Doutor, Herbert Kimura

Brasília – DF

2018

Neiva, Wagner Fonseca.
Uma avaliação dos métodos de inteligência artificial para a classificação de empresas corruptas através da modelagem das relações público-privadas/ Wagner da Fonseca Neiva. – Brasília, DF, 2018.

Monografia (bacharelado) – Universidade de Brasília, Faculdade de Economia, Administração e Contabilidade- FACE, 2º Semestre 2018

Orientador: Prof. Herbert Kimura, Departamento de Administração.

1. Aprendizados de Máquina. Inidoneidade. Gestão Pública.

WAGNER DA FONSECA NEIVA

UMA AVALIAÇÃO DOS MÉTODOS DE INTELIGÊNCIA ARTIFICIAL PARA A CLASSIFICAÇÃO DE EMPRESAS CORRUPTAS ATRAVÉS DA MODELAGEM DAS RELAÇÕES PÚBLICO-PRIVADAS

A Comissão Examinadora, abaixo identificada, aprova o Trabalho de Conclusão do Curso de Administração da Universidade de Brasília do aluno

Wagner da Fonseca Neiva

Doutor, Herbert Kimura
Professor-Orientador

Doutor, Vinicius Amorim,
Professor-Examinador

Doutor, Marcelo Wilbert
Professor-Examinador

Brasília, 7 de dezembro de 2018

Dedico este trabalho a minha família e aos amigos que sempre estiveram ao meu lado.

AGRADECIMENTOS

É com um prazer imenso que eu entrego esse trabalho e neste momento eu realmente só tenho a agradecer por ter participado dessa incrível experiencia chamada graduação. Primeiro eu gostaria de agradecer todos os funcionários que trabalham diariamente para que isso seja possível, dos terceirizados até os professores, é incrível o que vocês proporcionam aos alunos apesar de todas as limitações da instituição e principalmente sem prestígio que vocês merecem.

Em seguida eu gostaria de agradecer a População Brasileira que bancou indiretamente os meus estudos através dos impostos cobrados pelo governo, esse trabalho é uma das minhas tentativas de retornar para vocês aquilo que foi investido em mim.

Não poderia deixar de agradecer toda minha família, a aqueles que me apoiaram desde do inicio e também aqueles que questionaram a minha escolha de curso, afinal, a dúvida me fez refletir sobre os meus objetivos e com isso ter a certeza da minha decisão.

Gostaria de agradecer também a Empresa Junior de Administração que teve um papel importantíssimo na minha formação, trazendo para a mim a prática dos conhecimentos adquiridos na faculdade.

Por fim a todos os meus amigos que participaram dessa jornada, aqueles que me apoiaram e aqueles que me desafiaram e até mesmo aqueles que só estavam lá. Posso dizer com toda certeza que o meu caminho não foi fácil e muito menos perfeito, mas poucas coisas boas na vida são assim.

“Todo espírito aventureiro se lançará para a conquista do difícil prêmio e se verá mais estimulado do que desencorajado pelas falhas de seus predecessores, porquanto espera que a glória de terminar uma aventura tão difícil lhe é reservada.”

(David Hume, Investigação acerca do entendimento humano)

RESUMO

A Inteligência Artificial e sua aplicação nos mais diversos ramos de atividades, vêm sido vigorosamente estudada pela academia internacional em busca de soluções transformadoras para os problemas reais enfrentados pelas organizações. Nesse estudo, inspirado no trabalho de Barboza (2017) na predição de insolvência, buscamos realizar testes com métodos de aprendizagem de máquinas (support vector machine, bagging, boosting, random forest e artificial neural network) com intuito de classificar virtualmente aquelas empresas que realizaram contratos fraudados com o Governo através do favorecimento de agentes públicos. Coletamos dados de todas as despesas efetuadas pelo Governo Federal, no período de 2011 a 2017 e relacionamos com o Cadastro de Empresas Inidôneas e Suspensas e o Cadastro Nacional de Empresas Punidas com o objetivo de modelar objetivamente o relacionamento dessas empresas com a esfera pública. Através dessa modelagem foi possível alcançar métricas interessantes e conseguimos observar um modesto destaque para os métodos baseados em aprendizagem de máquina em relação aos métodos estatísticos, porém, uma diferença significativa ao acrescentarmos todas as variáveis coletadas, verificando a dificuldade desses modelos de trabalhar com um número acentuado de variáveis, contribuindo para o debate levantado por Tsai, Hsu e Yen (2014) acerca da superioridade dos métodos computacionais. Nesse trabalho percebemos uma superioridade na acurácia geral do Support Vector Machine alcançando 89,41% de acertos em nossa amostra de teste, porém, observamos um ótimo desempenho de Bagging e Random Forest na minimização de Falsos Positivos, um fator importante na implementação desses métodos, conforme discutido neste trabalho.

Palavras-chave: Aprendizados de Máquina. Empresas Inidôneas. Gestão Pública.

ABSTRACT

Artificial Intelligence and its applications in the fields of study had been vigorously investigated by the international academy in the search of solutions for real problems faced by organizations. In this paper, inspired by the work of Barboza (2017) in the development of insolvency predicting models, we seek to perform series of tests with machine learning algorithms (Support Vector Machine, Bagging, Boosting, Random Forest and Artificial Neural Inteligence) with the goal to virtually classificate those companies that were involved in fraudulent contracts with the Brazilian Government. We collected data from public expenses of the Federal Government from 2011 to 2017 and crossed that data with the National Record of Punished Companies and the Record of Suspends and Indone Companies with the goal to create an objective model that would translate the public-private relationship. Through that model, we were capable to reach interesting results and it was noticed a minor advantage by the machine learning models in relation to traditional statistics techniques. However, a meaningful difference when added all the variables collected, showing the struggle of those models to work with a higher number of variables. This paper contributes to the debate raised by Tsai, Hsu and Yen about superiority of computational methods. In this work, we perceive that Support Vector Machine lead to higher accuracy rates in our testing samples, but, a great performance by Bagging and Random Forest minimizing Falses Positives results, a relevant factor in the implementation of those methods.

Keywords: Machine Learning. Inapt Companies. Public Management.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN – *Artificial Neural Network* (Redes Neurais Artificiais)

CEIS – Cadastro de Empresas Inidôneas ou Suspensas

CNEP – Cadastro Nacional de Empresas Punidas

NC – Número de Contratos

RPC – Receita Pública Total

SVM – *Support Vector Machine*

SIAFI – Sistema Integrado de Administração Financeira do Governo Federal

NDC – Natureza da Despesa Orçamentária

ROC - Receiver Operating Characteristic Curve

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

O físico, Stephen Hawking, em uma palestra realizada na Universidade de Cambridge durante a inauguração do Centro Leverhulme para o Futuro da Inteligência, declarou suas ressalvas em relação aos recentes avanços significativos no campo da inteligência artificial. “IA pode ser a melhor coisa ou a pior coisa para a humanidade” essas foram as palavras do cosmólogo britânico, tal preocupação não foi uma surpresa para aqueles que o conhecem visto que não é a primeira vez que o próprio declara publicamente as suas preocupações com o desenvolvimento dessa tecnologia.

Suas preocupações derivam principalmente de dois fatores: a velocidade como os robôs aprendem e a forma como essa tecnologia será utilizada. O primeiro fator diz respeito a preocupação de que a construção de um sistema consciente com a capacidade de aprendizado semelhante aos seres humanos pode ameaçar a sua sobrevivência em detrimento da superioridade robótica na velocidade de processamento das informações, já o segundo fator diz respeito as intenções dos detentores dessa tecnologia e que tal ferramenta poderia ser utilizado como “novas formas para que os poucos oprimam muitos”. O presente artigo busca abordar o segundo fator.

Dentro dos principais componentes responsáveis pela ascensão da Inteligência Artificial nos últimos anos, algoritmos de aprendizagem de máquinas (*Machine Learning)* tem se apresentado como uma importante linha de pesquisa transversal aos diferentes temas da literatura acadêmica. Sua característica transversal se fundamenta em seu grande portfólio de aplicabilidades, desde campo da medicina aonde Sharma (2017) chegou a resultados pertinentes no diagnóstico de doenças cardíacas utilizando a vasta quantidade de dados acumulados pela indústria da saúde, até o campo das finanças onde temos diversos estudos relevantes como Whiting (2012) que buscou elaborar um modelo capaz de detectar fraude financeira de

grandes empresas pertencentes ao mercado financeiro utilizando somente a informação pública dessas corporações e Barboza (2017) que buscou realizar uma comparação entre os modelos estatísticos tradicionais (análise discriminante e regressão logística) utilizados para a predição de falência bancaria e inadimplência, e os modelos de aprendizados de máquinas (support vector machines, bagging, boosting e random forest) onde o estudo deflagrou uma superioridade relevante dos novos modelos sobre os tradicionais além de ter alcançado uma precisão significativa.

Para clarear o conceito, Jordan (2015) indica que de maneira sucinta que aprendizado de máquinas se dirige a possibilidade de construirmos máquinas que melhoram a sua performance automaticamente através da experiência e que os motivos para que esse tema tenha tido tanto progresso recentemente não se esgota aos recentes avanços em desenvolvimento de algoritmos. Dentre os principais fatores que permitiram tamanho progresso, temos os avanços no campo da computação que aumentaram exponencialmente a capacidade de processamento das máquinas que por consequência reduziram o custo destes equipamentos e o crescimento acelerado na última década da produção digital de informação gerada principalmente pelas pessoas, pelas máquinas digitais (*Internet of Things*) e as empresas por meio de *Business Intelligence*. Esse fenômeno que possibilita reunir, transportar e processar uma vasta quantidade de dados é delineado por Jordan (2015), entre outros autores, de *Big Data.*

Podemos observar a sua aplicabilidade como uma aproximação matemática da realidade, onde uma determinada tarefa é convertida em uma função cuja variável dependente costuma a ser um valor classificatório, que varia em função de diversos fatores determinados pelo modelo. No trabalho realizado por Sharma (2017) temos um modelo que busca identificar doença cardíaca nos pacientes, nesse caso a variável dependente (y) seria a existência ou não de doenças cardíacas que assumiria o valor de 0 ou 1 com base nos valores dos atributos mapeados no modelo como a idade, colesterol, pressão sanguínea, sexo e etc. Após a sua construção, o modelo será considerado valido com base na sua precisão e para isso é necessário realizar o processo de calibragem através de um conjunto de dados de treinamento onde o modelo é exposto a uma extensa série de casos onde para cada conjunto de atributos ele deve determinar a existência ou não da doença cardíaca, para cada tentativa o sistema recebe um feedback indicando a necessidade de alterações nos parâmetros

definidos previamente. Esse é um exemplo de utilização, com a crescente demanda de tecnologias dessa estrutura, aos mais diferentes casos, resultou em uma série de algoritmos que permitem uma maior flexibilidade como diferentes classificações e até mesmo a realização do processo de aprendizado sem supervisão.

