
UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO


Leonardo Alves da Silva

# Modelos de Aprendizagem de Máquina para a predição de valores de Impostos Estaduais

MOSSORÓ - RN

2023

**Leonardo Alves da Silva**

# Modelos de Aprendizagem de Máquina para a predição de valores de Impostos Estaduais

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, como requisito parcial para Defesa de Dissertação.

APROVADA EM:____ / ____ / _____.

BANCA EXAMINADORA

---

Prof. Dr. Sebastião Emidio Alves Filho
**(Orientador/UERN)**

---

Prof. Dr. Isaac de Lima Oliveira Filho
**(Examinador Interno/UERN)**

---

Prof. Dr. Lenardo Chaves e Silva
**(Examinador Interno/UFERSA)**

---

Prof. Dr. Aquiles Medeiros F. Burlamaqui
**(Examinador Externo/UFRN)**

# Resumo

Ao longo dos anos houve uma crescente alta na ocorrência de casos de sonegação fiscal, onde pessoas e empresas tentaram burlar a lei para não pagar ou pagar menos impostos do que é devido. Os impostos estaduais são as maiores fontes de arrecadação dos estados. Investir na sua fiscalização pode trazer grandes benefícios para a sociedade. Foi feita uma Revisão Sistemática da Literatura (RSL) e a partir dela foi observado que a utilização de modelos preditivos baseados em algoritmos de aprendizado de máquina se caracteriza como uma forma de auxiliar na identificação de empresas sonegadoras de impostos. Desta forma, o objetivo com este trabalho é desenvolver uma solução baseada em modelos de aprendizagem de máquina com foco na predição de valores de impostos estaduais, ou seja, para fazer a predição de quanto as empresas deveriam pagar de impostos estaduais para com base nisso melhorar a fiscalização. Neste estudo foram desenvolvidos dois modelos, direcionados na predição dos valores sonegados, através de bases de dados com e sem a identificação das empresas. Para isso, foram avaliados três algoritmos: *Extra Trees*, *Gradient Boosting* e *Random Forest*. Para o treinamento dos modelos foram utilizados alguns dados, entre eles: receitas com vendas de bens ou serviços, custo dos bens ou serviços vendidos, despesas e receitas operacionais, resultado financeiro e a distribuição do valor adicionado. Ao avaliar o desempenho desses algoritmos, foi observado através das métricas de desempenho que os modelos desenvolvidos a partir do algoritmo *Extra Trees* obtiveram os melhores resultados, alcançando um valor máximo no coeficiente de determinação $R^2$ que conseguiu explicar 86,8% da variância dos dados. Por fim, visando demonstrar como os modelos propostos poderiam ser utilizados na prática pelos agentes fiscais de tributos, uma aplicação *Web* foi desenvolvida. Com a realização deste trabalho espera-se contribuir com um aperfeiçoamento na forma de identificar empresas fraudulentas, descobrindo os bons candidatos à fiscalização para que assim possa tornar a tarefa dos auditores mais efetiva. Auxiliar o governo e contribuir para um processo de tomada de decisão, visando reduzir os casos de sonegação fiscal e a arrecadação dos impostos devidos, para que assim se possa ter mais serviços públicos de boa qualidade.

**Palavras-chave**: Aprendizagem de Máquina, Fiscalização, Impostos, Modelos Preditivos, Sonegação.

# Abstract

Over the years there has been a growing increase in the occurrence of cases of tax evasion, where people and companies try to circumvent the law to not pay or pay less taxes than which is due. State taxes are the largest sources of revenue for states. Investing in its supervision can bring great benefits to society. We carried out a Systematic Literature Review (SLR) that observed that the use of predictive models based on machine learning algorithms is characterized as a way to help identify tax evaders. So that, the objective of this work is to develop a solution based on machine learning with a focus on predicting state tax amounts, i.e. to predict how much companies should pay in state taxes to improve supervision. In this study, two models were developed, directed at predicting amounts withheld, through databases with and without the identification of companies. For this, three algorithms were evaluated: Extra Trees, Gradient Boosting and Random Forest. For training the models, we used some data, including: revenue from sales of goods or services, cost of goods or services sold, operating expenses and revenues, financial result and the distribution of the value added. When evaluating the performance of these algorithms, it was observed through the metrics of performance that the models developed from the Extra Trees algorithm obtained the best results, reaching a maximum value in the determination coefficient $R^2$ that managed to explain 86,8% of the data variance. Finally, in order to demonstrate how the proposed models could be used in practice by tax agents, a web application was developed. With the accomplishment of this work it is hoped to contribute with an improvement in the way of identifying fraudulent companies, discovering the good candidates for inspection so that it can make the auditors' task more effective. Assist the government and contribute to a decision-making process, aiming to reduce cases of tax evasion and the collection of taxes due, so that more good quality public services can be provided.

**Keywords**: Machine Learning, oversight, Taxes, Predictive Models, Tax Evasion.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| AM | Aprendizagem de Máquina |
| API | *Application Programming Interface* |
| CE | Critério de Exclusão |
| CI | Critério de Inclusão |
| CMC | Cadastro Municipal de Contribuintes |
| DE | Declaração de Exportação |
| DI | Declaração de Importação |
| DES | Declaração Eletrônica de Serviços |
| ET | *Extra Trees* |
| FGTS | Fundo de Garantia do Tempo de Serviço |
| GBR | *Gradient Boosting* |
| ICMS | Imposto sobre Circulação de Mercadorias e Serviços |
| IPVA | Imposto sobre a Propriedade de Veículos Automotores |
| ITCMD | Imposto sobre Transmissão Causa Mortis e Doação |
| IVC | Imposto de Vendas e Consignações |
| MAE | Erro Absoluto Médio |
| MAPE | Erro Percentual Absoluto Médio |
| MSE | Erro Quadrático Médio |
| QE | Questão Específica |
| QP | Questão de Pesquisa |
| RF | *Random Forest* |
| RFB | Receita Federal do Brasil |
| RL | Regressão Linear |

| | |
|---|---|
| RMSE | Raiz do Erro Quadrático Médio |
| RMSLE | Raiz do Erro Médio Quadrático e Logarítmico |
| RSL | Revisão Sistemática da Literatura |
| SMF | Secretaria Municipal da Fazenda |
| SVM | *Support Vector Machine* |
| WEB | *World Wide Web* |

# Sumário

# 1 Introdução

Para Moraes (1987) quando as pessoas de uma determinada região ou território se reúnem de forma jurídica, com um governo próprio, objetivando cumprir um determinado fim, é formado um estado que, com o intuito de atender as necessidades do grupo, deverá ser provido de soberania.

Essa soberania assegura ao estado alguns poderes, entre eles o poder fiscal que permite criar contribuições obrigatórias, entendidas como impostos, que servem para custear os serviços do estado. Dessa forma, pode-se dizer, que imposto é o preço que a população paga para que se tenha serviços públicos à disposição (CARVALHO, 1985).

A sonegação fiscal afeta a qualidade de vida da população e coloca em risco os projetos do estado. O combate à fraude tributária é de fundamental importância para o desenvolvimento do estado, pois permite ao governo disponibilizar mais serviços públicos e garantir a qualidade de vida do cidadão brasileiro (REGINA CLEIDE, 2005).

O tema de impostos estaduais foi escolhido neste estudo, devido a sonegação desse imposto impactar diretamente na vida das pessoas. Quando o contribuinte sonega impostos, está prejudicando a população em geral, ou seja, o atendimento de suas necessidades básicas, cuja assistência seria proporcionada com o dinheiro que foi sonegado. Portanto é interessante que sejam feitos estudos com o objetivo de desenvolver mecanismos ou ferramentas de auxílio para o governo, visando a redução e o combate à sonegação dos impostos estaduais.

De acordo com Wisaeng (2013) muitas indústrias estão investindo na análise de dados. Hoje em dia, uma grande quantidade de dados é coletada de várias fontes, para que possam ser analisados e utilizados para ajudar na tomada de decisões.

Espera-se que esses modelos possam ser utilizados na prática pelos agentes fiscais de tributos, para que possam auxiliá-los, tornando suas tarefas fiscais mais efetiva e possibilitando na descoberta dos bons candidatos à fiscalização. Que sirvam de alerta para que os contribuintes possam cumprir com as suas obrigações fiscais. Para que assim se possa garantir com a arrecadação dos impostos devidos, diminuindo os casos de sonegação e aumentando a quantidade e qualidade dos serviços públicos.

## 1.1 Objetivos

O objetivo geral deste trabalho é **desenvolver uma solução baseada em modelos de aprendizagem de máquina com foco na predição de valores de impostos estaduais que deveriam ser pagos pelas empresas**. Para alcançar o

objetivo geral, um conjunto de objetivos específicos devem ser contemplados, incluindo:

- investigar a literatura acerca do estado da arte para identificar trabalhos relacionados ao desenvolvimento de modelos de predição no âmbito da sonegação de impostos, visando compreender as metodologias, técnicas, ferramentas e resultados alcançados;
- propor uma solução baseada em modelos de aprendizagem de máquina que possibilite a predição de valores de impostos estaduais a partir de informações financeiras das empresas; e,
- disponibilizar os modelos desenvolvidos de modo a auxiliar os agentes fiscais de tributos na descoberta de empresas candidatas à fiscalização.

## 1.2 Justificativa

A fiscalização dos impostos estaduais é de suma importância, pois é um imposto que impacta diretamente na vida das pessoas e no desenvolvimento dos estados. A receita arrecadada por esses impostos retorna para os cidadãos por meio de serviços públicos melhores. Este estudo é importante para que se possa fazer a predição de valores de impostos estaduais devidos por empresas brasileiras e assim tornar a fiscalização cada vez mais eficiente e consequentemente um melhor desenvolvimento do estado.

Uma técnica de análise de dados bastante utilizada na área financeira é a aprendizagem de máquinas. Para Grus (2019) o conceito de aprendizagem de máquinas é usado para que se possa criar modelos que possam aprender a partir de dados analisados. Segundo Caetano (2015) na Aprendizagem de Máquina (AM), tem-se o aprendizado supervisionado, que é dividido em problemas de regressão, com variáveis-alvo quantitativas, e problemas de classificação, com variáveis-alvo qualitativas. Portanto neste trabalho será abordado a regressão, para fazer a predição de valores de impostos estaduais sonegados pelas empresas.

## 1.3 Metodologia

A elaboração deste trabalho é realizada com base em uma pesquisa aplicada, com o objetivo de gerar conhecimentos específicos da área de fiscalização dos Impostos Estaduais. A natureza empregada nesta pesquisa é a exploratória, logo, o estudo visa descrever como é realizado o processo de fiscalização dos Impostos Estaduais.

Este trabalho é de cunho experimental e sua metodologia foi dividida em algumas etapas, entre elas:

- (i) Fundamentação Teórica com o embasamento teórico sobre as áreas relacionadas ao escopo deste trabalho;

- (ii) Revisão Sistemática da Literatura com o objetivo de agregar as contribuições de estudos em relação a uma questão de pesquisa específica;
- (iii) Base de Dados e Pré-Processamento com as etapas da concepção do modelo de predição, estudo e transformação da base e pré-processamento dos dados;
- (iv) Análise de Desempenho apresentando os algoritmos que foram selecionados para serem avaliados em relação ao desenvolvimento dos modelos de predição;
- (v) Disponibilização dos Modelos de Predição na plataforma *Web*.

## 1.4 Organização do Documento

O restante desse trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada a fundamentação teórica. No Capítulo 3 é apresentado a revisão sistemática da literatura, com a temática deste trabalho. No Capítulo 4 é detalhada a base de dados e pré-processamento. No Capítulo 5 é apresentado a análise de desempenho e a elaboração dos modelos proposto. No Capítulo 6 é descrito o processo de disponibilização e acesso aos modelos desenvolvidos e, por fim, no Capitulo 7 são apresentadas as conclusões.

# 2 Fundamentação Teórica

Neste capítulo é apresentado o embasamento teórico sobre as áreas relacionadas ao escopo deste trabalho.

## 2.1 Impostos Estaduais

O imposto é uma obrigação que o contribuinte tem que pagar para o governo para custear determinados serviços. Existem três categorias de impostos no Brasil, o municipal, estadual e federal, porém este trabalho abordará os impostos estaduais.

Segundo Carvalho (1985) os principais impostos estaduais são Imposto de Transmissão Causa Mortis e Doação (ITCMD), Imposto sobre a Propriedade de Veículos Automotores (IPVA) e Impostos sobre Circulação de Mercadorias e Serviços (ICMS), conforme detalhado a seguir:

- **ITCMD:** Imposto de Transmissão Causa Mortis e Doação, é um imposto estadual que incide sobre o recebimento de herança ou de doações de qualquer natureza. Quem deve recolher esse imposto é o donatário no caso de recebimento de doações ou o herdeiro no caso de um processo de um inventário de uma herança recebida;
- **IPVA:** incide sobre a Propriedade de Veículos Automotores, ou seja, carros, motos, caminhões, etc. É um imposto obrigatório em que todo proprietário tem que fazer essa contribuição anual para o seu estado. A receita que é arrecadada pelo IPVA, além de ser usada para obras em estradas e rodovias, também é utilizada em outras áreas do estado, como Saúde, Educação, Segurança, entre outros;
- **ICMS:** é o Imposto sobre a Circulação de Mercadorias e Serviços, ele incide em produtos e serviços como: energia elétrica, alimentos em geral, transportes, telecomunicações, eletrodomésticos, ou seja, em praticamente todos os produtos consumidos pela população. Esse imposto é cobrado no momento da venda de um produto ou na prestação de serviço para empresas ou consumidor final.

## 2.2 Sonegação Fiscal

Para Lima (2002), sonegar significa burlar a lei, deixando de cumprir com as obrigações fiscais. Várias pesquisas estão sendo desenvolvidas com o objetivo de identificar as causas da sonegação fiscal e encontrar formas de combatê-la, pois a sonegação prejudica

o sistema econômico e impossibilita a viabilidade dos projetos do governo (ACHEK *et al.*, 2015).

Devido às variáveis que definem a base tributária não serem observadas com tanta frequência, muitos problemas de sonegação fiscal são causados. Muitas vezes os contribuintes levam vantagem tentando enganar a tributação, com informações fraudulentas, pois nem sempre é capaz de saber a verdadeira responsabilidade tributária de cada indivíduo (SIQUEIRA; RAMOS, 2005).