Por isso, essa tecnologia se apresenta como uma ótima alternativa para potencializar os resultados alcançados pelos órgãos de controle, principalmente para aqueles que atuam no Brasil, pois, ao observamos a realidade brasileira, devemos nos atentar a imensidão apresentada pelo nosso vasto território nacional, o que consequentemente leva a um número considerável de serviços realizados pela esfera pública que, em sua grande maioria, contam com a participação do mercado privado para atingir esses objetivos. Com isso, temos uma quantidade extraordinária de licitações para serem fiscalizadas pela esfera pública, somadas ao extenso número de organizações públicas que prestam contas anualmente para órgãos de controle, não é possível analisar profundamente todas essas contas dentro de um tempo hábil para que elas possam ser julgadas e aprovadas pelos Tribunais de Conta sem prejudicar o andamento do processo e por isso essas instituições buscam priorizar os casos mais sensíveis a esfera pública, essa incapacidade na fiscalização pública levou o país a nonagésima sexta posição no Índice de Percepção da Corrupção em 2017, o ranking foi realizado pela Transparência Internacional.

Esse excesso de informação que se apresenta como uma grande dificuldade para essas instituições, se torna um fator essencial para a aplicação dos algoritmos de aprendizados de máquina, visto que existe uma relação direta entre a quantidade de informação disponível para treinamento e a eficácia desses métodos. Vale ressaltar que esses modelos não devem substituir os profissionais responsáveis pelas auditorias e sim servir como ferramentas pertinentes para o seu contexto de trabalho, reduzindo o universo de controle desses profissionais, realizando uma priorização imparcial dos casos mais sensíveis a gestão pública, possibilitando a análise aprofundada dos casos que se apresentam como um risco as contas públicas.

1.2 Formulação do problema

Pela a sua capacidade de identificar padrões e perfis dentro uma vasta série de dados, essa tecnologia desempenha atualmente um papel crucial na identificação de fraude em transações financeiras de cartões de crédito. Esse tema foi abordado por diversos autores que alcançaram resultados significativos para a academia e para os bancos que aplicam esse conhecimento, Bekirev (2015) foi um dos primeiros a se aproximar de um modelo através do uso de redes neurais e clusterização, Dal Pozzolo (2015) buscou elaborar um modelo para detectar transações fraudulosas, que se calibrasse automaticamente de acordo com as transações em tempo real e Wedge (2017) teve uma grande contribuição a esse processo ao dedicar seu trabalho à redução de resultados falsos positivos gerados pelos modelos.

Tendo em vista o avanço significativo dessas tecnologias para detectar em tempo real uma transação financeira que não tenha sido realizado pelo usuário do cartão de crédito, seria possível a construção de um modelo que pudesse detectar em tempo real a realização de uma transação financeira de um órgão público para uma empresa privada, que se caracterizasse como um desvio de verba pública e se possível, qual método seria o mais apropriado para essa realidade.

1.3 Objetivo Geral

Neste artigo, buscaremos construir um modelo, através da utilização de algoritmos classificatórios de aprendizado de máquinas (bagging, boosting, random forest) com o objetivo de classificar virtualmente aquelas empresas que realizaram contratos fraudados com o Governo. Tal modelo pode ser utilizado semelhantemente a forma como os modelos de Dal Pozzolo (2015) e Wedge (2017) foram implementados, onde existe uma equipe responsável por constantemente analisar os resultados com o objetivo de determinar se um alerta gerado pelo modelo em uma nova transação representa ou não um falso positivo.

Por fim, este trabalho busca contribuir para a literatura acadêmica, levantando alternativas para a construção de novos modelos de aprendizado de máquinas e novas formas para potencializar a gestão pública com objetivo de que ela preste serviços de qualidade e com eficiência ao cidadão. Após a introdução, este artigo apresentará no **Capítulo 2** o Referencial teórico onde será discutido brevemente os algoritmos de aprendizado de máquinas que serão utilizados no decorrer desse trabalho, seguido pelo **Capítulo 3** onde será apresentado a metodologia do estudo e no **Capítulo 4** quais dados foram utilizados em sua realização. A discussão dos resultados apresentados pelos modelos se encontra no **Capítulo 5** e ao final o **Capítulo 6** para a desfecho do trabalho onde será apresentado as principais conclusões com o estudo e suas ressalvas.

1.4 Objetivos Específicos

I. Levantar os dados relevantes para a modelagem das relações Público-Privadas entre as empresas licitantes e o Governo Federal.
II. Compilar e tratar os dados para que eles estejam no formato exigido dos algoritmos de inteligência artificial.
III. A elaboração de um modelo que possa traduzir as relações público-privadas em uma configuração que possa ser analisada pelos métodos de classificação.
IV. A realização dos testes e a apuração desses resultados com o intuito de verificar qual a melhor abordagem para o problema proposto.
V. Analisar os resultados e propor os algoritmos mais adequados para o problema elaborado.

1.5 Justificativa

Esse tópico se apresenta como uma necessidade latente para o desenvolvimento do cenário econômico nacional, tendo em vista a quantidade de escândalos deflagrados por toda esfera pública brasileira após os eventos internacionais da Copa do Mundo e as Olimpíadas. É evidente que a corrupção, assim definida pelo World Bank como abuso do poder público praticado pelo agente para

benefício próprio, prejudica a sociedade pelos seus efeitos na redução dos insumos públicos, levando a redução na produtividade do capital privado, porém, tais escândalos colocaram a crise política brasileira na vitrine mundial, visto que vários escândalos tiveram repercussão pelos principais canais da mídia internacional o que leva a impactos adicionais à economia brasileira.

O trabalho de Campos e Pereira (2016) que tem como objetivo analisar a corrupção e a ineficiência no setor público evidencia o efeito da corrupção sobre a redução dos investimentos e do produto ao longo prazo. A teoria econômica que busca dar fundamento para tais conclusões, esclarece que tais atos aumentam o risco atribuído ao país em decorrência das despesas que encareceria as inversões, diminuindo a taxa de retorno dos investimentos o que desestimularia investidores estrangeiros a aplicar o capital no país em questão, além de levar o mercado privado interno a evitar investimentos. Neste sentido, o investimento em tecnologias que busquem minimizar a ocorrência de desvios de verbas publica poderia ser uma sinalização atrativa para o mercado internacional que possibilitaria o aumento dos investimentos, algo importante para o país visto que as previsões para os próximos anos mostram um crescimento discreto para economia.

2 REFERENCIAL TEÓRICO

Os algoritmos de aprendizado de máquinas desempenham um papel essencial para a sociedade contemporânea, é uma tecnologia que se encontra presente na vida das pessoas mesmo sem o conhecimento de tal. Além das aplicações discutidas na introdução, podemos presenciar o uso desses mecanismos na separação do correio eletrônico entre um e-mail verídico e um spam, nas recomendações de produtos em sites de vendas, no reconhecimento facial dos smartphones, em carros autônomos e na detecção de fraude financeira Whiting (2012)

Para que esse processo aconteça é necessário que o sistema aprenda através de uma série de observações onde o algoritmo irá extrair automaticamente desses dados padrões de comportamento e aplicar esse conhecimento para predição de novos efeitos, Dal Pozzolo (2015) definiu aprendizado de máquinas como o processo de extrair conhecimento a partir de dados. Nesse estudo os mecanismos de aprendizado de máquina serão utilizados para distinguir uma empresa fraudulenta, que por consequência será classificada como inidônea, de uma empresa não fraudulenta, com base em suas transações financeiras realizadas com a União.

No próximo tópico, será discutido brevemente o funcionamento dos algoritmos *Support vector Machines*, *Bagging*, *Boosting*, *Random Forest* e a utilização de métodos tradicionais de redes neurais artificiais como regressão logística e analise discriminante. Posteriormente, no tópico de discussões, o desempenho desses algoritmos será comparado para que assim tenhamos conhecimento do mais adequado para executar a determinada tarefa.

2.1 *Support Vector Machine*

O modelo Support vector Machine (SVM), assim explicitado por Meyer (2017), é centralizado nos vetores de suporte, tais vetores representam as observações dentro de uma sequência de dados que cujas classificações são distintas (Fraudulenta ou não Fraudulenta) porém possuem a menor distância gráfica entre si, ou seja, resumidamente as observações mais semelhantes que apresentam classificações

distintas. Basicamente, o modelo consiste em uma função matemática que busca delimitar uma linha que separa as observações em diferentes classes, essa função será otimizada com base na função conhecida como “kernel” que buscara maximizar a distância entre os vetores de suporte, aumentando a margem que separa as classes.

Pela estrutura do seu funcionamento, o modelo SVM costuma a ser utilizado para classificações binárias, embora não seja uma limitação para seu funcionamento visto que existem alternativas que flexibilizam o modelo. Um critério que merece destaque que em casos onde grupos são independentes, é possível alcançar uma precisão de até 100% embora não seja o caso deste trabalho visto que o objeto de estudo consiste em transações financeiras e empresas de mercado, que sofrem influência de infinitos fatores. Para isso o método SVM permite a inclusão de uma margem de erro (Zhou et al., 2014).

Em sua aplicação, o algoritmo que busca determinar a partir de uma série de dados quantitativos a classificação adequada. Para isso terá os seus parâmetros ajustados durante o conjunto de dados para treinamento, após ter os parâmetros ajustados pelo conjunto de treinamento, o modelo será exposto a um conjunto de dados de validação, totalmente independente do conjunto de treinamento, onde as classificações determinadas pelo modelo serão comparadas as classificações corretas, buscando estimar a sua precisão. A seguir, temos a estrutura matemática do modelo de otimização, de forma resumida por Li, Wang e He (2013).

Minimize: $\frac{1}{2}w^T w + C\sum_{i=1}^{M}\xi_i,$ (1)

Sujeito à: $y_i[w^T \emptyset(X_i) + b] \geq 1 - \xi_i$ (2)

Sendo que i = 1,2,...,M, $\xi_i \geq 0$ representam a margem de erro em relação a classificação estimada C e o y_i representa as classificações do conjunto de dados de treinamento, vale ressaltar que o termo Ø(x) não precisa ser necessariamente conhecido visto que de acordo com a aplicação da função de kernel

$\big(K(x) = K(Xi, Xj)\big) \quad \therefore \quad K(x) = \emptyset(Xi)^T . \emptyset(Xj)$ (1)

Sendo que:

$$K(Xi, Xj) = e^{(-y||Xi-Xj||^2)} \quad (2)$$

Onde a primeira equação representa a equação de kernel em sua forma linear, que possui a limitação de não oferecer uma precisão atraente quando se trata de modelos que utilizam dados não independentes e para isso existe a adaptação conhecida como “Função de base radial”. Por fim, deve-se salientar que y é uma constante com valor.

2.2 *Bagging*

O algoritmo *Bagging* (Bootstrap Aggregating) consiste em um método classificatório que funciona através da construção de diversos classificadores. Proposto por Breiman (1996), o método gera a partir do conjunto de dados de treinamento, uma série de subconjuntos onde essas observações são reorganizadas de forma aleatória, assim o algoritmo constrói um modelo de classificação para cada um desses subconjuntos, ao final do treinamento, quando o algoritmo estará pronto para processar os dados desejados, a classificação será realizada através de um sistema de votação que irá calcular a classificação que recebeu o maior número de votos, ou seja, a classificação prevista pela maior quantidade de modelos

A seguir, o seu funcionamento matemático:

1. Um subconjunto (bootstrap), t, é gerado a partir do banco de dados de treinamento
2. Um modelo classificatório C é construído a partir do subconjunto t, gerado no primeiro passo
3. O primeiro e o segundo passo são repetidos para t = 1, ... ,T
4. Cada modelo classificatório realiza um voto.

$$C(X) = T^{-1} \sum_{t=1}^{T} C_t(X)$$

2.3 *Boosting*

Boosting consiste em um algoritmo de aprendizado de máquina muito utilizado pela academia, Duarte (2009) descreve como uma técnica que utiliza uma combinação de modelos de classificação para construir um modelo final com maior precisão, tais modelos utilizados na combinação são conhecidos como modelos fracos. Freund (1999) relata que em seu funcionamento, o algoritmo expõe os modelos fracos a um subconjunto do conjunto de dados de treinamento, onde cada classificador buscará estimar a classe adequada dos dados com base em seus parâmetros, após a interação, os pesos de cada classificador são ajustados com base em sua taxa de erro e um novo modelo é construído. Com o objetivo de ampliar a precisão do modelo final, Freund e Schapire (1995) apresentaram o algoritmo *AdaBoost* que consiste essencialmente em uma ponderação das observações de acordo com o seu nível de dificuldade de classificação, forçando os modelos fracos a focarem nos casos mais sensíveis.

Inicialmente, todas as observações recebem um peso uniforme *1/m* onde m consiste no número total de observações, posteriormente uma amostra aleatória de treinamento é gerada para ser classificada pelos modelos fracos, a partir do resultado desse treinamento uma taxa de erro é calculada com base no número total de observações na amostra, caso essa taxa de erro seja maior do que a taxa de erro de uma série de palpites aleatórios, o subconjunto é descartado e o algoritmo processa um novo para realizar o treinamento, com os pesos originais. Caso a taxa de erro seja satisfatória o algoritmo irá atualizar os pesos de cada observação da amostra, definidos previamente como *1/m,* com base em quantos modelos classificaram incorretamente a observação e o peso de cada classificador que errou em sua estimação.

Utilizando o trabalho de Heoand Yang (2014), segue a descrição do algoritmo:

1. Realize uma distribuição de pesos, Wi(i) = 1/m é gerada, onde i = 1, 2,. . . . ., m; e Wt é a ponderação interativa (t = 1, ..., T), $W_{t+1}(i) = \frac{W_t(i)e^{a_t(2I(y_i \neq h_t)-1)}}{W_t(i)e^{a_t I(y_i \neq h_t)}}$, onde

Ht = argmax|0.5 - §t| é o erro tal que §t = $\sum_{t=1}^{m} wt(t)\, I(y_t \neq H_t(Xt))$| e I = 1 onde a medida de precisão é computada, caso contrário , 0

2. Em cada ciclo,$\alpha_t = \frac{1}{2}\ln(\frac{1-§t}{§t})$ é recalculado

3. A rotina é concluída quando |0.5 - $§_t$ | ≠ δ , onde δ é uma constante predefinida

4. Y(x) é avaliado para boost completo, onde: $Y(x) = sign \sum_{t=1}^{T} \alpha Ht(x)$

2.4 *Random Forest*

Proposta inicialmente por Breiman (2001), o modelo de classificação generalizado que utiliza algoritmos de aprendizado de máquina para gerar uma quantidade numerosa de modelos classificatórios, onde as amostras apresentam uma baixa correlação entre si, se mostrou uma técnica atrativa para problemas de aprendizado devido ao seu alto nível de precisão semelhante ao AdaBoost e em alguns casos, superior ao *Boosting* (Kruppa al., 2013). A principal diferença entre os modelos *Bagging* e *Random Forest*, é que ao extrair as observações do conjunto de treinamento com o intuito de gerar o subconjunto de dado Xi para o modelo F(xi), tal extração é determinada com base em um vetor de características especificas, definido de forma aleatória, resultando em uma baixa correlação entre os subconjuntos de dados Xi, onde i = 1, . . ., N.

Assim como no algoritmo proposto por Breiman (1996), após o treinamento dos modelos classificatórios, a classificação final para determinada observação será adquirida através da votação de cada um dos modelos, também conhecido como arvores classificatórias, a principal força do modelo se encontra na utilização de um vetor de características aleatórias na construção dos subconjuntos de dados, tal mecanismo aumenta a precisão do modelo e minimiza a ocorrência de sobre-ajuste. Além disso, o modelo apresenta o peso de cada variável na determinação de uma classificação, assim como na regressão linear, tal modelo segue uma estrutura matemática de profunda complexidade e sua discussão pode ser avaliada nos trabalhos de Booth al. (2014); Breiman (2001).

Yehet (2014) descreve o algoritmo como:

1. Criam-se subconjuntos a partir do conjunto de dados de treinamento, de forma aleatória, cujo número de observações é arbitrário e as características são determinadas por um vetor especifico a cada subgrupo.
2. Para cada subconjunto de dados é construído uma arvore classificatória, e todos os elementos de um conjunto recebem uma classificação (correta ou não)
3. Para cada elemento, é contabilizado as votações de todas as arvores, a classificação preferida é a escolhida

2.5 Redes Neurais Artificiais

Conforme Vellasco (2007) as redes neurais artificiais consistem em modelos utilizados na resolução de problemas complexos não-lineares, empregado principalmente em casos onde a solução não pode ser determinada por uma série de passos simples a serem seguidos por um algoritmo. Como o nome da a entender, o modelo se inspira no funcionamento dos neurônios do cérebro humano, tal vertente vem sendo estudado desde do século passado, a ideia foi concebida por Warren Mc Culloch e Walter Pitts que desenvolveram uma máquina inspirada no cérebro humano, denominada Psychon, porém, diferentemente dos modelos atuais, Psychon não possuí a capacidade de aprender pela experiência. Desde então, diversos avanços foram realizados neste campo de pesquisa resultando na criação de diferentes métodos, entre eles o algoritmo de retro propagação, que será utilizado neste trabalho, consiste em um dos algoritmos de treinamento supervisionado mais difundidos, cujo funcionamento utiliza a regra de correção do erro para que o sistema aprenda através da experiência.

Segundo Kshirsagar (2012 esse sistema é operacionalizado através de camadas de neurônios que comunicam entre si, um neurônio possuí a capacidade de receber informações através de seus conectores que são ponderadas através de pesos pré-determinados, essa informação será processada por uma função não-linear que resultará em um output, dependendo da camada em que esse neurônio se encontra, esse output será parte do resultado final da classificação ou a informação

de entrada para outro neurônio. O processo de retropropagação Vellasco (2007) verifica o erro entre o resultado final e o resultado esperado e modifica os pesos com a intenção de minimizar o erro da próxima saída, tal ajuste segue um método baseado no gradiente descendente que busca fazer modificações proporcionais ao gradiente do erro.

Neste estudo, a primeira camada responsável por receber as informações, receberá as variáveis do modelo com base nas transações financeiras de determinada empresa com o Estado, essa informação será entregue à uma camada intermediaria conhecida como camada oculta responsáveis por extrair as características mais significativas dos padrões de entrada, por fim a camada de saída processará a informação da camada oculta e irá produzir um resultado que consiste na classificação ou não de uma empresa como inidônea. Tal método possuí similaridade com o trabalho de Zhaoet al.(2014) que identificou certas limitações na técnica como baixa performance ao trabalhar com dados não balanceados, outra limitação relatada por Kshirsagar (2012) se encontra no fato que o sistema de redes neurais possuí o seu funcionamento de forma autônoma, o que inibe analises profundas sobre o método identificado pelas redes neurais para solucionar o problema.

2.6 Regressão Logística

As análises discriminantes e regressão logística são métodos tradicionais que trabalham com a minimização da variância dentro dos grupos classificatórios, apesar de terem sido um grande avanço no campo da estatística entre outros como Altman (1968) que apresentou um estudo com resultados relevantes na predição de falência bancária em empresas de manufaturados, os modelos de aprendizado de máquinas costumam a apresentar uma superioridade tanto na precisão como na eficiência, podemos observar esse fenômeno em diversos estudos como os de Kruppa et al. 2013 e Trustorff et al. 2010 que realizaram uma comparação entre os métodos para a classificação de risco de crédito. O método de análise discriminante produz um resultado para cada observação, esse resultado será movimentado para determinada classificação com base em um valor de corte arbitrário, um detalhe importante desse

modelo é que ele utiliza pressupostos como o de normalidade em sua utilização, o que se mostrou ser uma vantagem para a regressão logística pela sua maior flexibilização e um resultado de zero a um que condiz com a probabilidade de determinada observação ser a classificação indagada.

3 MÉTODOS E TÉCNICAS DE PESQUISA

O método utilizado na pesquisa se inspira na metodologia de Barboza (2017), por isso escolhemos realizar experimentos com diversas técnicas de aprendizado de máquinas com o intuito de identificar aquela com o melhor desempenho de acordo com o problema exposto. Para esse trabalho, utilizamos as seguintes técnicas

- *Bagging*
- *Boosting*
- *Random Forest*
- *Support Vector Machine*
- *Artificial Neural Network*
- Regressão Logística
- Analise Discriminante

Na realização do experimento com a técnica de *Support Vector Machine* foi utilizado a formula de kernel em sua forma linear e radial Os modelos deveriam julgar a probabilidade de uma empresa se tornar inidônea no período seguinte e utilizaria como base para esse julgamento as transações financeiras realizadas com os órgãos da esfera pública federal, onde o algoritmo deve analisar o comportamento dessas despesas, assim como as características intrínsecas de sua natureza como a origem dos recursos e o seu propósito oficial.

Esse estudo foi realizado através do software livre R por meio de seus pacotes estatísticos que se encontram em sua biblioteca padrão, foram eles: Ada, e1071, h2o, bagging, gbm, mboost, random forest, MASS, aod e nnet. Vale ressaltar que com a utilização dos pacotes estatísticos tradicionais de um software livre somado ao fato de que esse estudo foi realizado em dados abertos a população, torna a sua reprodução acessível para qualquer pesquisador que tenha pleno domínio das ferramentas estatísticas e o conhecimento sobre os principais algoritmos de aprendizado de máquinas. Para a realização dos estudados foram coletados dados a respeito das empresas que transacionaram com o Governo Federal (Ministérios, Presidência da República e o Gabinete da Vice-Presidência) no período de 2011 até 2017, isso resultou em um total de 305 mil empresas, essas empresas foram analisadas com base quesitos presentes na Tabela 1.

Essa tabela apresenta todas as variáveis que foram utilizadas n modelo. As variáveis categóricas foram apresentadas a partir da transformação dessas variáveis em *Dummy Variables*.