Santos (1996) estudou um modelo de sonegação em que há somente variáveis que irão determinar se o contribuinte irá sonegar ou não. Geralmente os contribuintes mal intencionados deixam de atender as exigências tributárias e preferem sonegar impostos, com a intenção de obter ganhos. Os economistas constantemente estão estudando características de possíveis fraudes com o objetivo de combater com a sonegação fiscal.

## 2.3 Ciência de Dados

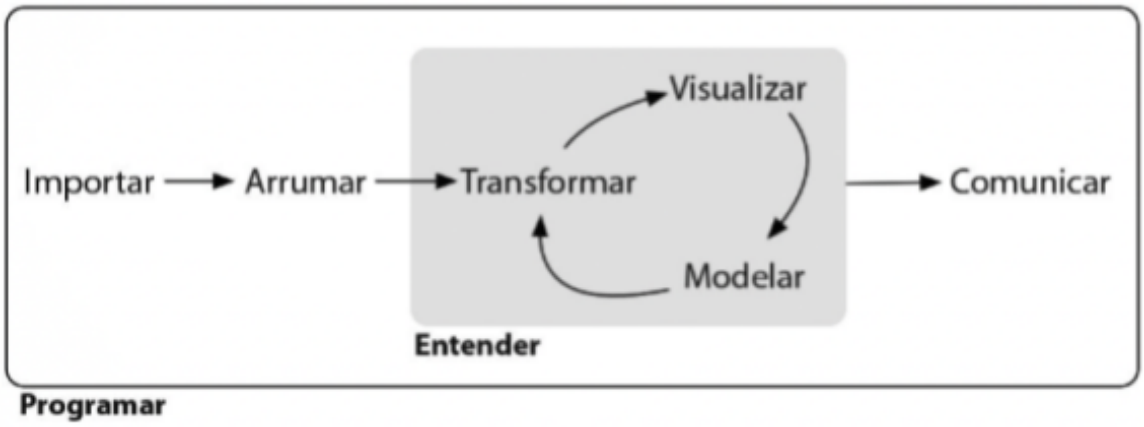
A Ciência de Dados é uma área responsável por transformar os dados em informação, ou seja, obter resultados a partir de dados brutos. Ela usa ferramentas, técnicas e algoritmos para descobrir padrões e relacionamento entre as variáveis (GROLEMUND; WICKHAM, 2017).

Na Figura 1 é apresentado o fluxo de etapas da Ciência de Dados, começando pela importação que é a etapa em que os dados são obtidos. Logo após a importação, os dados são arrumados e organizados de tal forma em que seja possível entender quais atributos estão disponíveis e em que formato foram armazenados.

Posteriormente, vem a etapa de transformação. Nessa etapa a base de dados é avaliada com o objetivo de extrair informações e remover inconsistências que possam causar impacto nas etapas seguintes. Após a etapa de transformação vem a de visualização para fazer a apresentação gráfica dos dados, de maneira que seja possível compreender de forma visual o perfil dos dados.

Em seguida, será possível usar o aprendizado de máquina na construção de um modelo, com o objetivo de identificar padrões e responder questões sobre os dados. Por fim, vem a etapa de comunicação que é direcionada na exposição dos resultados para a comunidade científica.

Figura 1 – Fluxo de etapas da Ciência de Dados




Fonte: Grolemund e Wickham (2017)

Dessa forma, a Ciência de Dados contém um fluxo de etapas que envolve a análise e visualização dos dados, com o objetivo de gerar informações úteis e facilitar a tomada de decisão (HAN; KAMBER; PEI, 2011). A Ciência de Dados dá possibilidades de construir modelos com base em aprendizagem de máquina para que se possa obter conhecimento através dos dados e extrair padrões para criar soluções (CERRI; CARVALHO, 2017). Na seção a seguir serão apresentados os conceitos inerentes à aprendizagem de máquina.

## 2.4 Aprendizagem de Máquina

A Aprendizagem de Máquina (AM) é uma área que surgiu com o objetivo de resolver problemas computacionais complexos e trabalhar com grandes quantidades de dados, tarefas difíceis de serem realizadas por seres humanos. O propósito da AM é criar técnicas computacionais que tornem viáveis ao computador tomar decisões com base em experiências aprendidas anteriormente, ou seja, tratar sobre a implementação de algoritmos que sejam capazes de obter conhecimento de forma automática, melhorando seu desempenho no auxílio a tomada de decisões através de outros problemas já solucionados (MONARD, 2003).

Segundo Carvalho *et al.* (2011) a aprendizagem de máquinas é categorizada em aprendizado supervisionado, aprendizado não supervisionado e aprendizado semi-supervisionado, conforme detalhado a seguir:

- **Aprendizado Supervisionado:** tem um supervisor externo, ou seja, um supervisor que vai conhecer a saída desejada para cada exemplo. O aprendizado supervisionado pode ser dividido em tarefas de regressão e classificação. Nas tarefas de classificação tem-se um número de rótulos discretos, ou seja, um número definido de elementos. Por exemplo, em um problema de diagnóstico médico os dados são representados

por um conjunto de pares *(x, y)* onde o $x$ poderia ser os sintomas do paciente e o $y$ o diagnóstico médico baseado nos sintomas para saber se o paciente estar doente ou não. Já na tarefa de regressão o conjunto de rótulos são contínuos, como peso, altura, valores dos impostos devidos, etc. Como exemplo de problema de regressão tem-se a identificação dos valores de impostos estaduais com base nas vendas de mercadorias e/ou serviços prestados;

- **Aprendizado Não Supervisionado:** diferentemente do aprendizado supervisionado, no não supervisionado os algoritmos não fazem uso do atributo de saída, eles vão explorar a regularidade dos dados e analisar o conjunto de dados de treinamento. O aprendizado não supervisionado pode ser subdividido em sumarização, associação e agrupamento. Na sumarização tenta encontrar uma descrição compacta para os dados, na associação encontrar padrões frequentes de associações entre os atributos e no agrupamento agrupar os dados de cordo com sua similaridade;

- **Aprendizado Semi-Supervisionado:** o aprendizado semi-supervisionado tem alguns dados que são rotulados e outros não. Ele combina uma pequena quantidade de dados que são rotulados e uma grande quantidade de dados não rotulados para o treinamento. Pega a pequena quantidade de dados rotulados e faz uma rotulação nos dados não rotulados para que se possa melhorar a precisão do aprendizado.

## 2.5 Aprendizado Supervisionado

Conforme Mitchell (1997) a aprendizagem supervisionada se resume na predição de um valor de saída com base em um conjunto de dados de entrada. Ela é dividida em tarefas de classificação e de regressão. A classificação é caracterizada pela atribuição de categorias predefinidas a dados, e a regressão se resume em prever o valor de uma variável numérica a partir de um conjunto de variáveis de entrada. Pode ser citado como exemplo de classificação: a classificação de um *dataset* de Marketing Bancário com base no histórico de ligações efetuadas pelos Clientes. Já como exemplo de regressão: pode-se ter a predição do valor do imposto dado pela regressão e o valor informado pela empresa. Este trabalho aborda um problema de regressão, portanto será detalhado na subseção a seguir, os algoritmos de regressão utilizados neste trabalho e em seguida as métricas de avaliação para regressão.

### 2.5.1 Algoritmos

Devido a relevância e a importância do aprendizado de máquina para o contexto de inúmeros problemas nas mais diferentes áreas, uma grande variedade de algoritmos pode ser

encontrada. Nesta subseção são descritos alguns algoritmos de aprendizado supervisionado.

### *Random Forest*

A floresta aleatória (do inglês, *Random Forest*) é um dos algoritmos de aprendizado de máquina mais utilizados. Ele busca a melhor característica em um subconjunto aleatório de características, ao invés de procurar pela melhor característica ao fazer a partição dos nós. Assim criando uma maior diversidade e gerando modelos melhores. Esse algoritmo é baseado em árvores de decisão, que geralmente começa com um único nó, que se divide em possíveis resultados. Existem três tipos de nós: nós de probabilidade, nós de decisão e nós de término. O *Random Forest* (RF) é um algoritmo que pode ser utilizado tanto para tarefas de classificação como para tarefas de regressão (HO, 1995). O critério Gini é utilizado na divisão binária entre os nós da árvore, sendo representado por meio da Equação 2.1.

$$Gini(n) = 1 - \sum_{j=1}^{2}(P_j)^2 \tag{2.1}$$

Sendo $Pj$ o número de vezes que a classe $j$ apareceu nas instâncias do conjunto de dados.

### *Extra Trees*

Arvore Extra ou Árvore Extremamente Randomizadas *(Extra Trees)*, trata-se de um conjunto de arvores de decisões. Ela é bem parecida com a floresta aleatória, que gera uma coleção de árvores de decisão. O *Extra Trees* (ET) diferencia-se do *Random Forest* (RF) por não utilizar uma amostra inicial igual para todas as árvores, e a divisão nos cortes para os nós é realizada de forma aleatória, enquanto no RF é realizado a divisão ideal. Assim possibilitando que essas diferenças façam com que as árvores extra possuam uma redução no viés e uma menor variância em relação a floresta aleatória (GEURTS; ERNST; WEHENKEL, 2006).

### *Gradient Boosting*

*Boosting* é uma forma genérica para aperfeiçoar o desempenho de qualquer algoritmo de aprendizado de máquina. Inicialmente, este método foi proposto para resolver problemas de classificação de padrões. Depois, a partir desta estratégia, foram surgindo diversas generalizações, dentre as quais, o algoritmo *Gradient Boosting*, o qual pode ser aplicado não somente em problemas de classificação como também de regressão. O gradiente usa um método numérico para encontrar um mínimo (local) de uma função. Este algoritmo é baseado na técnica de comitês e consiste em um conjunto $t$ de modelos treinados de

forma sequencial, no qual, o modelo $t$ tem como objetivo corrigir os erros do modelo *t-1* (MAYRINK, 2016). O erro do modelo na iteração $t$ é definido conforme a Equação 2.2.

$$L_t = l(y_i, \hat{y}_i + f_t(x_i)) \tag{2.2}$$

Sendo *yi* o valor alvo, *ŷi* a predição obtida pelo modelo $t$ para a instância *xi*, $l$ uma função de erro e $n$ o número total de instâncias.

### 2.5.2 Métricas de Avaliação para Regressão

Para que se possa avaliar os modelos de regressão, são necessários parâmetros que nos permitam reavaliar um modelo com ele mesmo. Existem diversas métricas aceitas pela comunidade de mineração de dados para avaliar os modelos de regressão. Essas métricas devem ser capazes de trabalhar em um conjunto de valores contínuos, tendo em vista a natureza dos dados. Dentre as métricas disponíveis, algumas foram destacadas por (AZANK, 2020): Erro Absoluto Médio, Erro Quadrático Médio, Raiz do Erro Quadrático Médio, R-Quadrado, Raiz do Erro Médio Quadrático e Logarítmico e Erro Percentual Absoluto Médio.

- **Erro Absoluto Médio (MAE)**: o erro absoluto médio calcula a diferença absoluta média entre os resultados observados *(yi)* e os previstos *(ŷi)*, e possibilita a identificação do erro médio em um conjunto de previsões de tamanho *N*. Seu erro não é afetado por valores *outliers*, ou seja, por valores que se diferenciam drasticamente dos demais. Esta métrica apresenta valor mínimo 0, sem valor máximo. Na sua fórmula sempre é adicionado um módulo entre a diferença dos valores, devido haver valores positivos e negativos. Quanto menor for o seu valor, mais preciso será o modelo de regressão. A fórmula para o cálculo do MAE pode ser vista na Equação 2.3.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2.3}$$

- **Erro Quadrático Médio (MSE)**: o erro quadrático médio é uma métrica bastante utilizada para verificar a acurácia de modelos. Ela consiste na média da somatória do erro ao quadrado e calcula a média de diferença entre o valor predito com o real, fazendo a subtração dos resultados observados *(yi)* com os previstos *(ŷi)*. Quanto menor for o seu valor, melhor será o modelo avaliado. A fórmula para o cálculo do MSE pode ser vista na Equação 2.4.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2.4}$$

- **Raiz do Erro Quadrático Médio (RMSE)**: é uma variação do MSE, como uma forma de melhorar a interpretabilidade da métrica, acertando a unidade. Utiliza basicamente o mesmo cálculo de MSE, porém, para lidar com o problema da diferença entre unidades, é aplicada a raiz quadrada. Uma das vantagens de utilizar esta métrica é que, ao aplicar a raiz quadrada, o erro passa a ter a mesma escala do indicador que estar trabalhando. Quanto mais baixo for o seu valor, significa que a performance do modelo foi boa, pois o erro se aproxima de zero. A fórmula para o cálculo do RMSE pode ser vista na Equação 2.5.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2.5}$$

- **R-Quadrado**: é uma métrica que representa a proporção da variância de uma variável dependente que é explicada por uma variável ou variáveis independentes. O $y$ com circunflexo detalha a estimativa do modelo de acordo com as variáveis. O $R^2$ é muito utilizado em modelos de regressão, seus resultados geralmente são dados em porcentagem e quanto maior for, melhor será o modelo. A fórmula para o cálculo do R-Quadrado pode ser vista na Equação 2.6.

$$R^2 = 1 - \frac{\text{Varianca Residual}}{\text{Varianca Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.6}$$

- **Raiz do Erro Médio Quadrático e Logarítmico (RMSLE)**: essa métrica faz um cálculo bem parecido ao do RMSE, ela é utilizada quando quer fazer a mensuração do viés e da variância, mas não quer penalizar erros que ocorram em magnitudes distintas. A aplicação de logarítmos tem o objetivo de evitar a penalização de uma grande diferença entre o valor predito e o real. A fórmula para o cálculo do RMSLE pode ser vista na Equação 2.7.

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - (\log \hat{y}_i + 1))} \tag{2.7}$$

- **Erro Percentual Absoluto Médio (MAPE)**: é uma das métricas mais usadas para reportar a performance do modelo, seu cálculo é bem parecido com o da métrica

MAE, porém tem o acréscimo de uma divisão por /y/, trazendo uma compreensão mais abrangente do resultado. Quanto menor for o seu valor, mais preciso será o modelo avaliado. A fórmula para o cálculo do MAPE pode ser vista na Equação 2.8.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2.8}$$

## 2.6 Considerações Finais

Através da definição dos conceitos abrangidos neste trabalho, foi possível perceber que o processo de Ciência de Dados possibilita o desenvolvimento de modelos baseados em aprendizagem de máquinas e permite a criação de modelos preditivos a partir da utilização de algoritmos em conjunto com o conhecimento adquirido através dos dados.