Tabela 1 - Variáveis quantitativas e qualitativas do modelo

Variáveis Quantitativas	Fórmula/Descrição
RPT	$\sum_{i=1}^{n} R = Receita\ Pública\ Total$
NC	$\sum_{i=1}^{n} \frac{R}{R} = Nº\ de\ Contratos$
Variáveis Categóricas	**Descrição**
Carteira de Cliente	Relação entre a empresa e os ministérios
Natureza da Despesa Orçamentária	Classificação Contábil para o Serviço da Empresa
Função	Função estratégica do estado ao qual a empresa prestava o serviço
Setor de Atuação	Tipo de serviço prestado pela organização

Nessa tabela podemos ver as variáveis responsáveis pelo modelo, temos Receita Pública Total (RPT), Número de Contratos (NC) como as nossas variáveis quantitativas e Carteira dos Clientes, Natureza da Despesa Orçamentária, Função e Setor de Atuação são as variáveis qualitativas do modelo que passaram por um processo de transformação, onde elas se tornaram Dummy Variables para que os

algoritmos fossem capaz de identificá-las. Como variável dependente utilizamos o Cadastro de Empresas Inidôneas e Suspensas (CEIS) e o Cadastro Nacional de Empresas Punidas (CNEP), ambos presentes no Portal da Transparência (www.portaltransparencia.gov.br), uma plataforma aberta para a população de dados que dizem respeito a gestão pública nacional. Esse Banco de dado apresentou mil e quinhentas e cinquenta e duas empresas que negociaram com os órgãos do Governo Federal, e foram categorizadas como empresas problemáticas e utilizadas no nosso modelo.

Conforme abordado por Tang (2009), os algoritmos classificatórios apresentam um desempenho limitado ao trabalharem com um conjunto de observações desbalanceado. Tendo em vista que a nossa amostra coletada possui um baixo número de empresas problemáticas a serem analisadas pelo modelo, o espaço amostral total foi reduzido para 2799 observações, onde primeiramente descartamos aquelas empresas que apresentaram um valor de receita inferior a dez mil e posteriormente buscamos selecionar aleatoriamente empresas presentes em nossa amostra, preservando um número considerável em empresas problemáticas para serem analisadas pelos algoritmos de inteligência artificial. O conjunto de dados de treinamento foi construído através de 2099 empresas, contendo 1024 empresas inidôneas, em contrapartida o nosso conjunto de dados teste buscou trabalhar com 700 empresas, sendo que 700 foram classificadas pelo Ministério da Transparência e Controladoria Geral da União como inidôneas.

Com o objetivo de potencializar os resultados dos nossos algoritmos de predição, buscamos realizar o processo de normalização de todas a variáveis presentes no modelo. Na Tabela 2 podemos analisar as informações relativa as estatísticas descritivas das variáveis quantitativas presentes no modelo e as correlações entre as variáveis foram exploradas na Tabela 3 e na Tabela 4. Para medir o desempenho dos algoritmos buscamos apurar a acurácia geral de cada um dos métodos, buscando contabilizar a porcentagem de acertos e levantamos o número de Falsos Positivos e Falsos Negativos.

Tabela 2 - Estatística Descritiva a respeito das Variáveis Quantitativas

Indicadores	Receita Pública Total	Nº de Contratos
Média	R$ 254.241.934	3114
Desvio Padrão	R$ 9.130.891.762	23982
Mínimo	R$ 10.008	1
Máximo	R$ 442.441.000.000	904996
Mediana	R$ 1.105.304	234
1º Quar	R$ 195.194	64
3º Quar	R$ 5.778.249	1085

Tabela 3 - Correlações das Variáveis no Modelo Completo

Modelo Completo	RPT	NC	NDC	Ministérios	Função	Setor de Atuação
RPT	1					
NC	,378**	1				
NDC	,074**	,138	1			
Ministérios	,092**	,299**	,407**	1		
Função	,077**	,257**	,437**	,958**	1	
Setor de Atuação	,277**	,424**	,401**	,544*	,496**	1

Observação: *** p<0.01*

Tabela 4 - Correlações das Variáveis no Modelo Enxuto

Modelo Enxuto	RPT	NC	NDC	Ministérios	Função	Setor de Atuação
RPT	1					
NC	,378**	1				
Inversões Financeira	,183**	,192**	1			
Outras Despesas Corrente	,000	,000	,000	1		
Investimentos	,000	,085**	,094**	-,191**	1	
Ministérios	,092**	,299**	,258**	,151**	,306**	1
Observação:	** p<0.01					

4 DADOS

A coleta dos dados foi possível graças a Lei da Transparência que determina a obrigação perante a gestão fiscal nacional de disponibilizar em tempo real as informações pormenorizadas sobre a execução orçamentária e financeira da União, dos Estados, do Distrito Federal e dos municípios. As transações financeiras foram extraídas do portal da transparência, gerenciado pelo Ministério da Transparência e Controladoria-Geral da União, tais transações são referentes aos pagamentos com aquisições e contratações de obras, compras governamentais dentre outras naturezas, tais transações contemplam uma série de variáveis qualitativas para serem analisadas pelo modelo, são elas:

1. O órgão responsável por autorizar a transação;
2. A secretária interna ao órgão responsável pela transação;
3. A categoria financeira em que a transação se encaixa;
4. A categoria genérica do gasto;
5. A função interna da determinada transação;
6. A função específica da transação;
7. O programa de governo que resultou na determinada transação;
8. A ação interna que resultou no gasto;
9. A empresa que prestou o serviço;
10. Outras informações para fins de documentação

A origem das informações são do Sistema Integrado de Administração Financeira do Governo Federal (Siafi) que consiste no principal instrumento utilizado para registro, acompanhamento e controle da execução orçamentária, financeira do governo federal, sua coletânea de informações fora dividida mensalmente e a disponibilidade dos dados são de janeiro de 2011 até dezembro de 2017, ao total, os dados incluem cerca de 99 milhões de transações. Esse banco de dados foi o insumo principal para a construção da tabela final a ser trabalhada pelo modelo, essa tabela buscou consolidar as principais informações a respeito das empresas que negociam com a esfera federal que serão posteriormente estudadas pelos algoritmos de

inteligência artificial, onde esses mecanismos vão identificar possíveis correlações entre essas informações e a inidoneidade das empresas.

Todas as noventa e nove milhões de transações foram armazenadas no MySQL, onde foi possível consolidar os registros e manipular o banco de dados como um todo, permitindo as alterações necessárias, como a remoção de registros cujas as propriedades são ocultadas por questão de segurança nacional. Com isso, foi possível extrair somente aquelas transações, cujo o favorecido era uma organização com CNPJ, deixando de lado as instituições públicas e outras organizações da esfera pública.

Com isso foi possível a consolidação de um banco de dados com trezentas e cinco mil e trezentas empresas, onde o relacionamento de todas essas empresas com a esfera federal pode ser facilmente analisado, levando com consideração os ministérios que negociam com essas empresas e os tipos de serviços que são prestados por ela. Dentre todas as informações presentes no banco de dados, somente algumas foram aproveitadas com o intuito de potencializar o nosso modelo de predição, devido à natureza abrangente de algumas variáveis, onde o grande número de possibilidades prejudicaria a sua efetividade.

Para esse banco de dados, priorizamos as seguintes variáveis independentes:

1. Receita Pública Total
2. Nº de Contratos
3. A carteira de clientes
4. Natureza da Despesa Orçamentária
5. Função
6. Setor de atuação

4.1 Receita Pública Total

O primeiro fator a ser analisado é o volume de recursos que foi destinada aquela organização e como isso foi relatado oficialmente na transparência federal.

Essa variável pode identificar facilmente os grandes contratos presentes no relacionamento público-privado, levando o nosso modelo a analisar os casos de maior impacto no orçamento público e que por consequência são de maior interesse aos agentes privados.

As grandes licitações são aquelas que costumam a atrair um maior número interessados e por isso são sucintas a possíveis negociações ilegais entre agentes públicos e privados, onde ocorre o enriquecimento ilícito do agente público em troca de um favorecimento a organização prestadora de serviço. Por esse favorecimento, temos uma quebra do processo de licitação tradicional, onde o agente público facilita a vitória da empresa, através dos mecanismos que estão sobre sua competência.

Para isso, para cada CNPJ dentro do banco de dados, foi levantado todas as transferências de capital entre a empresa favorecida e o governo federal, do período de 2011 a 2017 e através de um somatório, levando em consideração o valor de cada uma dessas transações, podemos identificar a **Receita Pública Total** de cada uma dessas empresas nesses sete anos.

$$\sum_{I=1}^{n} R = Receita\ Pública\ Total$$

4.2 Número de Contratos

Para que ocorra o enriquecimento ilícito do agente público, por conta de um favorecimento a uma organização no processo licitatório, é necessário que esse agente público seja capaz de burlar o sistema licitatório, criando métodos de beneficiar a empresa perante a sua concorrência tornando-a a mais elegível perante os requisitos pré-estabelecidos. No trabalho elaborado pelo Centro de Pesquisa de Corrupção de Budapeste Mihály (2013), podemos observar uma árdua tentativa de encontrar um método efetivo de identificar a corrupção através de um indicador objetivo e quantitativo visto que tal comportamento é extremamente imperceptível para os dados brutos e por isso, em sua grande maioria, os indicadores de corrupção

possuem um caráter subjetivo como observado em Rohwer, Anja (2009) e no *User´s Guide to Measuring Corruption* realizado pela *Global Integrity*.

Para isso, Rohwer (2009) declara que a eliminação da concorrência seria um pré-requisito para a ocorrência deste processo e por isso teve o seu trabalho direcionado ao estudo do processo licitatório, observando as características intrínsecas da concorrência como o número de licitantes permitidos, a determinação dos pré-requisitos e as alterações das condições de performance após a determinação do contrato. Para esse trabalho, não foi possível coletar informações referentes ao processo licitatório de cada uma dessas empresas, mas tendo como foco o estudo da concorrência dentro da gestão pública federal, foi identificado o número total de transações entre a organização analisada e o governo federal, levando em consideração todas as suas instâncias.