# 3 Revisão Sistemática da Literatura

De acordo com Dresch, Lacerda e Valle (2015), uma Revisão Sistemática da Literatura (RSL) é caracterizada como um estudo que tem como objetivo agregar as contribuições de estudos em relação a uma questão de pesquisa específica. Neste capítulo é apresentado uma RSL com intuito de identificar na literatura estratégias preditivas voltadas a identificação de valores de impostos estaduais sonegados por empresas.

Esta RSL segue um protocolo planejado e fundamentado, assim possibilitando que o processo possa ser replicado em outros trabalhos. O protocolo utilizado foi baseado no trabalho de Keele *et al.* (2007) e é composto por cinco etapas: i) delimitar a questão de pesquisa e questões específicas; ii) definir as fontes de busca; iii) selecionar os trabalhos relevantes; iv) avaliar os trabalhos, e v) responder à questão de pesquisa.

## 3.1 Planejamento

Nesta seção serão especificadas as questões de pesquisa, processo de busca e os critérios de inclusão e exclusão envolvidos no processo de seleção dos trabalhos.

### 3.1.1 Questões de Pesquisa

A Questão de Pesquisa (QP) desta RSL se caracteriza como: "Quais as abordagens estão sendo utilizadas na predição de valores de impostos estaduais sonegados pelas empresas?". Para responder este questionamento, foram elaboradas cinco Questões Específicas (QE):

- **QE1:** Quais as principais abordagens (técnicas e algoritmos) utilizadas para análise de dados?
- **QE2:** Quais recursos tecnológicos (linguagens e ferramentas) foram utilizados na análise de dados?
- **QE3:** Qual a origem (fonte dos dados) das bases de dados empregadas na análise de dados?
- **QE4:** Quais variáveis foram consideradas mais relevantes na análise de dados?
- **QE5:** Quais as limitações identificadas no trabalho?

### 3.1.2 Processo de Busca

A pesquisa foi direcionada em trabalhos publicados entre 2012 e 2022, em quatro fontes relevantes na área de tecnologia, a citar:

- *ACM Digital Library* <https://dl.acm.org>;
- *IEEE Xplore Digital Library* <https://ieeexplore.ieee.org>;
- *Scielo* <https://search.scielo.org/>;
- Google Acadêmico <https://scholar.google.com.br>.

As fontes foram escolhidas baseando-se na disponibilidade em acessar o conteúdo dos artigos, importância das fontes e indexação de artigos publicados em periódicos, eventos e conferências voltadas para o objetivo desta pesquisa. Para realizar a pesquisa, foram definidos termos atrelados as áreas financeira (Sonegação Fiscal, Imposto Estadual, Impostos), tecnológica (Aprendizagem de Máquina, Regressão, Inteligência Artificial, Detecção, Ciência de Dados) e estatística (Predição, Modelo).

Para realizar o processo foi necessário gerar uma *string* de busca no idioma português e outra em inglês, tendo em vista que as fontes de pesquisa predominantemente indexam artigos científicos escritos em tal idioma. A *string* de busca **I** foi aplicada nas fontes *Scielo* e Google Acadêmico. Na fonte *ACM* e *IEEE Xplore* foi utilizada a *string* **II**. A *string* **I** ficou definida da seguinte forma: ((Sonegação Fiscal OU Imposto Estadual OU Impostos) E (Aprendizagem de Máquina OU Regressão OU Inteligência Artificial OU Detecção OU Ciência de Dados) E (Predição OU Modelo)). Já a *string* **II** foi composta por: ((*Tax Evasion OR State Tax OR Taxes*) *AND* (*Machine Learning OR Regression OR Artificial Intelligence OR Detection OR Data Science*) *AND* (*Prediction OR Model*)).

### 3.1.3 Critérios para inclusão e exclusão

Para realização da seleção dos trabalhos foram definidos um conjunto de Critérios de Inclusão (CI) e Critérios de Exclusão (CE).

**Critérios de Inclusão:**

CI1) Estudos com foco no objetivo da pesquisa;

CI2) Estudos que foram publicados entre os anos 2012 e 2022;

CI3) Estudos publicados no idioma Inglês;

CI4) Estudos publicados no idioma Português;

CI5) Estudos que abordaram pelo menos um modelo estatístico/matemático.

**Critérios de Exclusão:**

CE1) Estudos que não tem relevância com a pesquisa;

CE2) Estudos repetidos em mais de uma fonte de pesquisa;

CE3) Estudos cuja base de dados utilizada não é do tipo estruturada;

CE4) Estudos que não fornecem dados suficientes para responder a nenhuma das questões de pesquisa.

## 3.2 Seleção dos Trabalhos

O processo de escolha e análise dos artigos foi dividido em quatro estágios. O Estágio I destinou-se à pré-seleção dos artigos, caracterizando-se na obtenção de trabalhos nas fontes de pesquisa, dada a *string* de busca. A pré-seleção se concretizou com a leitura do título e das palavras-chave de cada trabalho obtido e com verificação de alguns critérios de inclusão (CI1, CI3 e CI4). Ao final, 90 trabalhos foram selecionados[1].

Na seleção do estágio II foram analisados o título, palavras-chaves e resumos dos trabalhos. Também foram verificados o critério de inclusão (CI1, CI2, CI3 e CI4) e alguns critérios de exclusão (CE1 e CE3). Os trabalhos que atenderam aos critérios de inclusão foram adicionados à Lista de Incluídos (LI1), enquanto os excluídos foram organizados em uma Lista de Excluídos (LE1). Ao final da execução do Estágio II, 28 trabalhos foram selecionados[2].

No Estágio III foi realizada a leitura da introdução, metodologia adotada e conclusão dos trabalhos da LI1. Para este estágio, verificou-se o critério de exclusão (CE4). Com isso, os estudos excluídos foram inseridos na Lista de Excluídos (LE2), enquanto os trabalhos que permaneceram foram adicionados em uma segunda Lista de Incluídos (LI2). Com o término do Estágio III apenas 9 trabalhos permaneceram na análise[3].

No Estágio IV foram analisados os trabalhos da LI2. Neste caso, foi feita uma leitura completa dos trabalhos e uma revisão de todos os critérios de inclusão e exclusão. Ao fim deste estágio foram selecionados 04 trabalhos para responder as QEs[4].

Na Tabela 1 é apresentada a quantidade de artigos obtidos em cada fonte de pesquisa, para cada estágio do processo de análise.

[1] Trabalhos do Estágio I: <https://cutt.ly/kLhaJMI>
[2] Trabalhos do Estágio II: <https://cutt.ly/6Lha59d>
[3] Trabalhos do Estágio III: <https://cutt.ly/WLhsoQm>
[4] Trabalhos do Estágio IV: <https://cutt.ly/TLhshCo>

Tabela 1 – Número de trabalhos por fonte de busca

| **Fonte** | **Estágio I** | **Estágio II** | **Estágio III** | **Estágio IV (%)** |
|---|---|---|---|---|
| *ACM* | 25 | 0 | 0 | 0 |
| *IEEE* | 18 | 0 | 0 | 0 |
| *SCIELO* | 5 | 2 | 0 | 0 |
| GOOGLE ACADEMICO | 42 | 26 | 9 | 4 (9.5%) |
| TOTAL | 90 | 28 | 9 | 4 (4.4%) |

Fonte: Autoria própria.

De acordo com a RSL pode-se perceber que nos últimos anos foram publicados poucos estudos sobre a identificação de sonegadores de impostos, alguns destes estudos serão apresentados a seguir.

Segundo Rocha (2016) foram feitos experimentos de concepção de classificadores de contribuintes sonegadores, levando em consideração o porte da sonegação, se seria um valor alto ou baixo. Ele analisou mais de 30 atributos que teriam potencial para ser discriminantes em relação à sonegação fiscal, chegando à conclusão de que seis dos atributos analisados foram discriminantes para determinar as classes estudadas.

De acordo com PICCIRILLI *et al.* (2013) dados obtidos de um setor de auditoria localizado na cidade de Goiânia foram usados para identificar características de empresas irregulares. Para prever se a empresa seria irregular foi utilizado os atributos socioeconômicos das empresas, através do uso do algoritmo de árvore de decisão obtendo resultado de 92,03% de precisão.

Coelho (2012) explorou o conceito de sonegação fiscal, com o intuito de otimizar o uso de técnicas de mineração de dados para que polibilitasse detectar empresas fraudulentas. Com base em características de empresas autuadas procurou-se encontrar empresas candidatas que praticam a sonegação fiscal. Normalmente as abordagens de detecção de fraudes baseadas em autuações fiscais resultam em uma coleção de dados de treino muito limitado, pois a porcentagem de autuações geralmente é pequena em comparação ao total de empresas.

Sharma e Panigrahi (2013) executaram uma RSL sobre a aplicação de técnicas de mineração de dados para detectar fraudes na área da contabilidade financeira. Esta RSL teve o objetivo de dar um embasamento teórico sobre esta área de pesquisa. Através da revisão sistemática chegou-se à conclusão de que as técnicas de mineração de dados mais utilizadas para solucionar os problemas pertinentes a detecção e classificação de dados fraudulentos são árvores de decisão, algoritmos de regressão linear, modelos logísticos, redes bayesianas e redes neurais artificiais.

Levando em consideração estes trabalhos relacionados, evidencia-se que o tema de detecção de fraude fiscal, de identificar valores de impostos estaduais sonegados por empresas nos mais diferentes tipos de setores, é de bastante interesse. Portanto, a realização

deste trabalho é de fundamental importância, pois contribui para um processo de tomada de decisão, visando reduzir os casos de sonegação fiscal.

Dessa forma, com o término do processo de seleção, foi possível analisar as contribuições dos trabalhos de modo a buscar respostas às Questões Específicas.

## 3.3 Resultados

Os quatro trabalhos selecionados foram lidos por completo com intuito de identificar nos resultados alcançados respostas para as QP. No Apêndice A é apresentada uma síntese das principais informações dos artigos. A seguir são especificadas as respostas obtidas mediante as Questões de Pesquisa.

### 3.3.1 QE1: Quais as principais abordagens (técnicas e algoritmos) utilizadas para análise de dados?

Dentre as principais técnicas e algoritmos identificados nos trabalhos selecionados, tem-se: SVM (ROCHA, 2016), algoritmo J48 (PICCIRILLI *et al.*, 2013), algoritmo Apriori (COELHO, 2012), modelos logísticos, redes neurais, redes bayesianas, *Gradiente Boosting* e Regressão Linear (SHARMA; PANIGRAHI, 2013).

### 3.3.2 QE2. Quais recursos tecnológicos (linguagens e ferramentas) foram utilizados na análise de dados?

Baseando-se nos estudos selecionados, foram utilizadas na análise dos dados e criação dos modelos as seguintes linguagens/ferramentas: o software WEKA[5] (ROCHA, 2016); software Tamanduá (COELHO, 2012) e linguagem *Python*[6] e R[7] (SHARMA; PANIGRAHI, 2013).

### 3.3.3 QE3: Qual a origem (fonte dos dados) das bases de dados empregadas na análise de dados?

Para o treinamento dos modelos desenvolvidos, os autores utilizaram informações advindas de empresas, destacando-se a origem de: I) Empresas prestadoras de serviço em conjunto com a Secretaria Municipal da Fazenda (SMF) do município de Porto Alegre (Procempa) (ROCHA, 2016); II) Dados de empresas do setor atacadista de pequeno, médio e grande porte, situadas no município de Goiânia. No período de 2013 a 2014 foram coletados vários dados dos Sistemas de Auto de Infração e Cadastro de Contribuintes

[5] WEKA <https://www.cs.waikato.ac.nz/ml/weka/>
[6] Python <https://www.python.org/>
[7] R <https://www.r-project.org/>

através dos sistemas de informação mantidos pela SEFAZ-GO (PICCIRILLI *et al.*, 2013); III) Dados extraídos da Declaração Eletrônica de Serviços (DES), declarada por empresas de Belo Horizonte, cadastradas no Cadastro Municipal de Contribuintes (CMC) que tomaram serviços de empresas sediadas fora de Belo Horizonte (COELHO, 2012) e IV) Informações eletrônicas disponíveis na Receita Federal do Brasil (RFB), através de reuniões com especialistas da RFB na área de investigação dos crimes de lavagem de dinheiro no comércio exterior e das áreas de fiscalização aduaneira e de vigilância e repressão aduaneira. Para isso foram utilizadas a seguintes bases de dados: Arrecadação (BArr), Cadastros (BCad), Comércio Exterior (BCE), Contribuições, Empregados (BEmp), Movimentações Financeiras (BMF), Notas Fiscais Eletrônicas (BNFe) e Retenções de Impostos na Fonte (BRIF) (SHARMA; PANIGRAHI, 2013).

### 3.3.4 QE4: Quais variáveis foram consideradas mais relevantes na análise de dados?

As variáveis mais relevantes utilizadas pelos autores para realizar a análise dos dados foram: I) Notas Fiscais (ROCHA, 2016); II) Porte, Natureza Jurídica, Classe da Atividade Econômica, Valores de Crédito, Débito, Saldo Credor e Saldo Apurado Devedor por Período, Deduções, ICMS a Recolher, Valores de Ajuste de Débito e Estorno de Débito e Crédito por ano (PICCIRILLI *et al.*, 2013); III) Nome do Tomador de Serviços, Inscrição Municipal do Tomador, Nome do Prestador de Serviços, CNPJ do Prestador, Cidade do Prestador, Valor Total do Serviço e Valor do ISS Retido (COELHO, 2012) e IV) Declarações diversas prestadas por contribuintes à RFB, Demonstrativo de apuração de contribuições sociais (Dacon), identificação da empresa exportadora, o Tipo de atividade econômica realizada, sua Situação Cadastral Atual e Passada (ativa, inativa ou suspensa), Movimentações Realizadas no Comércio Exterior pelas Empresas Exportadoras, Informações sobre os Valores e Quantitativos Exportados e Importados em cada Declaração de Exportação (DE) e Declaração de Importação (DI), Valores Declarados como Devidos (SHARMA; PANIGRAHI, 2013).

### 3.3.5 QE5: Quais as limitações identificadas no trabalho?

Foi possível identificar que alguns trabalhos relataram problemas nos procedimentos envolvidos na construção dos modelos preditivos, a citar: I) a baixa quantidade de amostras (instâncias) presentes nas bases de dados utilizadas na composição do modelo (PICCIRILLI *et al.*, 2013) e II) escassez de variáveis, mesmo em bases com muitas instâncias (COELHO, 2012).

Na Tabela 2 é apresentada a disposição das características das bases de dados utilizadas, incluindo: I) quantidade de instâncias, II) número de variáveis e III) período de coleta dos dados.

Tabela 2 – Características das Bases de Dados

| ID | Instâncias | Atributos | Período de Coleta dos Dados |
|---|---|---|---|
| 54 | 5.128 | 74 | 01/01/2016 - 01/10/2016 |
| 66 | 143 | 32 | 05/02/2013 - 05/02/2014 |
| 67 | 1.020.890 | 8 | 10/01/2010 - 01/02/2011 |
| 68 | 2.719 | 77 | Não Informado |

Fonte: Autoria própria.

## 3.4 Considerações Finais

Esta RSL auxiliou na identificação do cenário atual de estudos científicos sobre o desenvolvimento de soluções preditivas relacionadas a predição de valores de impostos estaduais sonegados por empresas. Por meio dos resultados apresentados foi constatado que: i) um conjunto de técnicas baseadas na aprendizagem de máquinas vêm sendo utilizadas como abordagem fundamental na criação dos modelos, a citar: SVM, *Gradiente Boosting* e Regressão Linear; ii) as linguagens de programação *Python* e R estão sendo utilizadas com frequência na análise e no desenvolvimento dos modelos preditivos; iii) além das linguagens de programação, alguns softwares e algoritmos específicos estão sendo empregados, como: Ferramenta WEKA e os algoritmos Apriori e J48; iv) mesmo com a disponibilização de bases de dados relacionadas aos casos de sonegação de impostos nos estudos analisados, em sua grande maioria já pré-processadas, estudos vêm utilizando bases de empresas com o objetivo de coletar e avaliar uma maior variedade de informações de uma região geográfica específica. As bases usadas nos trabalhos selecionados tinham informações privadas, não acessíveis facilmente, apenas por alguns orgãos públios.

# 4 Base de Dados e Pré-Processamento

Neste capítulo são apresentadas as etapas necessárias para a concepção do modelo de predição (Seção 4.1). Na sequência é apresentada a base de dados utilizada nesta pesquisa (Seção 4.2). Em seguida são apesentados os procedimentos utilizados no pré-processamento (Seção 4.3). Na sequência é descrito a nova base gerada após a realização das etapas de pré-processamento (Seção 4.4). Por fim, são apresentadas as considerações finais do capítulo (Seção 4.5).

## 4.1 Etapas da Concepção do Modelo de Predição

Para a concepção do modelo de predição foram definidas as seguintes etapas: i) escolha da base de dados; ii) seleção dos atributos; iii) pré-processamento dos dados; iv) definição de novas bases de dados; v) seleção dos algoritmos; vi) treinamento e teste dos modelos; vii) avaliação dos modelos a partir das métricas de desempenho; viii) avaliação dos principais atributos, e; ix) disponibilização dos modelos através uma aplicação *Web*.

Todos os experimentos deste trabalho foram realizados em um ambiente de programação *Python* com uso da ferramenta Google Colab[1]. Dentre as bibliotecas utilizadas no experimento tem-se a *Numpy*[2] e *Pandas*[3] para fazer a preparação e análise dos dados e o *PyCaret*[4] para criar os modelos de aprendizado de máquina.

O *pycaret* é uma biblioteca do *python* baseada no aprendizado de máquina automatizado que faz a comparação de vários algoritmos e destaca a melhor pontuação de cada métrica. Ela pode ser usada em problemas de agrupamento, detecção de anomalias, processamento da linguagem natural, regras de associação, classificação e regressão (ALI, 2020). Ela foi utilizada para fazer a seleção dos algoritmos e a partir dos resultados obtidos foi observado que os algoritmos *Extra Trees*, *Gradient Boosting* e *Random Forest* obtiveram os melhores resultados.

Após isso, foi utilizada a técnica *hold-out* para a realização do treinamento e teste dos três algoritmos selecionados. Para isso, foi utilizada uma proporção de 70/30, sendo 70% das instâncias para treinamento e 30% para teste. Na etapa seguinte, visando avaliar os resultados alcançados pelos algoritmos, foram utilizadas métricas de desempenho da Subseção 2.5.2.

[1] Google Colab <https://colab.research.google.com/>
[2] Numpy <https://numpy.org/>
[3] Pandas <https://pandas.pydata.org/>
[4] PyCaret <https://pycaret.gitbook.io/docs/>

## 4.2 Base de Dados

As bases utilizadas nesse trabalho tratam-se de dados financeiros de empresas brasileiras de médio e grande porte, com setores diversificados, tais como setor Financeiro, Bens Industriais, Comunicações, Petróleo e Gás, Biocombustíveis, Saúde, Tecnologia da Informação e Utilidade Pública. Entre as empresas do setor financeiro tem-se banco Itaú, Santander, Bradesco e Banco do Brasil.

Os dados foram obtidos através de uma *Application Programming Interface* (API) disponível no Github[5], que baixa e processa os arquivos *Comma Separated Values* (CSV) da Comissão de Valores Mobiliários (CVM) no site Portal Dados Abertos e os armazena em um banco de dados local SQLite, onde são extraídos os dados consolidados do balanço patrimonial, fluxo de caixa e demonstração de valor adicionado.

Para esse trabalho foram extraídas duas bases de dados através da API, após isso foi feito a união delas para obter informações que de certa forma seriam relevantes para o problema de pesquisa. Foi extraído a base de nome *companies* contendo 608 instâncias e três atributos, com informações de cada empresa, possuindo os atributos *ID*, *CNPJ* e o *NOME* da empresa.

Foi extraído também a principal base de dados, de nome dfp, que contém dados financeiros de empresas brasileiras do período de 2009 a 2020 e é composta por 1.288.052 instâncias e doze atributos, essa base contém os atributos: *ID*, *ID_CIA*, *CODE*, *YEAR*, *DATA_TYPE*, *VERSAO*, *MOEDA*, *ESCALA_MOEDA*, *DT_FIM_EXERC*, *CD_CONTA*, *DS_CONTA* e *VL_CONTA*, que serão explicados posteriormente.

Entre os tipos de contas dessa base tem-se, por exemplo: Passivo Total, Passivo Circulante, Receita de Venda de Bens e/ou Serviços, Custo dos Bens e/ou Serviços Vendidos, Despesas/Receitas Operacionais, Resultado Financeiro, Receitas, Distribuição do Valor Adicionado, Impostos, Taxas e Contribuições.

Na Figura 2 é apresentada a base de dados original. Conforme pode-se ver, as contas estão divididas em várias linhas através do atributo *DS_CONTA* (descrição da conta).

[5] API <https://github.com/dude333/rapina>

Figura 2 – Visão da Base de Dados Original

| | ID | ID_CIA | CODE | YEAR | DATA_TYPE | VERSAO | MOEDA | ESCALA_MOEDA | DT_FIM_EXERC | CD_CONTA | DS_CONTA | VL_CONTA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3964637342 | 516 | 10 | 2020 | BPA | 2 | REAL | MIL | 1609372800 | 1 | Ativo Total | 1.693794e+09 |
| **1** | 1659546626 | 516 | 2792736795 | 2020 | BPA | 2 | REAL | MIL | 1609372800 | 1.01 | Caixa e Equivalentes de Caixa | 1.761895e+08 |
| **2** | 2861785205 | 516 | 3102149090 | 2020 | BPA | 2 | REAL | MIL | 1609372800 | 1.01.01 | Caixa | 1.678456e+07 |
| **3** | 2096417537 | 516 | 1787761393 | 2020 | BPA | 2 | REAL | MIL | 1609372800 | 1.01.02 | Aplicações de Liquidez | 1.594049e+08 |
| **4** | 316176377 | 516 | 2546186903 | 2020 | BPA | 2 | REAL | MIL | 1609372800 | 1.02 | Ativos Financeiros | 1.392258e+09 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1288047** | 1888438359 | 507 | 1729268281 | 2010 | DVA | 2 | REAL | MIL | 1293753600 | 7.08.04.03 | Lucros Retidos / Prejuízo do Período | 3.122080e+05 |
| **1288048** | 440871288 | 507 | 2451075925 | 2009 | DVA | 2 | REAL | MIL | 1262217600 | 7.08.04.04 | Part. Não Controladores nos Lucros Retidos | 1.618000e+03 |
| **1288049** | 4064021602 | 507 | 2451075925 | 2010 | DVA | 2 | REAL | MIL | 1293753600 | 7.08.04.04 | Part. Não Controladores nos Lucros Retidos | 3.520000e+02 |
| **1288050** | 1396703795 | 507 | 3426904189 | 2009 | DVA | 2 | REAL | MIL | 1262217600 | 7.08.05 | Outros | 0.000000e+00 |
| **1288051** | 2986400361 | 507 | 3426904189 | 2010 | DVA | 2 | REAL | MIL | 1293753600 | 7.08.05 | Outros | 0.000000e+00 |

1288052 rows × 12 columns

Fonte: Autoria própria.

## 4.3 Pré-processamento

O pré-processamento serve para minimizar e eliminar problemas nos dados através de um conjunto de atividades que envolvem preparação, organização e estruturação dos dados. É uma técnica de fundamental importância, pois será determinante para a qualidade final dos dados que serão analisados.

Nesta fase foi aplicado um conjunto de operações sobre os dados, tais como a normalização, feito uma verificação para saber se tinha valores nulos (foi constatado que não, por isso não foi feita nenhuma substituição de valores), valores faltosos, duplicados (a única base que tinha valores duplicados era a dfp, porém foi feito a eliminação desses registros através da biblioteca *pandas*), *outliers*, linhas com Not a Number (NaN), o tipo de dado de cada atributo, dados estatísticos das bases de dados (referente aos atributos numéricos), correlação das variáveis, entre outras operações.

Dentre os atributos disponíveis na base de dados dfp, 5 foram selecionados como sendo mais importantes: *ID_CIA* (armazena o código da empresa), *CD_CONTA* (registra o código identificador de cada conta), *DS_CONTA* (contém a descrição da conta), *YEAR* (mostra o ano em que as informações se referem variando entre 2009 a 2020) e *VL_CONTA* (que armazena os valores de todas as contas em escala moeda de MIL).

É importante destacar que a seleção dos atributos para compor a nova base de dados da Figura 4 foi feita através de uma pesquisa com contabilista especialista na área financeira. Durante a pesquisa foi definido que as variáveis escolhidas para compor a base tinham uma correlação forte com os impostas estaduais, conforme pode-se ver na Figura 3.

## 4.4 Nova Base de Dados Gerada

Após as etapas de pré-processamento foi feito o agrupamento entre as bases de dados *companies* e dfp para que se pudesse cruzar as informações referentes a cada empresa e gerar uma base de dados definitiva. É importante destacar que a base de dados dfp passou por um processo de transformação para que se pudesse chegar aos resultados alcançados, um deles foi o atributo DS_CONTA que armazena a descrição de cada conta, e estava mostrando essas informações em várias linhas, daí foi preciso fazer uma tratativa agrupando essas contas em colunas, assim permitindo gerar uma base de dados em que os algoritmos de regressão pudessem trabalhar de forma eficiente.

Antes de aplicar a regressão foi feito uma análise na nova base gerada, obtendo a correlação de variáveis preditoras em relação a variável alvo. Na Figura 3 a seguir é apresentado a correlação dos atributos que foram selecionados e que tem uma maior influência nos valores dos impostos estaduais. Durante a análise verificou-se que a correlação das receitas com vendas de bens ou serviços vendidos foi de (0.77), custo dos bens ou serviços vendidos (-0.73), despesas e receitas operacionais (-0.71), resultado financeiro (-0.44), receitas (0.85) e a correlação da distribuição do valor adicionado foi de (0.88) em relação aos impostos estaduais.

Figura 3 – Correlação dos atributos

| | IMPOSTOS_ESTADUAIS | RECEITA_VENDAS_BENS_OU_SERVICOS | CUSTO_DOS_BENS_OU_SERVICOS_VENDIDOS | DESPESAS_RECEITAS_OPERACIONAIS | RESULTADO_FINANCEIRO | RECEITAS | DISTRIBUICAO_DO_VALOR_ADICIONADO |
|---|---|---|---|---|---|---|---|
| IMPOSTOS_ESTADUAIS | 1.00 | 0.77 | -0.73 | -0.71 | -0.44 | 0.85 | 0.88 |
| RECEITA_VENDAS_BENS_OU_SERVICOS | 0.77 | 1.00 | -0.98 | -0.83 | -0.60 | 0.97 | 0.88 |
| CUSTO_DOS_BENS_OU_SERVICOS_VENDIDOS | -0.73 | -0.98 | 1.00 | 0.75 | 0.54 | -0.95 | -0.82 |
| DESPESAS_RECEITAS_OPERACIONAIS | -0.71 | -0.83 | 0.75 | 1.00 | 0.56 | -0.82 | -0.75 |
| RESULTADO_FINANCEIRO | -0.44 | -0.60 | 0.54 | 0.56 | 1.00 | -0.58 | -0.58 |
| RECEITAS | 0.85 | 0.97 | -0.95 | -0.82 | -0.58 | 1.00 | 0.94 |
| DISTRIBUICAO_DO_VALOR_ADICIONADO | 0.88 | 0.88 | -0.82 | -0.75 | -0.58 | 0.94 | 1.00 |

Fonte: Autoria própria.

Conforme podemos ver, com exceção do resultado financeiro que apresentou uma correlação um pouco inferior, todos os outros atributos da base apresentaram uma correlação forte acima dos 70% (positiva ou negativa) em relação a variável alvo, pois quanto mais os valores delas sobem, os valores dos impostos estaduais também aumentam.