O objetivo é de que o modelo possa verificar uma possível correlação entre o número de contratos premiados a uma única empresa e a possibilidade dessa empresa ter um relacionamento ilícito com algum agente público, onde ele estaria facilitando o processo em virtude desse relacionamento, eliminando a concorrência de outras empresas. Para isso, foi realizado um processo automatizado, onde para cada uma das trezentas e cinco mil e trezentas empresas, foi levantada o número de vezes que essa organização recebeu recursos públicos federais e com isso, poderíamos medir a presença dessa empresa dentro da gestão financeira nacional, tendo como esperança, que as empresas favorecidas de forma ilegal, tenham uma presença assimétrica perante as empresas regulares.

anos.

$$\sum_{I=1}^{n} \frac{R}{R} = Nº\ de\ Contratos$$

4.3 A carteira de Clientes

O foco desse estudado se encontra na tentativa de transforma o relacionamento de uma empresa com o governo em algo quantificável e mensurável, para que posteriormente, esse relacionamento seja analisado em busca de

correlações com inidoneidade, tendo como foco a inidoneidade relacionada a troca de favores entre agentes públicos e organizações privadas. Para isso, foi coletado do Sistema Integrado de Administração Financeira do Governo Federal (Siafi) o relacionamento de cada uma dessas empresas com os órgãos integrantes da administração pública federal, são eles:

1. Ministérios
2. Presidências da República
3. Gabinete da Vice-Presidência

Como limitação do estudo, esse relacionamento foi descrito através de uma variável binária, onde para cada uma das empresas, temos uma sequência de “0” e “1” que determina se ela já recebeu recursos de um determinado ministério ou não, sendo o “1” a ocorrência dessa transação. Ao estudar as correlações, o sistema buscara determinar um coeficiente para cada ministério, que poderá impactar positivamente ou negativamente na probabilidade de uma determinada empresa ser inidônea, tendo em vista o número de empresas inidôneas que negociaram com o determinado órgão. A esperança é que seja possível observar o comportamento de cada ministério no que diz respeito ao ato de corrupção em si, usando os dados históricos para observar aqueles que estão mais relacionados a essa prática, criando assim, um risco intrínseco de cada ente da esfera pública federal.

4.4 Natureza da Despesa Orçamentária

Para entendermos o relacionamento da empresa com a esfera pública, buscamos analisar como a empresa se encaixa dentro da execução orçamentária, dentre as possíveis categorias a empresa pode ser classificada como uma despesa corrente ou como uma despesa de capital, sendo que a despesa de capital pode ser um investimento ou uma inversão financeira. As despesas correntes são aquelas que dão base para a manutenção e o funcionamento dos serviços públicos em geral, ou seja, são despesas que não contribuem diretamente para a formação ou aquisição de bem de capital.

Como despesas correntes, podemos encontrar diversos tipos de serviços como material gráfico, reposição de material de escritório, manutenção de software e prestadores de serviços. Diferentemente das Despesas de Capital que por definição, são aquelas que contribuirão para a produção ou geração de novos bens ou serviços e integrarão o patrimônio público, ou seja, são aqueles gastos que geraram um bem de capital, seja ele um ativo imobilizado como um novo imóvel ou a melhoria de um serviço como a aquisição de um software.

Para isso as empresas serão classificadas de acordo com a natureza orçamentária que elas representam para o estado, atribuindo um coeficiente para cada uma dessas classificações de acordo com o seu relacionamento com as empresas inidôneas, onde uma empresa poderá ser classificada como um gasto corrente ou um investimento ou até mesmo ambas.

4.5 Função

Essa categoria busca determinar para qual esfera do governo aquela empresa está prestando o seu serviço, levando em consideração as funções estratégicas do estado, determinando algumas categorias mais abrangentes que buscam contemplar as responsabilidades do governo perante a população. A Subsecretária do Tesouro Municipal de São Paulo registou em seu Manual de Contabilidade Aplicada ao Setor Público que “A função pode ser traduzida como o maior nível de agregação das diversas áreas de atuação do setor público. A função se relaciona com a missão institucional do órgão, por exemplo, cultura, educação, saúde, defesa.”

Resumidamente, podemos verificar no banco de dados final a relação de cada empresa com as funções estratégicas do governo, e com isso podemos determinar um coeficiente para cada área de atuação da administração federal em relação ao nosso modelo. O objetivo em mente é averiguar se existe alguma correlação entre as mais diversas áreas de atuação de um governo e a prática de corrupção, com uma análise bem estruturada será possível observar quais áreas estão mais sucintas a esse tipo de ocorrência.

4.6 Setor de Atuação

A administração pública possui uma infinidade de responsabilidades perante a sua nação o que leva o próprio governo a demandar diversos serviços que por sua vez são executados pelas empresas do setor privado. Uma das consequências da corrupção é a realização de contratos superfaturados que por sua vez levam a máquina pública a despender mais recursos do que o necessário para a realização de um determinado serviço ou a aquisição de um determinado bem.

Tendo em vista a magnitude do serviço público e número de licitações realizadas em todo o território nacional, se torna um grande desafio para os órgãos de controle verificar se os valores pagos estão condizentes com os serviços entregues. Isso pois, assim como foi descrito por Rohwer, Anja (2009), para identificarmos a quantidade de recurso a mais que foi despendido em um contrato, teríamos que ter uma noção completa a respeito da quantidade entregue, o preço cobrado por quantidade e pôr fim a qualidade do que foi entregue e embora o preço e a quantidade sejam informações públicas, a quantidade só é determinante para produtos homogêneos e a qualidade não pode ser objetivamente quantificada.

Levando em consideração esse raciocínio, podemos levantar a hipótese de que o tipo de serviço prestado pela organização pode ser um fator correlacionado com o risco dessa organização praticar atos ilícitos que causam danos ao erário visto que para cada licitação, temos um risco particular desse contrato ser superfaturado de acordo com o tipo de serviço ou produto requisitado por essa licitação. Para isso, foi coletado para cada uma das empresas presentes no banco de dados, o tipo de serviço prestado por essa organização, dentre as mais variadas categorias como Obras e Instalações e Serviços de consultoria, onde cada categoria receberá um coeficiente indicando a sua relação com as empresas que praticaram atos inidôneos.

4.7 Inidoneidade

A Lei nº 8.666/1993, foi primeiro passo para uma regulação efetiva da relação entre o serviço público e os entes privados visto que ela buscava instituir normas sobre

as licitações e contratos da Administração Pública. Dentro de sua competência, temos a criação de sanções administrativas para empresas, concedendo a administração pública poder legal para punir as empresas que apresentarem um comportamento desalinhado com o que é esperado dessas organizações, tais sanções podem ser observadas no Artigo 87.

> Art. 87. Pela inexecução total ou parcial do contrato a Administração poderá, garantida a prévia defesa, aplicar ao contratado as seguintes sanções:
>
> I - advertência;
>
> II - multa, na forma prevista no instrumento convocatório ou no contrato;
>
> III - suspensão temporária de participação em licitação e impedimento de contratar com a Administração, por prazo não superior a 2 (dois) anos;
>
> IV - declaração de inidoneidade para licitar ou contratar com a Administração Pública enquanto perdurarem os motivos determinantes da punição ou até que seja promovida a reabilitação perante a própria autoridade que aplicou a penalidade, que será concedida sempre que o contratado ressarcir a Administração pelos prejuízos resultantes e após decorrido o prazo da sanção aplicada com base no inciso anterior.

Apesar de ser um passo importante para a formação de um relacionamento saudável e ético entre o poder público e os entes privados, essa regulação era superficial e permitia diversas brechas que foram exploradas ao decorrer dos anos. No trabalho de Menezes (2009) ele aponta que a ausência de ampla e irrestrita publicidade dos atos declaratórios de inidoneidade expedido pelos órgãos públicos de todo o país neutraliza os seus efeitos suspensivos em outras localidades, ou seja que embora as empresas fossem punidas e suspensas de acordo com a Lei nº 8.666/1993, a falta de uma administração pública integrada e de uma divulgação adequada dessas punições, resultava em um punição simbólica visto que em sua grande maioria, essas empresas voltavam a negociar com a esfera pública por meio de outros órgãos. Por consequência, a Controladoria-Geral da União que representa o Órgão Central do Sistema de Controle Interno do Poder Executivo Federal, criou o Cadastro Nacional de Empresas Inidôneas ou Suspensas (CEIS) que tem como objetivo unir em um só banco de dados, todas as informações compiladas das instituições federais a respeito de seus fornecedores responsáveis por irregularidades.

Esse banco de dados foi um avanço importante para a administração pública visto que era uma forma sistêmica de se monitorar as relações entre o poder público e as empresas privadas visto que na maioria nas vezes não existia uma punição concreta para aquelas empresas que falhavam na execução de uma licitação. Porém, como declarado pela própria lei, a sua legislação tinha como foco monitorar as empresas que falhavam na execução do contrato, não existia uma forte base legal para julgar as empresas pela prática de corrupção e pela realização de contratos superfaturados, por consequência, a administração pública viveu um período de impunidade no Brasil, onde as licitações em sua grande maioria visavam enriquecer aqueles que participavam do projeto e não a realização de um projeto que visava a melhoria dos serviços básicos para a população.

A consequência dessa impunidade, levou a população brasileira a viver durante anos sobre o domínio de uma máquina pública infestada de agente corruptos que visavam o seu próprio interesse, como corrupção é um termo amplo e frequentemente usado na mídia internacional, vale ressaltar que nesse artigo nos apegamos a definição proposta por Klitgaard:

> Corrupção é o comportamento que se desvia dos deveres formais de uma função pública devido a interesses privados (pessoais, familiares, de grupo fechado) de natureza pecuniária ou para melhorar o status, ou que viola regras contra o exercício de certos tipos de comportamentos ligados a interesses privados. (1994, p. 40)

Apesar do descontentamento da sociedade como um todo, existiam poucos movimentos para manifestar essa indignação, somente em junho de 2013, onde Campos (2015) defende que a população brasileira iniciou um movimento em todo o território nacional, se alastrando pelas ruas das cidades reivindicando de modo especial uma gestão pública que fosse ética, moral e com integridade perante os seus governantes. Dentro desse cenário, a gestão pública nacional precisava dar uma resposta, tanto para a população que estava enfurecida em meio à vasta corrupção espalhada por todo o território nacional, como para atender os compromissos internacionais, entre eles, aqueles estabelecidos na Convenção das Nações Unidas contra a Corrupção (ONU).

Como resposta, o Senado Brasileiro resolveu dar continuidade na aprovação da Lei nº 12.846/2013, um projeto de lei que foi aprovada pela Câmara dos deputados em maio de 2011 e estava paralisada no Senado, essa lei ficou conhecida como a Lei Anticorrupção. Campos (2015) pontua que os principais pontos dessa nova lei são a responsabilidade de pessoas jurídicas, que até então raramente sofriam sanções pelas suas ações corruptas, a institucionalização do *compliance* seguindo as diretrizes internacionais, os acordos de leniência e o Cadastro Nacional de Empresas Punidas (CNEP) que consistia em um banco de dados de todas as empresas que foram julgadas por atos de corrupção.

Esse trabalho visa a elaboração de um modelo que possa identificar a inidoneidade de uma empresa, tendo como foco a inidoneidade contraída por atos de corrupção, através do seu relacionamento com a máquina pública. Para isso, utilizaremos como insumo o Cadastro Nacional de Empresas Inidôneas ou Suspensas (CEIS) e o Cadastro Nacional de Empresas Punidas (CNEP) para definir a variável dependente do nosso modelo.

O Cadastro Nacional de Empresas Inidôneas ou Suspensas (CEIS) consiste em um banco de dados com o objetivo de armazenar todas as empresas e pessoas físicas que falharam na prestação de serviços a máquina pública e por isso foram impedidos de negociar com outros órgãos. Atualmente o banco de dados conta com cinquenta e dois mil e sessenta e seis registros, desses cinquenta e dois mil, somente vinte seis mil cento e oitenta dois são referentes a pessoa jurídica, ou seja, uma empresa enquanto os outros registros são de pessoas físicas.

Tendo em vista o objetivo do trabalho é detectar aquelas empresas que possuem um caráter imoral perante a sua relação com a esfera pública, foi decidido retirar aquelas empresas que foram punidas pela má execução do contrato ou até mesmo pela sua não entrega. Com isso, temos um total de dezesseis mil e oitocentas e setenta empresas, enquanto o Cadastro Nacional de Empresas Punidas (CNEP) possuí um tamanho bem menor, contendo somente setenta e dois registros, como o seu propósito é registrar aquelas empresas que sofreram alguma sanção contida Lei nº 12.846/2013 (Lei Anticorrupção) ele será acrescentado como um todo ao nosso conjunto de empresas problemáticas.

5 RESULTADOS E DISCUSSÃO

Para a realização do método utilizamos um DESKTOP-VCRHVKS Samsung (8 GB de Memória RAM, 980 GB de Memória Física,AMD Phenom(tm) II X6 1090T Processor, 3200 Mhz, 6 Núcleo, 6 Processadores Lógicos), todos os algoritmos de classificação foram implementados no software livre R, versão 3.4.4, instalado com todos os pacotes mencionados anteriormente. Na Tabela 5 e na Tabela 6 podemos ver os resultados da implementação dos modelos de inteligência artificial e da Regressão Logística.

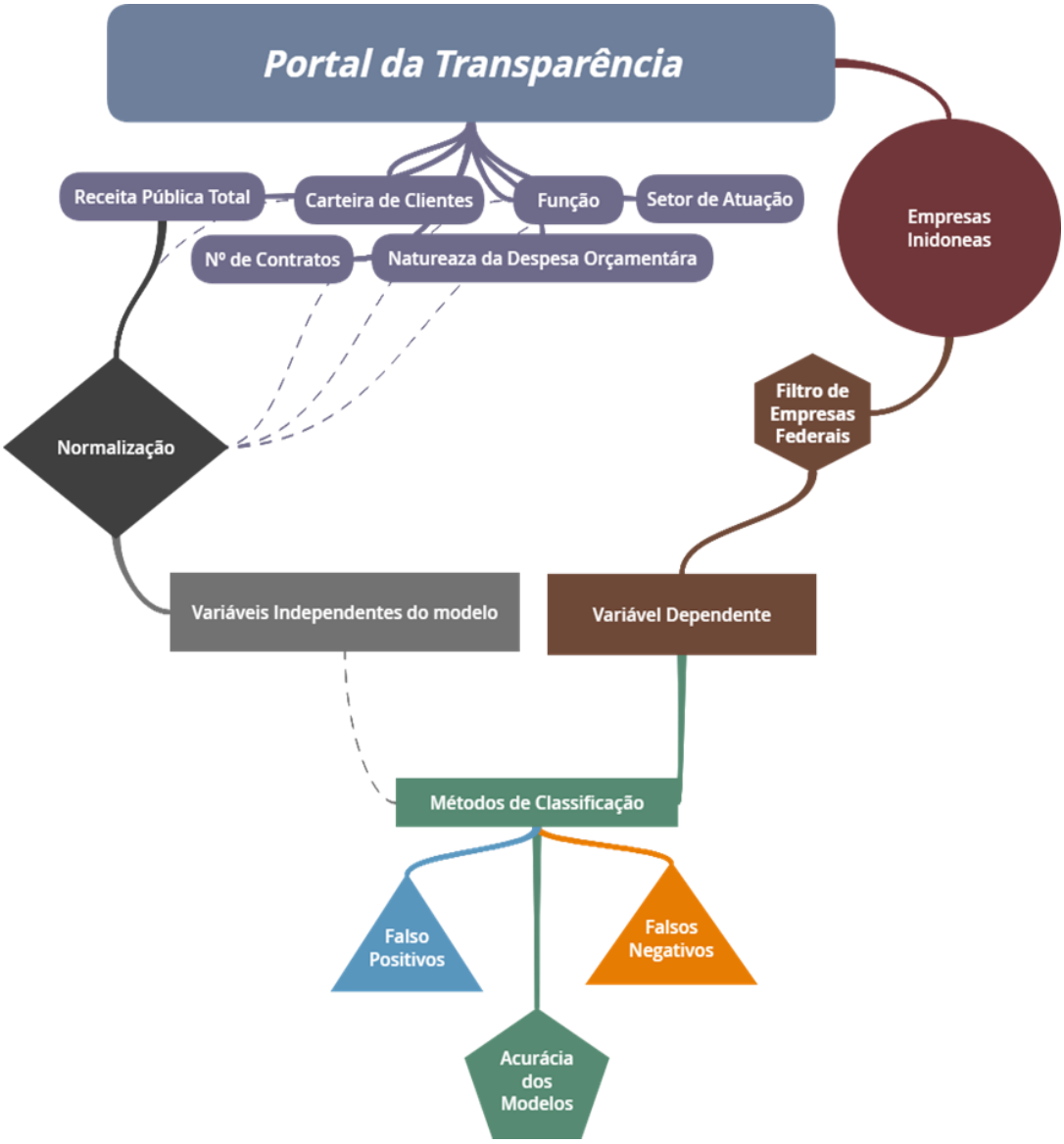

Fig. 1. Ilustração Gráfica da execução do trabalho.

Seis variáveis foram selecionadas para o modelo, todas elas extraídas do Portal da Transparência assim como os bancos de dados das Empresas Inidôneas que passaram por um filtro visto que este trabalho priorizou as transações do governo federal. Todas as variáveis passaram pelo método de normalização para depois serem computadas e trabalhadas pelos modelos de classificação.

Tabela 5 – Resultados do Modelo Enxuto

Modelo Enxuto	Positivo Verdadeiro	Negativo Verdadeiro	Falso Positivo	Falso Negativo	Erro Tipo I (%)	Erro Tipo II (%)	ACC (%)
SVM Radial	285	340	11	63	3,72%	15,63%	89,41%
SVM Linear	286	335	16	62	5,30%	15,62%	88,48%
Regressão Logística	294	332	29	54	8,98%	14,36%	88,13%
Random Forest	252	361	5	82	1,95%	18,51%	87,57%
Bagging	262	351	1	86	0,38%	19,68%	87,57%
ANN	288	318	33	60	10,28%	15,87%	86,70%
Boosting	274	328	41	63	13,02%	16,11%	85,27%

Tabela 6 – Resultados do Modelo Completo

Modelo Completo	Positivo Verdadeiro	Negativo Verdadeiro	Falso Positivo	Falso Negativo	Erro Tipo I (%)	Erro Tipo II (%)	ACC (%)
Random Forest	268	343	6	83	2,19%	19,48%	87,29%
SVM Radial	283	327	22	68	7,21%	17,22%	87,14%
Bagging	267	340	9	84	3,26%	19,81%	86,71%
ANN	298	305	44	53	12,87%	14,80%	86,14%
SVM Linear	290	311	38	61	11,59%	16,40%	85,86%
Boosting	271	303	34	73	11,15%	19,41%	84,29%
Regressão Logística	327	202	147	24	31,01%	10,62%	75,57%

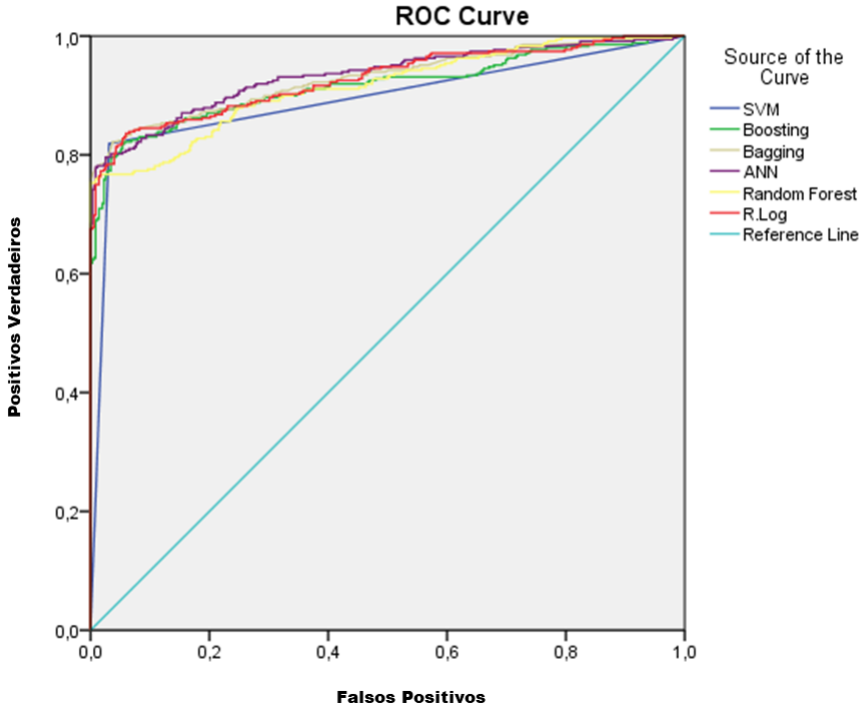

Fig. 2. Curva ROC (Modelo Enxuto).

Para avaliarmos os resultados, construímos a Curva de ROC (*Receiver Operating Characteristic*). Essa curva consiste em uma visualização gráfica da relação de Sensibilidade (Positivos Verdadeiros) e Especificidade (Falsos Positivos), o modelo SVM Radial teve o seu comportamento linear visto que diferentemente dos outros modelos, o seu *output* é binário.

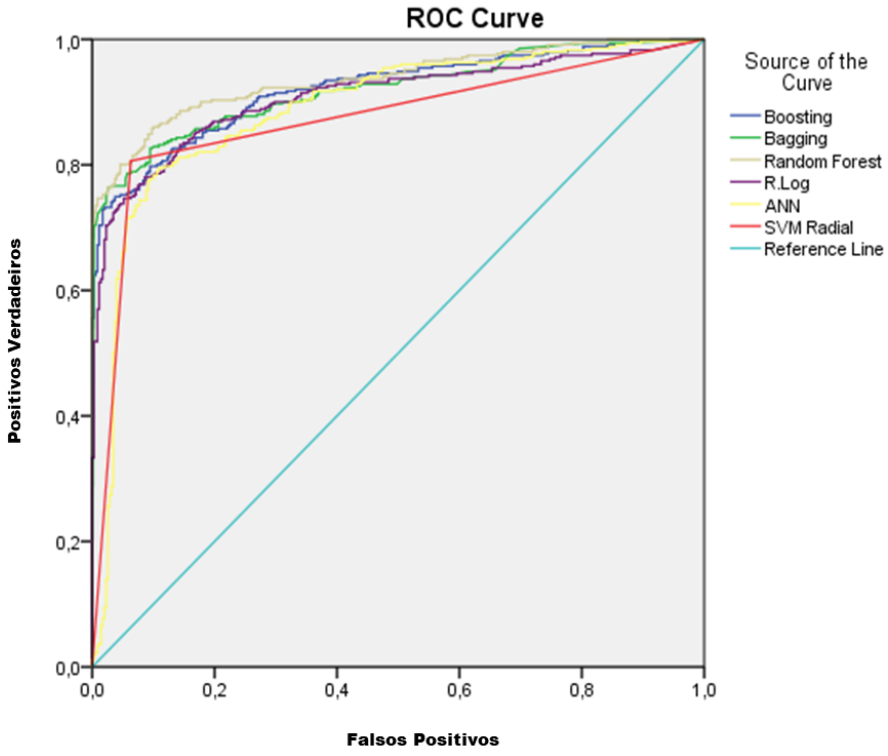

Fig. 3. Curva ROC (Modelo Completo).