Para gerar a nova base foram selecionados os atributos *ID* (código identificador da empresa) e *EMPRESA* (nome da empresa) da base de dados *companies*. Já dentre os atributos disponíveis na base de dados dfp, foram selecionadas algumas contas do atributo *DS_CONTA* como sendo mais importantes para o problema de pesquisa, a citar:

- *ANO:* variando entre 2009 a 2020;
- *IMPOSTOS_ESTADUAIS:* contendo a soma de todos os impostos estaduais (ICMS, IPVA e ITCMD);

- *RECEITA_VENDAS_BENS_OU_SERVICOS:* registrando as receitas das vendas dos bens ou serviços prestados;
- *CUSTO_DOS_BENS_OU_SERVICOS_VENDIDOS:* mensurando se as despesas com bens dão lucro nas vendas;
- *DESPESAS_RECEITAS_OPERACIONAIS:* registrando os gastos essenciais para que o negócio da empresa consiga operar;
- *RESULTADO_FINANCEIRO:* armazenando os lucros ou prejuízos existentes entre as atividades não operacionais da empresa;
- *RECEITAS:* registrando os valores que a empresa recebe por seus produtos ou serviços;
- *DISTRIBUICAO_DO_VALOR_ADICIONADO:* mensurando quanta riqueza a empresa produziu em um determinado período de tempo.

Na Figura 4 pode-se ver a base de dados definitiva para análise de dados, contendo 3909 instâncias e 10 atributos. As contas foram divididas em colunas e as informações agrupadas pelo *ID* (código identificador da empresa) e pelo *ANO*. Vale lembrar que na Figura 4 são apresentadas informações de todas as empresas que têm os tipos de contas selecionadas acima respectivamente, contendo os valores das contas anual, do período de 2009 a 2020.

Figura 4 – Base de Dados Transformada

| | ID | EMPRESA | ANO | DISTRIBUICAO_DO_VALOR_ADICIONADO | RECEITAS | RESULTADO_FINANCEIRO | DESPESAS_RECEITAS_OPERACIONAIS | CUSTO_DOS_BENS_OU_SERVICOS_VENDIDOS | IMPOSTOS_ESTADUAIS | RECEITA_VENDAS_BENS_OU_SERVICOS |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 100 | CENTRAIS ELET BRAS S.A. - ELETROBRAS | 2009 | 15440138.0 | 27037574.0 | -3638097.0 | -17883163.0 | -3896669.0 | 0.0 | 25831183.0 |
| 1 | 100 | CENTRAIS ELET BRAS S.A. - ELETROBRAS | 2010 | 15939588.0 | 31015307.0 | -364123.0 | -21137374.0 | -4265905.0 | 0.0 | 29814662.0 |
| 2 | 100 | CENTRAIS ELET BRAS S.A. - ELETROBRAS | 2011 | 17716825.0 | 34248491.0 | 234453.0 | -23719982.0 | -4715747.0 | 0.0 | 33061356.0 |
| 3 | 100 | CENTRAIS ELET BRAS S.A. - ELETROBRAS | 2012 | 6160137.0 | 39583627.0 | 632509.0 | -42013112.0 | -5474384.0 | 0.0 | 39538861.0 |
| 4 | 100 | CENTRAIS ELET BRAS S.A. - ELETROBRAS | 2013 | 8984118.0 | 28188399.0 | 285946.0 | -29026287.0 | -4350755.0 | 0.0 | 28188399.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3904 | 703 | CECRISA REVESTIMENTOS CERAMICOS SA | 2010 | 232124.0 | 602807.0 | -70107.0 | -129794.0 | -313778.0 | 30075.0 | 498891.0 |
| 3905 | 704 | BUDDEMEYER S/A | 2009 | 66663.0 | 144194.0 | 121.0 | -24557.0 | -72745.0 | 9096.0 | 115268.0 |
| 3906 | 704 | BUDDEMEYER S/A | 2010 | 68593.0 | 163485.0 | -1571.0 | -29419.0 | -85063.0 | 9717.0 | 129944.0 |
| 3907 | 706 | YARA BRASIL FERTILIZANTES | 2009 | 254597.0 | 2173519.0 | 125071.0 | -216653.0 | -2116686.0 | 43783.0 | 2096268.0 |
| 3908 | 706 | YARA BRASIL FERTILIZANTES | 2010 | 1496159.0 | 3119650.0 | -43835.0 | 1098464.0 | -1593480.0 | 13008.0 | 1818561.0 |

3909 rows × 10 columns

Fonte: Autoria própria.

## 4.5 Considerações Finais

Através da definição dos conceitos abrangidos neste capítulo, foi possível perceber que o processo de Ciência de Dados é de fundamental importância, pois nos fornece ferramentas e etapas necessárias para o pré-processamento e análise das bases de dados de modo a extrair conhecimento dos dados ao avaliar os diferentes tipos de informações.

No próximo Capítulo é apresentado a análise de desempenho com o detalhamento dos melhores algoritmos selecionados.

# 5 Análise de Desempenho

Neste capítulo são apresentados os algoritmos que foram selecionados para serem avaliados em relação ao desenvolvimento dos modelos de predição (Seção 5.1). Na sequência é apresentada uma análise nos dados das empresas, com intuito de identificar similaridades (Seção 5.2). Em seguida, são apresentados os resultados alcançados na criação do modelo de predição através de uma base de dados sem a identificação das empresas (Seção 5.3). Na sequência são ilustrados os resultados obtidos através do desenvolvimento do modelo de predição através de uma base de dados com a identificação das empresas (Seção 5.4). Por fim, são discutidos os resultados alcançados (Seção 5.5).

## 5.1 Algoritmos Selecionados

Para seleção dos algoritmos foi utilizado a biblioteca *PyCaret*, pois ela fornece um conjunto de algoritmos úteis para problemas de regressão, sendo alguns deles: *Extra Trees*, *Gradient Boosting*, *Random Forest*, *Lasso*, *Lasso Least Angle*, *Ridge*, *Orthogonal Matching Pursuit*, *Decision Tree*, *AdaBoost*, *K Neighbors*, *Elastic Net*, *Bayesian Ridge*, *Linear Regression*, *Huber Regressor*, *Light Gradient Boosting Machine*, *Dummy*, *Passive Aggressive* e *Least Angle*.

Ao utilizar a base de dados gerada na Seção 4.4 foi possível avaliar o desempenho dos algoritmos da biblioteca *PyCaret*. Para fazer a comparação dos modelos foi utilizado a função *compare_models*[1], que efetuou a validação cruzada[2] com 10 *folds* de diversos modelos e ordenou a saída do modelo em ordem decrescente do coeficiente de determinação $R^2$. Avaliando os resultados, foi constatado através das métricas de desempenho que os algoritmos *Extra Trees*, *Gradient Boosting* e *Random Forest* obtiveram os melhores resultados dentre os demais algoritmos. Nos Apêndices B e C é apresentado o desempenho dos dez melhores algoritmos em relação à cada uma das novas bases de dados.

Nas Tabelas 3 e 4 são apresentados os algoritmos e seus parâmetros a serem utilizados na construção dos modelos de predição dos valores de impostos estaduais, através das bases de dados com e sem a identificação das empresas respectivamente. A partir dessa definição, os algoritmos selecionados foram avaliados considerando as respectivas bases de dados, com intuito de compreender os atributos estatisticamente mais relevantes e os resultados alcançados.

[1] compare_models <https://pycaret.org/compare-models/>
[2] validação cruzada <https://pt.wikipedia.org/wiki/ValidaÃğÃčo_cruzada>

Tabela 3 – Algoritmos e seus respectivos parâmetros a serem avaliados no desenvolvimento do Modelo de Predição com a identificação das empresas

| **Algoritmo** | **Parâmetros** |
|---|---|
| *Extra Trees* (ET) | bootstrap = False, ccp_alpha = 0.0, criterion = 'mse', max_depth = 11, max_features = 1.0, max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.001, min_impurity_split = None, min_samples_leaf = 2, min_samples_split = 5, min_weight_fraction_leaf = 0.0, n_estimators = 60, n_jobs = -1, oob_score = False, random_state = 1835, verbose = 0, warm_start = False. |
| *Gradient Boosting* (GBR) | alpha = 0.9, ccp_alpha = 0.0, criterion = 'friedman_mse', init = None, learning_rate = 0.3, loss = 'ls', max_depth = 3, max_features = 'sqrt', max_leaf_nodes = None, min_impurity_decrease = 0.0005, min_impurity_split = None, min_samples_leaf = 2, min_samples_split = 10, min_weight_fraction_leaf = 0.0, n_estimators = 80, n_iter_no_change = None, presort = 'deprecated', random_state = 1835, subsample = 0.75, tol = 0.0001, validation_fraction = 0.1, verbose = 0, warm_start = False. |
| *Random Forest* (RF) | bootstrap = True, ccp_alpha = 0.0, criterion = 'mse', max_depth = 9, max_features = 'log2', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.5, min_impurity_split = None, min_samples_leaf = 2, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 110, n_jobs = -1, oob_score = False, random_state = 3174, verbose = 0, warm_start = False. |

Fonte: Autoria própria.

Tabela 4 – Algoritmos e seus respectivos parâmetros a serem avaliados no desenvolvimento do Modelo de Predição sem a identificação das empresas

| **Algoritmo** | **Parâmetros** |
|---|---|
| *Extra Trees* (ET) | bootstrap = True, ccp_alpha = 0.0, criterion = 'mae', max_depth = 10, max_features = 1.0, max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.5, min_impurity_split = None, min_samples_leaf = 3, min_samples_split = 10, min_weight_fraction_leaf = 0.0, n_estimators = 230, n_jobs = -1, oob_score = False, random_state = 8756, verbose = 0, warm_start = False. |
| *Gradient Boosting* (GBR) | alpha = 0.9, ccp_alpha = 0.0, criterion = 'friedman_mse', init = None, learning_rate = 0.2, loss = 'ls', max_depth = 7, max_features = 'log2', max_leaf_nodes = None, min_impurity_decrease = 0.01, min_impurity_split = None, min_samples_leaf=3, min_samples_split=4, min_weight_fraction_leaf = 0.0, n_estimators = 30, n_iter_no_change = None, presort = 'deprecated', random_state = 8756, subsample = 0.4, tol = 0.0001, validation_fraction = 0.1, verbose = 0, warm_start = False. |
| *Random Forest* (RF) | bootstrap = False, ccp_alpha = 0.0, criterion = 'mae', max_depth = 11, max_features = 'sqrt', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.005, min_impurity_split = None, min_samples_leaf = 6, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 170, n_jobs=-1, oob_score = False, random_state = 8756, verbose = 0, warm_start = False. |

Fonte: Autoria própria.

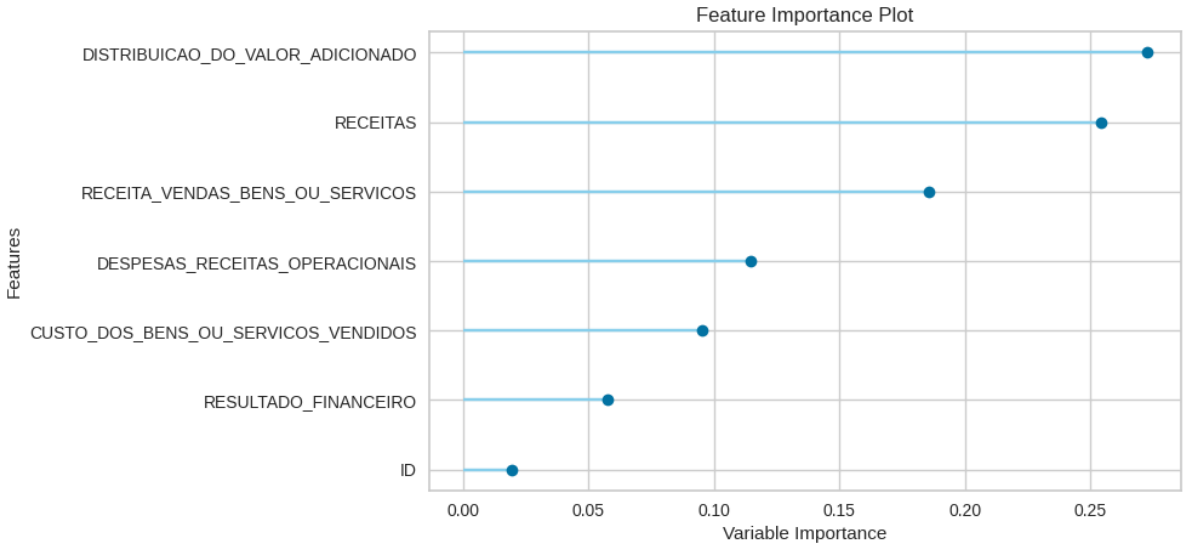
## 5.2 Análise dos Dados

Durante a realização do experimento foi feita a análise exploratória dos dados, definindo as melhores variáveis que tinham correlação com os impostos estaduais. Portanto, para os testes foram utilizados como entrada o identificador das empresas, distribuição do valor adicionado, receitas, resultado financeiro, despesas e receitas operacionais, custo dos bens ou serviços vendidos e as receitas com vendas de bens ou serviços para predizer na variável alvo os valores dos impostos estaduais. Após isso foi aplicado os algoritmos *Extra Trees*, *Gradient Boosting* e *Random Forest* da biblioteca *pycaret* que resultou na geração de alguns gráficos, que são apresentados a seguir, permitindo uma maior compreensão dos resultados.

### 5.2.1 Gráficos do Algoritmo *Random Forest*

Na Figura 5 é apresentado um gráfico com a distribuição das variáveis mais importantes, em ordem crescente, para o algoritmo *Random Forest* (RF). É importante destacar que o *pycaret* usa o atributo *feature_importance_* para definir o grau de importância de uma variável. Esta função se baseia em um conjunto de parâmetros pré-definidos para selecionar as features que mais contribuem para prever a variável de destino. Ela retorna um *array* onde cada elemento é uma *feature*, e irá dizer, em proporções, quão importante a *feature* é para o modelo. Por meio deste gráfico, é possível perceber que a variável *DISTRIBUICAO_DO_VALOR_ADICIONADO* é a que mais influencia nos impostos estaduais e de forma proporcional, ou seja, quanto maior o valor da distribuição do valor adicionado, maior será o valor do imposto estadual. Isso também ocorre com as outras variáveis, porém, com menor grau de influência.