Para cada algoritmo de classificação, buscamos a construção de dois modelos, um contendo todas as variáveis que foram levantadas com base nos dados coletados, e outro modelo mais enxuto, que buscou priorizar as variáveis que tinham maior impacto na predição dos resultados. Para a realização desse modelo mais enxuto foram priorizadas as variáveis intervalares, sendo elas a Receita Pública Total e o Nº de Contratos, e as variáveis classificatórias que apresentaram um número menor de possibilidades sendo elas o Natureza da Despesa Orçamentária que era responsável por determinar que tipo de gasto contábil que aquela empresa representava para o governo e a Carteira de clientes que tinha relação com quais ministérios aquela empresa havia transacionado.

Em todos os casos o modelo enxuto apresentou uma acurácia superior ao modelo que continha todas as variáveis, em sua grande maioria não houve uma grande disparidade entre os modelos com a exceção da Regressão Logística que apresentou uma diferença de 12,55% de acurácia entre os modelos, mostrando mais uma vez a dificuldade desse algoritmo de trabalhar com modelos complexos, onde o algoritmo deve analisar um número extenso de variáveis. Porém, apesar da baixa acurácia em comparação com os outros modelos e algoritmos, o modelo completo implementado através da Regressão Logística apresentou a maior quantidade de Positivos Verdadeiros, ou seja, foi o modelo que apresentou a menor ocorrência de erros Tipo II, sendo o responsável por identificar corretamente o maior número de empresas problemáticas. Em termos de acurácia, o algoritmo *Support Vector Machine* apresentou os melhores resultados, sendo que o método radial alcançou um acurácia de 89,41% de acerto enquanto o método linear, surpreendentemente, alcançou uma acurácia de 88,84%, ocupando uma posição superior a outros métodos que utilizam uma abordagem mais complexa. Outro método linear que apresentou um ótimo rendimento foi a Regressão Logística, onde o modelo enxuto alcançou 88,13% de acertos sendo assim, o terceiro modelo com a maior acurácia, seguido de *Random Forest* com 87,57%.

Logicamente que em termos de implementação, a escolha de um algoritmo não deve ser totalmente baseada em sua acurácia, é necessária uma análise profunda a respeito dos erros cometidos por cada algoritmo para que isso se encaixe adequadamente ao contexto em que está inserido. Se priorizarmos aqueles algoritmos que apresentaram um desempenho satisfatório na minimização dos Falso Positivos, ou dos erros Tipo I, inevitavelmente buscaríamos trabalhar com o *Bagging* que demonstrou enorme eficiência na minimização dos falsos positivos, alcançando uma acurácia de quase 100%, logo em seguida temos o *Random Forest* que teve uma acurácia geral idêntica ao *Bagging* com 87,57%, porém resultou em uma taxa de Falsos Positivos de 2,19%, um pouco superior ao *Bagging*.

Esses algoritmos seriam interessantes para um cenário que buscasse minimizar ao máximo os casos onde o modelo classifica uma empresa como problemática de forma equivocada, poderia ser o caso para uma secretária da Controladoria Geral da União que possuí um orçamento limitado e por isso necessita de um sistema capaz de direcionar de forma eficaz a atuação das cúpulas de

investigação. Caso existisse uma certa flexibilidade em relação a ocorrência desses erros, poderíamos ponderar a possibilidade de trabalharmos com o *Support Vector Machine Radial*, visto que além de ser o algoritmo com a maior acurácia geral, ele apresentou uma taxa relativamente baixa de Falsos Positivos, sendo ela de apenas 4%.

Por outro lado, temos aqueles com o destaque positivo no desempenho relativo a minimização de Falsos Negativos, ou erros Tipo II, que são os casos daquelas empresas problemáticas que passariam despercebidas pelo nosso modelo, como dito anteriormente, o modelo completo implementado através da Regressão logística teve o melhor desempenho nesse quesito, porém a sua implementação seria ponderada por conta de sua baixa acurácia geral ao compararmos com os outros modelos e algoritmos. Se o interesse é minimizar a ocorrência de Falsos Negativos, através de um modelo eficiente e equilibrado, devemos considerar o modelo mais enxuto da Regressão Logística visto que ele apresenta uma acurácia geral de 88,13% e uma taxa de Falsos Negativos de 14%. Apesar da superioridade numérica da Regressão Logística, em relação a ocorrência de Falsos Negativos, o modelo completo implementado pelas Redes Neurais Artificiais talvez se apresente como uma alternativa mais atrativa por conta de sua natureza não-linear.

Primeiramente podemos observar que a diferença entre os dois métodos em relação a ocorrência de Falso Negativos e bem pequena visto que as Redes Neurais Artificiais alcançaram uma métrica de 15%, além disso, o algoritmo teve um bom desempenho implementado o modelo completo visto que alcançou um acurácia geral de 86,14%, o que demonstra a capacidade do algoritmo de lidar com modelos complexos que trabalham com um número considerável de variáveis. Esse fator é importante visto que em uma implementação futura, outras variáveis podem ser identificadas e acrescentadas ao modelo, e ao trabalharmos com a Regressão Logística podemos ser limitados pela dificuldade desse método de realizar predições com base em uma longa série de fatores. Além disso podemos observar e comparar o desempenho dos métodos através da visualização gráfica exposta pela curva de ROC, Figura 2 e Figura 3, que tem como objetivo, ilustrar o custo de oportunidade apresentado por essas técnicas de classificação entre a Sensibilidade (Positivos Verdadeiros) e a Especificidade (Falsos Positivos), por isso, aqueles métodos que se localizam próximos do canto esquerdo são aqueles com o desempenho mais atrativos.

Interessante observar a diferença no comportamento da Regressão Logística, visto que no Modelo Enxuto apresentou um desempenho semelhante aos métodos de Inteligência Artificial, algo que se diferencia bastante dos resultados apurados por Barboza (2017), porém ao trabalharmos com o um modelo robusto podemos identificar uma distância mais significativa dos outros modelos mais refinados como *Support Vector Machine* e *Random Forest*, que apresentaram bons desempenho em ambos os modelos.

Por isso devemos sempre levar em consideração os mais diversos fatores na escolha de um método, porém, podemos afirmar que dentre os métodos analisados, aqueles que tiveram mais destaque foram *Support Vector Machine*, *Random Forest*, *Bagging* e Redes Neurais Artificiais, um resultado semelhante a outros trabalhos mencionados nesse artigo, na pesquisa realizada por Whiting (2012) na detecção de fraudes financeiras, observamos uma vantagem significativa de Random Forest em comparação aos métodos tradicionais estatísticos e podemos observar também um destaque dos algoritmos *Support Vector Machine*, *Random Forest* e *Bagging* no trabalho de Barboza (2017) em seu modelo de predição de insolvência, talvez esse seja um sintoma de que o problema levantado nesse artigo não esteja tão distante daqueles que estão em destaque dentro da academia internacional.

6 CONCLUSÕES E RECOMENDAÇÕES

A análise de dados e os métodos de inteligência artificial são uma tendência mundial que muitos especialistas alegam que vão mudar completamente a realidade em que vivemos. Atualmente, essa tendência já é uma realidade para as principais empresas de tecnologias que dominam o mercado mundial, podemos ver a sua aplicação nas estratégias de marketing realizadas pela Amazon, nos anúncios direcionados do Facebook e nos carros autônomos da Tesla dentre várias outras aplicações e por isso eu acredito que esse tema deve ser aprofundado pela academia brasileira em busca de soluções que possam causar um impacto significativo no cenário nacional.

Dentre os maiores desafios enfrentados pela gestão pública se encontra a grandiosidade do território brasileiro o que dificulta a realização de um monitoramento aprofundado das finanças públicas, de uma avaliação significativa dos resultados obtidos e de uma fiscalização eficiente de todas as licitações realizadas por todo o nosso vasto território. Precisamos investir em tecnologias que possibilitem uma gestão eficaz dos interesses públicos através integração nacional e digital de toda informação que possa impactar no atingimento dos interesses estratégicos da nação e principalmente em meios que possam proteger o patrimônio nacional de agentes públicos e privados que visam o próprio enriquecimento por meio de danos ao erário. Esse trabalho buscou verificar a aplicação desses novos conhecimentos a realidade brasileira, principalmente no que se diz respeito a proteção do patrimônio público com a esperança de que esse seja somente um ponto de partida para a realização de novas pesquisas dentro desse tema e que com isso a população brasileira possa colher os frutos de uma academia que visa o desenvolvimento nacional em prol do crescimento da economia e a diminuição da enorme disparidade social que se espalha perante todo o nosso território nacional.

Para isso, enfrentamos o desafio de buscar modelar, por meio de dados reais e acessíveis a toda população, a relação entre as empresas brasileiras e o poder público visto que somente por meio de dados objetivos que podemos aplicar os algoritmos de inteligência artificial com o objetivo de diferenciar uma empresa que presta serviços legais a população de uma empresa que está buscando enriquecer os

seus acionistas através de contratos superfaturados e serviços de baixa qualidade. Vale ressaltar que tal desafio é reconhecido por toda academia internacional devido ao caráter subjetivo das relações corruptas entre os agentes públicos e privados o que resulta numa dificuldade de se apurar, identificar e quantificar a corrupção como algo palpável e passível de analises, porém, esse trabalho, com as suas devidas limitações, buscou identificar alternativas que possam auxiliar no combate a corrupção, alternativas que possam ser implementadas pelos órgãos de controle como ferramentas que visam restringir o universo de controle, otimizando o trabalho dos auditores públicos devido ao direcionamento da atuação desses profissionais aos casos mais sensíveis da administração pública.

Esse projeto também poderia ser aplicado nas instituições policiais com o caráter consultivo, onde ao conduzir uma investigação em determinado órgão da administração pública, a equipe alocada poderia consultar o risco apresentado por cada uma das empresas que negociaram com o determinado órgão público nos últimos anos. Esse risco se apresentaria como a probabilidade de uma organização se tornar inidônea, levando em consideração os contratos celebrados com a administração pública, ao consultar essa informação os agentes buscariam priorizar as empresas que apresentaram um risco elevado, evitando o desperdício de tempo em casos inofensivos a administração pública. Esse tipo de direcionamento aumentaria a efetividade dessas investigações, além de diminuir o tempo e os recursos alocados para o tal.