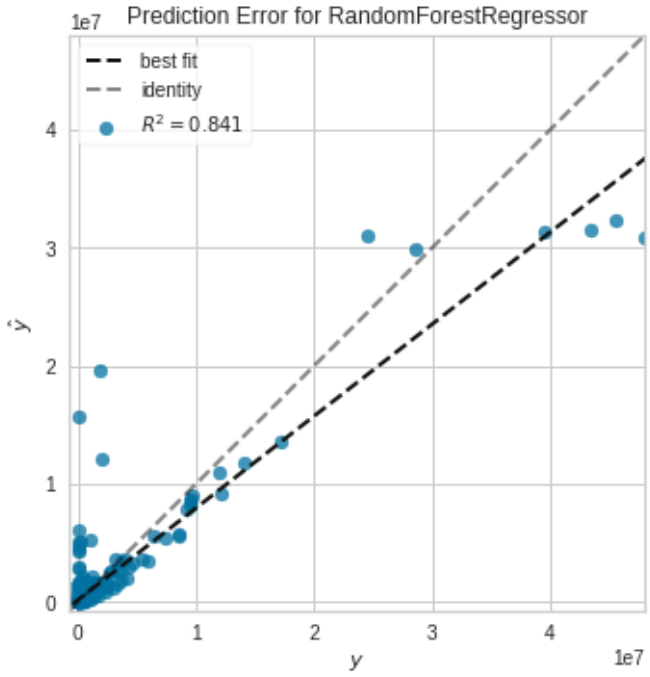
Figura 5 – Variáveis Importantes (*Random Forest*)




Fonte: Autoria própria.

Na Figura 6 é apresentado um gráfico com a predição de erro em que mostra uma linha reta traçada pelo *identity* e o *best fit* com o melhor ajuste do modelo, fazendo um alinhamento ideal do melhor segmento de similaridade e minimizando a distancia entre os pontos de dados. Por meio do gráfico é possível perceber no coeficiente de determinação $R^2$ que (84,1%) da variância dos dados pode ser explicada através do modelo construído pelo algoritmo RF, enquanto os outros (15,9%), teoricamente se trataria de uma variância residual, que é a diferença entre aquilo que queremos prever (Impostos Estaduais) com base nos valores que damos como entrada.

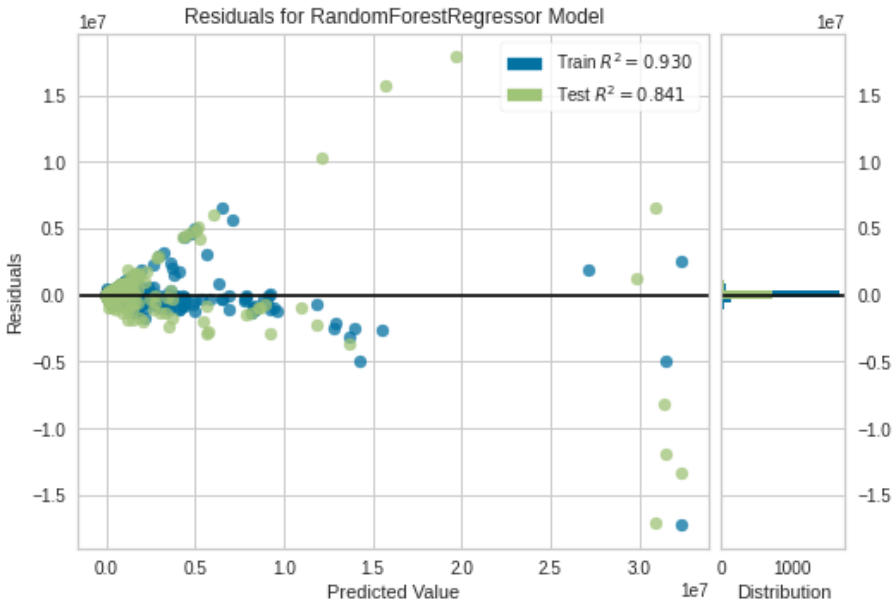
Figura 6 – Predição de Erro (*Random Forest*)




Fonte: Autoria própria.

No gráfico da Figura 7 a seguir, pode-se analisar a variação dos resíduos e a distribuição deles, representando a diferença entre o valor real e o valor estimado de y. O modelo RF está bem ajustado, apresentando bons resultados, com resíduos de treino e teste bastante elevados.

Figura 7 – Distribuição dos Resíduos (*Random Forest*)



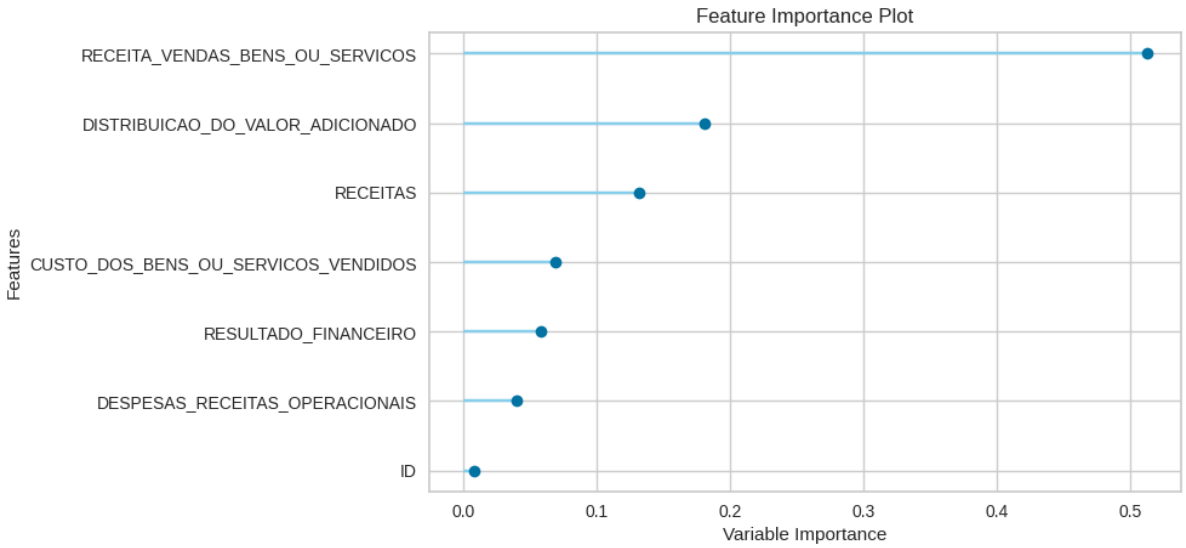

Fonte: Autoria própria.

### 5.2.2 Gráficos do Algoritmo *Gradient Boosting*

Na Figura 8 é apresentado um gráfico com as variáveis mais importantes para o algoritmo *Gradient Boosting* (GBR). Através do gráfico é possível perceber que a variável *RECEITA_VENDAS_BENS_OU_SERVICOS* é a que mais influencia de forma proporcional nos valores dos impostos estaduais.

Pode-se ver uma diferença entre as variáveis mais importantes para o algoritmo *Random Forest*, analisadas anteriormente no gráfico da Figura 5. Pois a variável que foi definida como sendo a mais importante para o GBR (*RECEITA_VENDAS_BENS_OU_SERVICOS*) foi escolhida como sendo a terceira variável mais importante para o RF, e a primeira variável mais importante para o RF (*DISTRIBUICAO_DO_VALOR_ADICIONADO*) foi escolhida como sendo a segunda mais importante para o GBR.

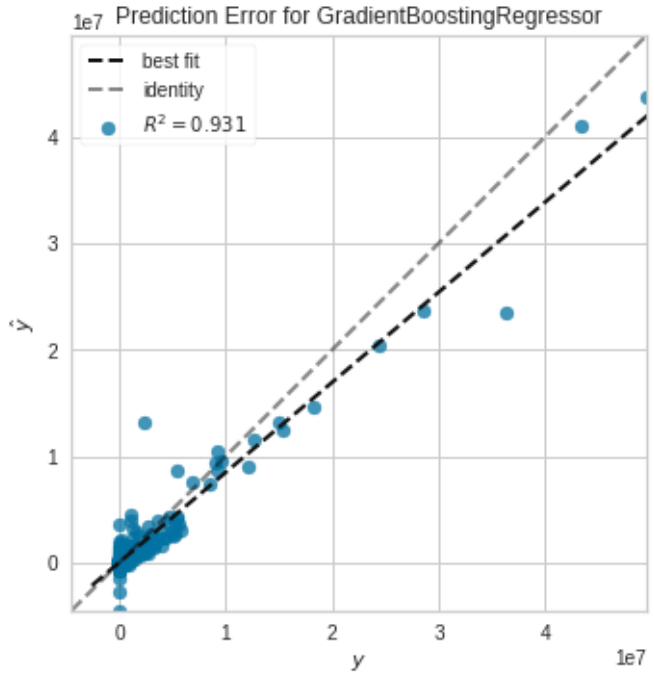
Figura 8 – Variáveis Importantes (*Gradient Boosting*)




Fonte: Autoria própria.

Na Figura 9 é apresentado um gráfico com a predição de erro para o algoritmo GBR. Através do gráfico é possível notar no coeficiente de determinação $R^2$ que 93,1% da variância dos dados pode ser explicada pelo modelo construído, enquanto os outros 6,9%, teoricamente se trataria de uma variância residual. Comparando a predição de erro do GBR com a do RF apresentado no gráfico da Figura 6, pode-se ver que o GBR apresentou um resultado melhor, tanto no coeficiente de determinação $R^2$ como no ajuste do modelo.

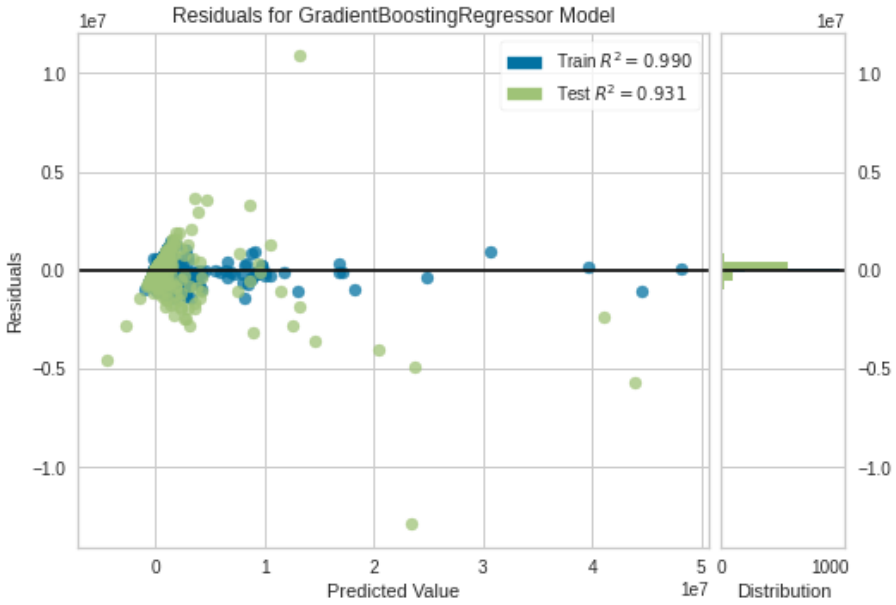
Figura 9 – Predição de Erro (*Gradient Boosting*)




Fonte: Autoria própria.

No gráfico da Figura 10, pode-se ver uma variação dos resíduos e a distribuição deles. Com os resultados do modelo GBR bem ajustados. Fazendo uma comparação com a distribuição dos resíduos do RF apresentado no gráfico da Figura 7, pode-se ver através da métrica $R^2$ que ambos os algoritmos apresentaram resultados acima de 90% para os resíduos do treino. Já para os resíduos de testes o GBR foi melhor, apresentando uma diferença de 9% em relação ao RF.

Figura 10 – Distribuição dos Resíduos (*Gradient Boosting*)

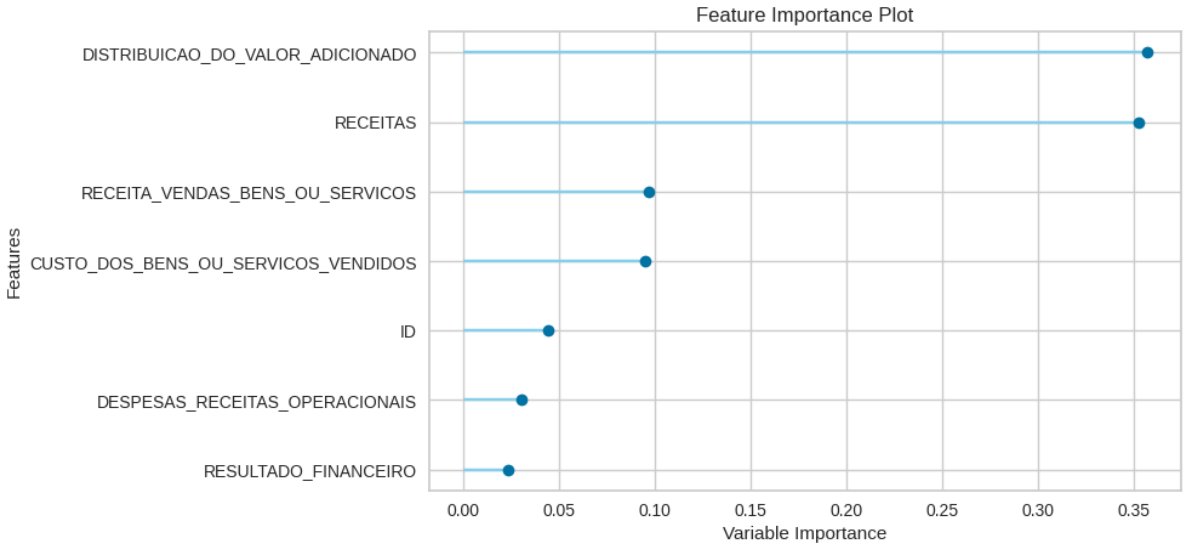



Fonte: Autoria própria.

### 5.2.3 Gráficos do Algoritmo *Extra Trees*

Na Figura 11 é apresentado um gráfico com a distribuição das variáveis mais interessantes para o algoritmo *Extra Trees* (ET). Para definir o grau de importância dessas variáveis o *pycaret* usou o atributo *feature_importance_*, que retorna um *array* com as *features* mais importantes para o modelo. Através do gráfico é possível perceber que a variável *DISTRIBUICAO_DO_VALOR_ADICIONADO* é a que mais influencia de forma proporcional nos valores dos impostos estaduais. Pode-se ver uma diferença entre as variáveis mais importantes para o algoritmo GBR analisadas anteriormente no gráfico da Figura 8. Pois a variável que foi definida como sendo a mais importante para o ET (*DISTRIBUICAO_DO_VALOR_ADICIONADO*) foi escolhida como sendo a segunda variável mais importante para o GBR, e a primeira variável mais importante para o GBR (*RECEITA_VENDAS_BENS_OU_SERVICOS*) foi escolhida como sendo a terceira mais importante para o ET.

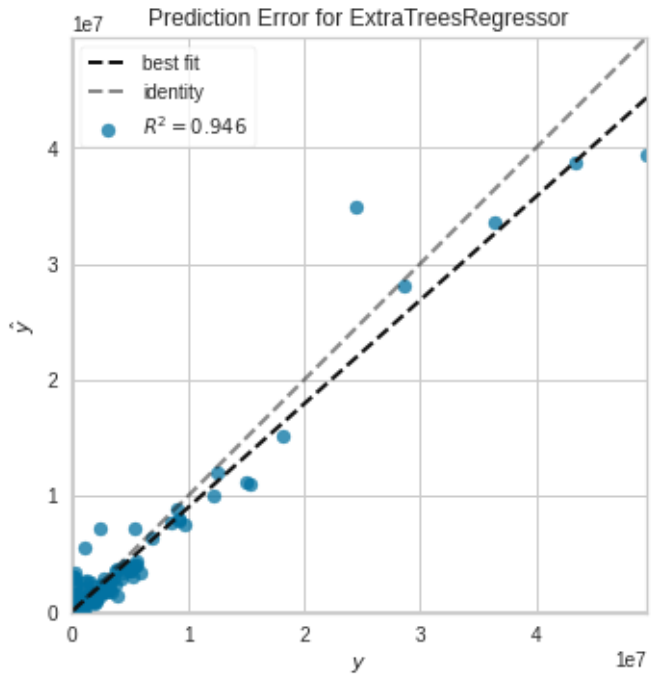
Figura 11 – Variáveis Importantes (*Extra Trees*)




Fonte: Autoria própria.

Na Figura 12 é apresentado o gráfico com a predição de erro do *Extra Trees*. Pode-se perceber através do coeficiente de determinação $R^2$, que o modelo do algoritmo ET performou muito bem, ou seja, apresentou o melhor ajuste, em comparação com os modelos dos algoritmos RF e GBR apresentados nos gráficos das Figuras 6 e 9, respectivamente. Ele conseguiu explicar 94,6% da variância dos dados, enquanto que o RF conseguiu explicar 84,1% e o GBR 93,1%.

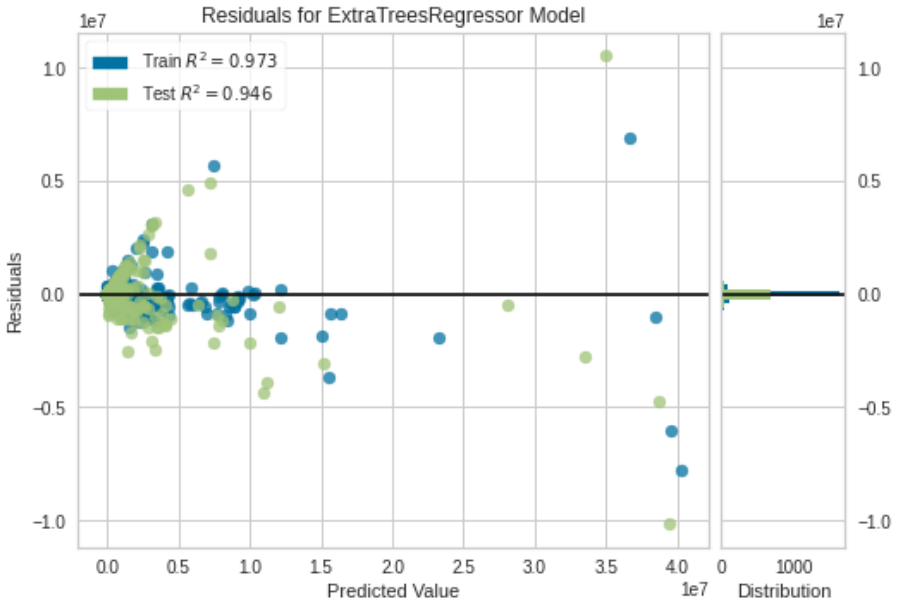
Figura 12 – Predição de Erro (*Extra Trees*)




Fonte: Autoria própria.

No gráfico da Figura 13, pode-se ver a distribuição dos resíduos para o modelo *Extra Trees*, representando a diferença entre o valor real e o valor estimado de y. Por meio do gráfico pode-se perceber que o ET apresentou um melhor resultado em comparação com os outros algoritmos avaliados nos gráficos das Figuras 7 e 10. Em relação aos resíduos de teste, o ET obteve um coeficiente de determinação $R^2$ de 94,6%. Enquanto o GBR e o RF apresentaram um $R^2$ de 93,1% e 84,1%, respectivamente.

Figura 13 – Distribuição dos Resíduos (*Extra Trees*)




Fonte: Autoria própria.

## 5.3 Modelo de predição sem a identificação das empresas

Na Tabela 5 são apresentados os resultados obtidos pelo *pycaret*, utilizando a base de dados sem a identificação das empresas, com 70% dos dados para treino e 30% para testes. Foi utilizado o método *tune_model*[3] do *pycaret* para tunar os três melhores modelos escolhidos pelo *compare_models*, e ao analisar as métricas de desempenho percebeu-se que os resultados dos algortmos melhoraram. Nota-se que o algoritmo que obteve o melhor desempenho foi o *Extra Trees* (ET) com um coeficiente de determinação R² de 79,6% e desvio padrão de 0.15%. Ja o *Random Forest* (RF) e o *Gradient Boosting* (GBR) apresentaram desempenhos semelhantes, com R² de 78,2% e 78,7%, e desvio padrão de 0,17% e 0,13%, respectivamente.

Tabela 5 – Desempenho dos três melhores algoritmos sem a identificação das empresas

| # | ALGORITMO | R² (%) | STD (%) |
|---|---|---|---|
| 1 | Extra Trees | 79,6 | 0,15 |
| 2 | Gradient Boosting | 78,7 | 0,13 |
| 3 | Random Forest | 78,2 | 0,17 |

Fonte: Autoria própria.

[3] tune_model <https://pycaret.org/tune-model/>

## 5.4 Modelo de predição com a identificação das empresas

Objetivando uma maior explanação dos resultados obtidos, na Tabela 6 faz-se uma comparação dos resultados obtidos pelos algoritmos utilizando a base de dados com a identificação das empresas. Nessa base de dados foi utilizado 70% dos dados para treino e 30% para testes. Também foi utilizado o método *tune_model* do *pycaret* para tunar os modelos e ao analisar as métricas de desempenho. Nesse sentido, percebeu-se que os resultados dos algoritmos melhoraram muito em relação a base de dados sem a identificação das empresas. Nota-se que o algoritmo ET continuou obtendo o melhor desempenho, com um coeficiente de determinação $R^2$ de 86,8% e desvio padrão de 0,11%. Já o RF oteve um $R^2$ de 76,9% e desvio padrão de 0,11%. Enquanto o GBR teve um $R^2$ de 83,2% e desvio padrão de 0,13%. Através da análise pode-se concluir que todos os algoritmos obtiveram resultados melhores em relação a base de dados sem a identificação das empresas. Porém o ET apresentou um desempenho bem superior se comparado com os outros algoritmos, pois teve um resultado de $R^2$ que conseguiu explicar 86,8% da variância dos dados.

Tabela 6 – Desempenho dos três melhores algoritmos com a identificação das empresas

| # | **ALGORITMO** | **R² (%)** | **STD (%)** |
|---|---|---|---|
| 1 | Extra Trees | 86,8 | 0,11 |
| 2 | Gradient Boosting | 83,2 | 0,13 |
| 3 | Random Forest | 76,9 | 0,25 |

Fonte: Autoria própria.

## 5.5 Discussões dos Resultados

Como pode-se verificar, o *PyCaret* é uma ótima ferramenta que tem muitas funções e bibliotecas que nos permite gerar modelos de aprendizado de máquina com poucas linhas de código. Através dele pode-se treinar vários algoritmos com muita facilidade, e assim saber quais os que apresentam os melhores resultados. Após a análise dos resultados, verificou-se através das métricas de desempenho que o modelo *Extra Trees* (ET) foi o que obteve o melhor resultado em ambas as bases de dados. Porém o valor da métrica de desempenho MAPE foi um pouco elevado em ambas as bases de dados, pois o trabalho não é eficiente para encontrar pequenas inconsistências, especialmente para valores pequenos, mas poderia detectar fraudes de grandes sonegadores. Na base de dados sem a identificação das empresas o modelo ET conseguiu explicar através do $R^2$, 79,6% da variância dos dados. Já na base de dados com a identificação das empresas ele se sobressaiu, conseguindo explicar 86,8% da variância dos dados. De acordo com a análise, pode-se concluir que a identificação das empresas na base de dados ajudou bastante no processo de regressão, fazendo com que o modelo do algoritmo *Extra Trees* apresentasse um resultado bem melhor.

Pois quanto maior o valor do coeficiente de determinação $R^2$, melhor é o resultado, já que indica que o valor do imposto predito pelo algoritmo foi próximo do valor real.

# 6 Disponibilização dos Modelos de Predição

Após a implementação dos modelos de predição, foi realizado o processo de disponibilização. Foi utilizado o pacote *Streamlit* para desenvolver uma aplicação *Web* e disponibilizar os modelos em rede. Pois é um dos principais pacotes gratuitos e que facilita a visualização e compartilhamento de dados a partir de códigos implementados na linguagem *Python* (STREAMLIT, 2022).

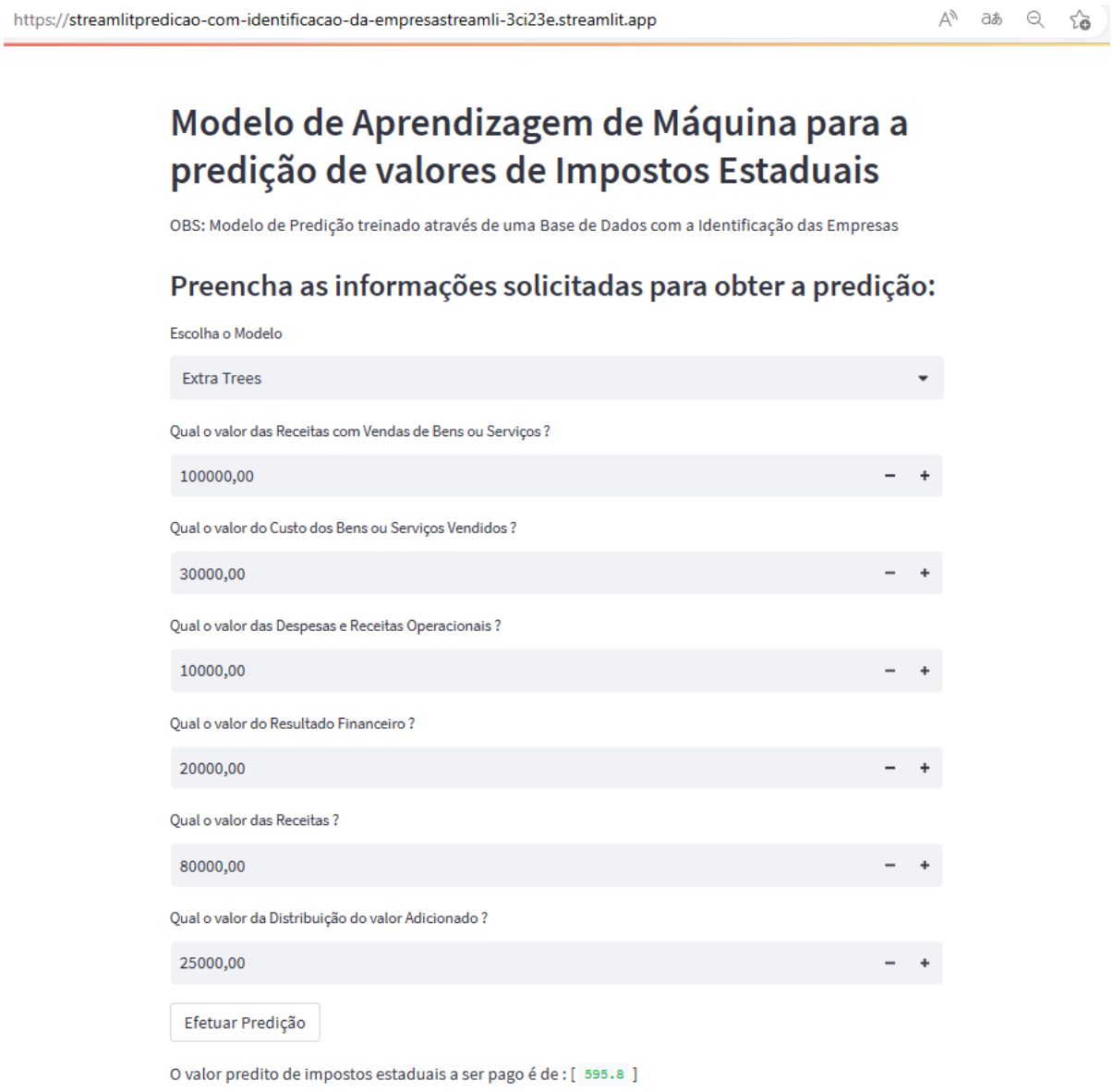
Os dois modelos foram disponibilizados[1] [2] para plataforma *Web*, permitindo o acesso através de qualquer navegador em *desktop* ou dispositivos móveis, assim, possibilitando o acesso a um maior número de usuários.

Na Figura 14 é apresentado a aplicação desenvolvida, com o modelo de predição treinado através de uma base de dados com a identificação das empresas. Através do formulário os usuários podem informar os dados de entrada necessários para a predição do modelo, tais como: receitas com vendas de bens ou serviços vendidos, custo dos bens ou serviços vendidos, despesas e receitas operacionais, resultado financeiro, receitas e distribuição do valor adicionado.