Por isso vale ressaltar que o modelo trabalhado por esse projeto possuí as suas devidas limitações e certamente ele poderia ser aperfeiçoado com o intuito de potencializar os resultados aqui obtidos. Esse aperfeiçoamento poderia ser adquirido através da inserção de novas variáveis como o capital social dessas empresas, o número de funcionários e a data de abertura, tais variáveis facilitariam o trabalho dos nossos métodos de classificação, aumentando assim, a capacidade desses algoritmos de distinguir uma empresa problemática de uma empresa regular. Outro ponto que deve ser explorado é a realização de métodos adaptados à realidade desse problema, visto que para esse cenário temos uma amostra extremamente desbalanceada de empresas regulares e empresas problemáticas o que se apresenta como um grande desafio para os métodos tradicionais implementados nesse trabalho.

Mas apesar das limitações que foram expostas acimas o trabalho foi desenvolvido adequadamente por meio de uma amostra balanceada e por meio desse trabalho foi possível a obtenção de resultados satisfatórios.

Como apresentado no capítulo anterior, foi possível alcançarmos uma acurácia geral de 89,41% através do método de classificação *Support Vector Machine*, mesmo com as limitações do modelo, o que destaca ainda mais a capacidade dessas novas tecnologias e a oportunidade que isso representa para os órgãos de controle como uma poderosa ferramenta de combate a corrupção.

Porém, assim como havíamos destacado, não devemos nos atentar somente para a acurácia geral do modelo, e por isso devemos enfatizar resultado extraordinário desempenhando pelo método *Bagging* onde podemos observar uma quantidade quase que nula de falsos positivos. Tive a oportunidade de participar de palestras durante a execução desse trabalho que destacavam as dificuldades das secretárias que trabalham com ciência de dados na Controladoria Geral da União e nessa ocasião pude presenciar a importância de minimizar os Falsos Positivos visto que uma investigação equivocada é algo extremamente oneroso para a administração pública, devido ao tempo gastos dos profissionais alocados nessa investigação e os recursos necessários para o tal. Por isso é extremamente motivador saber que a aplicação desses métodos pode fornecer a um futuro auditor uma lista de empresas problemáticas, com a segurança de que o tempo aplicado nessas investigações não será algo em vão.

Porém, com os avanços dessa tecnologia e de sua implementação na gestão pública, acredito que inevitavelmente as prioridades vão se inverter e que com isso, os profissionais de analise de dados alocados nos órgãos de controle buscaram priorizar aquelas técnicas que minimizam a quantidade de Falsos Negativos para que seja possível coletar todo o espaço amostral de empresas problemáticas para o governo, mesmo que dentro dessa amostra tenhamos uma certa quantidade de Falsos Positivos. Com esse objetivo em mente, recomendo a utilização dos métodos de Redes Neurais Artificias, devido a sua natureza não linear visto que isso permite ao modelo a possibilidade de trabalhar com um número considerável de variáveis, podemos observar a ocorrência dessa na natureza nos resultados do trabalho em virtude dos números alcançados pelo método ao se utilizar todas as variáveis levantadas pelo trabalho, como foi observado no capítulo anterior, foi possível

alcançar uma acurácia geral de 86,14% e somente 14,80% de Falsos Negativos, sendo assim o segundo melhor desempenho na minimização de Falsos Negativos, o melhor desempenho se formos considerar somente aqueles resultados que alcançaram uma métrica satisfatória.

Porém, com estes resultados, podemos concluir que é possível investigarmos objetivamente a corrupção e que, apesar do seu caráter subjetivo, podemos analisar as finanças públicas em busca de rastros e indícios de corrupção e acredito que se dedicarmos o mesmo empenho que foi investido nos últimos anos para consolidar a segurança financeira dos bancos, poderemos sonhar com a construção de sistemas capazes de identificar em tempo real atos que represente uma ameaça para o patrimônio público. Além disso, devemos ressaltar que os prejuízos que estas práticas causam ao patrimônio público é somente um dos fatores que prejudicam a economia do país visto que uma gestão pública de alto risco reduz os investimentos das empresas nacionais e os investimentos internacionais a nossa economia, pelo aumento significativo do risco atrelado ao país por conta dessas práticas, por isso o investimento nessas tecnologias e uma implementação integra e transparente destes trabalhos, deve recuperar a confiança local da população e das empresas nacionais, além de ser uma sinalização positiva para o mercado externo, que por consequência, traria mais investimentos.

Por fim, recomendo a futuros trabalhos a realização de métodos regressivos de inteligência artificial para que seja possível realizar um diagnóstico nas estruturas de nossa gestão pública. Um trabalho com esse caráter poderia alertar os gestores da administração federal quais ministérios estão mais propensos a realizar transações dessa categoria, quais funções estratégicas do governo devem ser observadas com maior atenção ou até mesmo qual tipo de serviço costuma a ser fraudado com facilidade, um trabalho nesse formato poderia trazer aos futuros auditores da gestão pública uma discernimento sobre os riscos presentes em cada estrutura de nossa máquina pública.

REFERÊNCIAS

A USERS' GUIDE TO MEASURING CORRUPTION. Norway: Undp Oslo Governance Centre, set. 2008.

BARBOZA, Flavio; KIMURA, Herbert; ALTMAN, Edward. **Machine learning models and bankruptcy prediction**. v. 83, p.405-417, out. 2017.

BEKIREVA, A. S. et al. **Payment Card Fraud Detection Using Neural Network Committee and Clustering**. v. 24, n. 4, p. 193-200. fev. 2015.

BRASIL. Constituição (1993). Lei nº 8666, de 21 de junho de 1993. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. **Lei de Licitações**. Planalto,

BRASIL. Constituição (2013). Lei nº 12846, de 1 de agosto de 2013. Dispõe sobre a responsabilização administrativa e civil de pessoas jurídicas pela prática de atos contra a administração pública, nacional ou estrangeira, e dá outras providências.. **Lei da Empresa Limpa**. Planalto,

BRASIL. TESOURO NACIONAL. **020332 - CLASSIFICAÇÕES ORÇAMENTÁRIAS.** 2018. Disponível em: <http://manualsiafi.tesouro.fazenda.gov.br/020000/020300/020332>. Acesso em: 20 nov. 2018.

BREIMAN, Leo. Bagging Predictors. **Kluwer Academic Publishers**. Boston, v. 24, p. 123-140. nov. 1996.

BREIMAN, Leo. Random Forests. **Kluwer Academic Publishers**. Berkeley, v. 45, p. 5-32. abr. 2001.

CAMPOS, Francisco de Assis Oliveira; PEREIRA, Ricardo A. de Castro. Corrupção e ineficiência no Brasil: Uma análise de equilíbrio geral. **Estudos Econômicos.** Fortaleza, p. 373-408. jul. 2016.

CAMPOS, Patrícia Toledo de. Comentários à Lei nº 12.846/2013 – Lei anticorrupção. **Revista Digital de Direito Administrativo.** São Paulo, p. 160-185. jul. 2014.

DIETTERICH, Thomas G. **An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization**. v. 40, n. 2, p.139-157, ago. 2000.

FAZEKAS, Mihály; TÓTH, István János; KING, Lawrence Peter. Anatomy of grand corruption: A composite corruption risk index based on objective data. **The Corruption Research Center Budapest**. Budapest, v. 07, n. 3 p. 1-46. nov. 2013.

FREUND, Yoav; SCHAPIRE, Robert E.**A Short Introduction to Boosting**. New Jersey, v. 14, n. 5, p. 771-780. set. 1999.

JORDAN, M. I.; MITCHE, T. M. **Machine learning:** Trends, perspectives, and prospects. Pittsburgh, v. 349, n. 6245 p. 255-261. jul. 2015.

H.FRIEDMAN, Jerome. Stochastic gradient boosting. **Sciencedirect**. v. 38, n. 4, p.367-378, fev. 2002.

KHADJAVI, Menusch; LANGE, Andreas; NICKLISCH, Andreas. **How transparency may corrupt − experimental evidence from asymmetric public goods games**. v. 142, p. 468-481. ago. 2017.

KLITGAARD, R. **A corrupção sob controle**. Rio de Janeiro: Jorge Zahar Editora, 1994.

LEAL, Rogério Gesta; KAERCHER, Jonathan Augustus Kellermann. OS IMPACTOS DA CORRUPÇÃO FRENTE À VIOLAÇÃO DOS DIREITOS HUMANOS E DE CIDADANIA: UM DEBATE A SER COMPREENDIDO. **Barbarói.** Santa Cruz do Sul, p. 271-288. jun. 2016.

MAGALHÃES, João Marcelo Rego. Aspectos relevantes da lei anticorrupção empresarial brasileira (Lei nº 12.846/2013). **Revista Controle: Doutrinas e Artigos.** Ceará, p. 24-46. nov. 2013.

MAGUIRE, Karen. **MACHINE LEARNING, HUMAN FRAUD**. South Carolina, p. 70-71. mar. 2017.

MEYER, David. **Support Vector Machines:** The Interface to libsvm in package e1071. 2017. Disponível em: <https://r-forge.r-project.org/>. Acesso em: 1 fev. 2017.

MUNIZ, Francisco Arthur de Siqueira. Os Limites das Sanções Administrativas na Lei 8.666/99. **Revista Controle.** Ceará, p. 171-188. jun. 2011.

Prefeitura de São Paulo. **Manual de Contabilidade Aplicada ao Setor Público, Parte I – Procedimentos Contábeis Orçamentários:** Despesa Orçamentária: conceitos, classificação e etapas. 2012. Disponível em: <http://transparencia.prefeitura.sp.gov.br/contas/Documents/Despesas_detalhamento_municipal.pdf>. Acesso em: 20 nov. 2018.

POZZOLO, Andrea dal. **Adaptive Machine Learning for Credit Card Fraud Detection.** 2015. 199 f. Tese (Doutorado) - Curso de Computer Science, Computer Science Department, Université Libre de Bruxelles, Bruxelles, 2015.

ROHWER, Anja. MEASURING CORRUPTION: A COMPARISON BETWEEN THE TRANSPARENCY INTERNATIONAL'S CORRUPTION PERCEPTIONS INDEX AND THE WORLD BANK'S WORLDWIDE GOVERNANCE INDICATORS. **Cesifo Dice Report.** Munich, p. 42-52. mar. 2009.

SHARMA, Tanvi; VERMA, Sahil; KAVITA. **Prediction of Heart Disease Using Cleveland Dataset:** A Machine Learning Approach. Haryana, v. 4, n. 3, p. 17-21. set. 2017.

TANG, Yuchun et al. SVMs Modeling for Highly Imbalanced Classification. v. 30, n. 1, p. 281-288. fev. 2009.

WEDGE, Roy; KANTER, James Max; VEERAMACHANENI, Kalyan. **Solving the "false positives" problem in fraud prediction**. Cambridge, p. 1-14. out. 2017.

WHITING, David G. et al. **MACHINE LEARNING METHODS FOR DETECTING PATTERNS OF MANAGEMENT FRAUD**. Utah, p. 505-527. maio 2012.