O usuário tem a opção de escolher o algoritmo que fará a predição logo antes de preencher os dados de entrada no formulário. Porém, é importante destacar que o algoritmo *Extra Trees* foi o que obteve os melhores resultados em ambos os modelos. Depois que o usuário informar os dados de entrada, será possível solicitar através dos modelos desenvolvidos, a predição do valor de impostos estaduais a serem pagos, para isso só precisa clicar no botão "Efetuar Predição". O valor da predição nos dois modelos é calculado a partir do método *predict*[3] do pacote *Sklearn*[4].

[1] Modelo de Predição sem a Identificação da Empresa <https://cutt.ly/g9oAFMD>
[2] Modelo de Predição com a Identificação da Empresa <https://cutt.ly/L9oAZvU>
[3] Predict <https://www.askpython.com/python/examples/python-predict-function>
[4] Scikit-Learn <https://scikit-learn.org/stable/index.html>

Figura 14 – Modelo de Aprendizagem de Máquina para a predição de valores de Impostos Estaduais




Fonte: Autoria própria.

# 7 Conclusões

A sonegação fiscal afeta a qualidade de vida da população e coloca em risco os projetos do estado. Por outro lado, as técnicas de AM oferecem vantagens para análise de dados financeiros, possibilitando prever se determinada empresa deixou ou não de cumprir com suas obrigações fiscais. Em meio às dificuldades causadas pela sonegação fiscal, este trabalho propôs o desenvolvimento de dois modelos para a predição de valores de impostos estaduais a serem pagos pelas empresas brasileiras. Sendo um modelo gerado através de uma base de dados sem a identificação das empresas e outro com a identificação. Para isso, foram avaliados três algoritmos de aprendizado de máquina: *Extra Trees*, *Gradient Boosting* e *Random Forest*.

É importante lembrar que há diferentes tipos de impostos nos estados, mas os valores disponíveis nas bases de dados só tinham o total dos impostos estaduais. Antes de treinar os modelos, foi feito um estudo de quais variáveis preditoras da base de dados tinham uma correlação mais forte com a variável alvo (impostos estaduais) para o problema de pesquisa. Então os modelos foram treinados a partir do identificador da empresa, receitas com vendas de bens ou serviços vendidos, custo dos bens ou serviços vendidos, despesas e receitas operacionais, resultado financeiro, receitas e a distribuição do valor adicionado.

A partir da etapa de treinamento e teste em ambas as bases de dados, foi observado que o algoritmo *Extra Trees*, obteve os melhores resultados em relação as métricas de desempenho, alcançando um resultado no coeficiente de determinação $R^2$ que conseguiu explicar (79,6%) da variância dos dados na base de dados sem a identificação das empresas e (86,8%) na base de dados com a identificação das empresas.

Para a continuação deste trabalho pretende-se: i) coletar uma versão mais atualizada dos dados para treinar os modelos de aprendizagem de máquina; ii) avaliar a utilização da técnica *Transfer Learning*, que permite reutilizar um modelo pré-treinado em um novo problema com um conjunto de dados distinto, e iii) verificar a utilização de outros algoritmos e comitês no novo conjunto de dados. Para que assim, os algoritmos de regressão possam trabalhar de forma mais eficiente e dar resultados melhores, descobrindo empresas suspeitas para que os auditores possam realizar a fiscalização e arrecadação dos impostos necessários. Pois quando o contribuinte sonega está provocando prejuízos a população em geral e colaborando para um péssimo atendimento das necessidades básicas, cuja assistência é de obrigação do estado e deveria ser proporcionada com o dinheiro público sonegado.

# Referências

ACHEK, I. *et al.* The determinants of tax evasion: a literature review. *International Journal of Law and Management*, Emerald Group Publishing Limited, 2015. Citado na página 14.

ALI, M. Pycaret: An open source, low-code machine learning library in python. *PyCaret version*, v. 2, 2020. Citado na página 28.

AZANK, F. *Como avaliar seu modelo de regressão*. 2020. Disponível em: <https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regress%C3%A3o-c2c8d73dab96>. Citado na página 18.

CAETANO, M. e. a. Modelos de classificação: aplicações no setor bancário. 2015. Citado na página 11.

CARVALHO, A. *et al.* Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, p. 45, 2011. Citado na página 15.

CARVALHO, P. d. B. *Curso de direito tributário*. [S.l.]: Saraiva Educação SA, 1985. Citado 2 vezes nas páginas 10 e 13.

CERRI, C.; CARVALHO, A. C. P. L. F. *Impostos Federais, Estaduais e Municipais*. 2017. Disponível em: <https://seer.sct.embrapa.br/index.php/cct/article/view/26381>. Citado na página 15.

COELHO, E. d. M. P. Ontologias difusas no suporte à mineração de dados: aplicações na secretaria de finanças da prefeitura municipal de belo horizonte. Universidade Federal de Minas Gerais, 2012. Citado 3 vezes nas páginas 24, 25 e 26.

DRESCH, A.; LACERDA, D. P.; VALLE. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. [S.l.]: Bookman Editora, 2015. Citado na página 21.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine learning*, Springer, v. 63, n. 1, p. 3–42, 2006. Citado na página 17.

GROLEMUND, G.; WICKHAM, H. *R für Data Science: Daten importieren, bereinigen, umformen, modellieren und visualisieren*. [S.l.]: O'Reilly, 2017. Citado 2 vezes nas páginas 14 e 15.

GRUS, J. *Data science from scratch: first principles with python*. [S.l.]: O'Reilly Media, 2019. Citado na página 11.

HAN; KAMBER; PEI. *Mineração de dados: conceitos e técnicas*. [S.l.: s.n.], 2011. Citado na página 15.

HO, T. K. Random decision forests. In: IEEE. *Proceedings of 3rd international conference on document analysis and recognition*. [S.l.], 1995. v. 1, p. 278–282. Citado na página 17.

KEELE, S. *et al. Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007. Citado na página 21.

LIMA, W. C. Colóquio temático sobre ética e sonegação. *São. Paulo: IPEP*, 2002. Citado na página 13.

MAYRINK, V. T. d. M. *Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo*. Tese (Doutorado) — Dissertação (Mestrado)—Brasil: Universidade Federal de Juiz de Fora (UFJF . . . , 2016. Citado na página 18.

MITCHELL, T. Machine learning.-new york, ny, usa: Mcgraw hill. *Inc. isbn*, v. 70428077, 1997. Citado na página 16.

MONARD. Capítulo 4: Conceitos sobre aprendizado de máquina. *Sistemas inteligentes: Fundamentos e Aplicações. São Paulo: Manole*, p. 39–56, 2003. Citado na página 15.

MORAES, B. R. de. *Compêndio de direito tributário*. [S.l.]: Forense, 1987. Citado na página 10.

PICCIRILLI, T. L. *et al.* Mineração de dados aplicada à classificação dos contribuintes do iss. *Goiânia: PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS*, 2013. Citado 3 vezes nas páginas 24, 25 e 26.

REGINA CLEIDE, F. P. *Os Municípios, o ISS e os Crimes Contra a Ordem Tributária*. 2005. Disponível em: <http://www.fiscosoft.com.br/main_artigos_index.php?PID=129504&printpage=_#nota_avaliacao>. Citado na página 10.

ROCHA, S. M. Mineração de dados aplicada à classificação de contribuintes de icms. *In XXXVI ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO*, 2016. Citado 3 vezes nas páginas 24, 25 e 26.

SANTOS, J. Evasão e fraude fiscais–uma perspectiva económica do fenómeno. *Prospectiva e Planeamento*, v. 2, p. 183–199, 1996. Citado na página 14.

SHARMA, A.; PANIGRAHI, P. K. A review of financial accounting fraud detection based on data mining techniques. *arXiv preprint arXiv:1309.3944*, 2013. Citado 3 vezes nas páginas 24, 25 e 26.

SIQUEIRA, M. L.; RAMOS, F. S. A economia da sonegação: teorias e evidências empíricas. *Revista de Economia contemporânea*, SciELO Brasil, v. 9, p. 555–581, 2005. Citado na página 14.

STREAMLIT. *Streamlit: The fastest way to build and share data apps*. 2022. Disponível em: <https://github.com/streamlit/streamlit>. Citado na página 48.

WISAENG, K. A comparison of different classification techniques for bank direct marketing. *International Journal of Soft Computing and Engineering (IJSCE)*, Citeseer, v. 3, n. 4, p. 116–119, 2013. Citado na página 10.

# Apêndices

# APÊNDICE A – Trabalhos Selecionados na Revisão Sistemática da Literatura

Figura 15 – Trabalhos Selecionados

| Autor | Algoritmo | Software | Principais Variáveis |
|---|---|---|---|
| Leite, Luana Priscilla Carreiro Varão | SVM | Weka | Notas Fiscais. |
| Ebberth Lopes de Paula | Algoritmo J48 e Arvore de decisão | Tamanduá | Porte, Natureza Jurídica, Classe da Atividade Econômica, Valores de Crédito, Débito, Saldo Credor e Saldo Apurado Devedor por Período, Deduções, ICMS a Recolher, Valores de Ajuste de Débito e Estorno de Débito e Crédito por ano |
| Abdulaal, Ahmed; Patel, Aatish; Charani, Esmita; Denny, Sarah; Mughal, Nabeela; Moore, Luke | Algoritmo Apriori | Python | Nome do Tomador de Serviços, Inscrição Municipal do Tomador, Nome do Prestador de Serviços, CNPJ do Prestador, Cidade do Prestador, Valor Total do Serviço e Valor do ISS Retido |
| Francisco Nobre de Oliveira, Luis Paulo Guimarães dos Santos. | Arvore de decisão, modelos logísticos, redes neurais, redes bayesianas, Gradiente Boosting Machine e Regressão Linear. | R | Declarações diversas prestadas por contribuintes à RFB, Demonstrativo de apuração de contribuições sociais (Dacon), identificação da empresa exportadora, o Tipo de atividade econômica realizada, Sua Situação Cadastral Atual e Passada (ativa, inativa ou suspensa), Movimentações Realizadas no Comércio Exterior pelas Empresas Exportadoras, Informações sobre os Valores e Quantitativos Exportados e Importados em cada Declaração de Exportação (DE) e Declaração de Importação (DI), Valores Declarados como Devidos |

Fonte: Autoria própria.

# APÊNDICE B – Desempenho dos algoritmos em relação aos dados sem a identificação das empresas

Figura 16 – Desempenho dos algoritmos sem a identificação das empresas

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE | TT (Sec) |
|---|---|---|---|---|---|---|---|---|
| **et** | Extra Trees Regressor | 177129.5093 | 5.852505e+11 | 6.817095e+05 | 0.8262 | 5.2013 | 253.9651 | 0.535 |
| **gbr** | Gradient Boosting Regressor | 237254.5319 | 7.889644e+11 | 8.459926e+05 | 0.7509 | 6.3941 | 412.0125 | 0.263 |
| **rf** | Random Forest Regressor | 214126.8747 | 8.702525e+11 | 8.665083e+05 | 0.7174 | 5.2525 | 292.4589 | 0.917 |
| **dt** | Decision Tree Regressor | 216820.2972 | 1.359611e+12 | 9.387214e+05 | 0.6545 | 5.2494 | 101.9148 | 0.018 |
| **knn** | K Neighbors Regressor | 239156.1547 | 1.416551e+12 | 1.103850e+06 | 0.6260 | 4.9771 | 271.9931 | 0.015 |
| **omp** | Orthogonal Matching Pursuit | 377516.3525 | 1.659612e+12 | 1.194411e+06 | 0.5536 | 6.6157 | 972.0693 | 0.009 |
| **lightgbm** | Light Gradient Boosting Machine | 297256.7660 | 2.275687e+12 | 1.376222e+06 | 0.3854 | 5.7396 | 279.0384 | 0.070 |
| **ada** | AdaBoost Regressor | 839194.6229 | 1.557590e+12 | 1.224468e+06 | 0.3414 | 8.5500 | 13051.2074 | 0.175 |
| **par** | Passive Aggressive Regressor | 409198.0379 | 3.244561e+12 | 1.702516e+06 | 0.1072 | 5.5479 | 422.6971 | 0.011 |
| **br** | Bayesian Ridge | 395815.1536 | 1.964286e+12 | 1.316341e+06 | 0.0213 | 6.9450 | 1323.9512 | 0.010 |

Fonte: Autoria própria.

# APÊNDICE C – Desempenho dos algoritmos em relação aos dados com a identificação das empresas

Figura 17 – Desempenho dos algoritmos com a identificação das empresas

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE | TT (Sec) |
|---|---|---|---|---|---|---|---|---|
| **et** | Extra Trees Regressor | 184251.3040 | 7.128883e+11 | 7.803948e+05 | 0.8293 | 5.0794 | 108.6437 | 0.855 |
| **gbr** | Gradient Boosting Regressor | 246152.9933 | 9.837657e+11 | 9.285629e+05 | 0.7298 | 6.2815 | 379.2048 | 0.625 |
| **rf** | Random Forest Regressor | 236241.2403 | 1.221138e+12 | 1.021708e+06 | 0.6211 | 5.1411 | 129.2201 | 1.076 |
| **huber** | Huber Regressor | 325804.7413 | 1.490854e+12 | 1.166083e+06 | 0.6051 | 5.9335 | 380.0402 | 0.118 |
| **knn** | K Neighbors Regressor | 247199.3000 | 1.709177e+12 | 1.234876e+06 | 0.5987 | 4.9152 | 117.1613 | 0.030 |
| **omp** | Orthogonal Matching Pursuit | 390327.6069 | 1.854900e+12 | 1.296778e+06 | 0.5358 | 6.5473 | 832.0792 | 0.018 |
| **ada** | AdaBoost Regressor | 794410.1619 | 1.611641e+12 | 1.241659e+06 | 0.5331 | 8.3681 | 10078.9551 | 0.257 |
| **br** | Bayesian Ridge | 386760.9127 | 1.719122e+12 | 1.248644e+06 | 0.5261 | 6.8339 | 1068.5195 | 0.020 |
| **llar** | Lasso Least Angle Regression | 388039.5978 | 1.720495e+12 | 1.248414e+06 | 0.5247 | 6.8326 | 1156.5334 | 0.024 |
| **en** | Elastic Net | 388051.4375 | 1.720391e+12 | 1.248297e+06 | 0.5246 | 6.8328 | 1157.8149 | 0.033 |

Fonte: Autoria própria.