Análise preditiva da incidência e mortalidade da COVID-19 em municípios brasileiros usando aprendizagem de máquina

Jorge W. G. Felix[1], Marcus W. A. de Carvalho[1], Raquel Vigolvino Lopes[1]

[1]Departamento de Ciências Exatas (DCX) - Universidade Federal da Paraíba (UFPB)

Rua da Mangueira, s/n, Companhia de Tecidos - Rio Tinto - Brasil

jorge.wellison@dcx.ufpb.br

marcuswac@dcx.ufpb.br

raquel@dcx.ufpb.br

Abstract.

This article analyses the influence of socioeconomic factors and mobility on the incidence and mortality of COVID-19 in brazilian municipalities using machine learning techniques. To that, from the case databases of COVID-19 of Brasil.IO, socioeconomic of the Instituto Brasileiro de Geografia e Estatística and mobility of InLoco, three classification algorithms were used: decision tree, Naive Bayes and SVM (support vectors machine). Then, it was made a effectiveness analysis using metrics of, accuracy, precision, recall and F1 score.

Resumo.

Este artigo busca analisar a influência de fatores socioeconômicos e de mobilidade na incidência e mortalidade da COVID-19 em municípios brasileiros usando técnicas de aprendizagem de máquina. Para tal, a partir das bases de dados de casos da COVID-19 do Brasil.io, socioeconômicos do Instituto Brasileiro de Geografia e Estatística (IBGE) e de mobilidade da InLoco, utilizou-se três algoritmos de classificação do nível de incidência e mortalidade da COVID-19 nos municípios, sendo eles: Árvore de Decisão, Naive Bayes e SVM (máquina de vetores de suporte). Em seguida, foi realizada uma análise da eficácia dos classificadores utilizando as métricas de acurácia, precisão, revocação e F1 score.

---

1 ”Trabalho de conclusão de curso, sob orientação do professor Marcus W. A. de Carvalho e co-orientação da professora Raquel Vigolvino Lopes, submetido ao Curso de Bacharelado em Sistemas de Informação do Centro de Ciências Aplicadas e Educação (CCAE) da Universidade Federal da Paraíba, como parte dos requisitos necessários para obtenção do grau de BACHAREL EM SISTEMAS DE INFORMAÇÃO.”

1. INTRODUÇÃO

A COVID-19 é uma doença definida como uma infecção viral de alta transmissibilidade que afeta majoritariamente o trato respiratório e gastrointestinal, causada pelo vírus SARS-CoV-2 - Síndrome Respiratória Aguda Grave Coronavírus 2 (ZHU et al, 2020).

A COVID-19 é uma zoonose, onde sua transmissão ocorre de humanos para humanos, principalmente pelo contato de gotículas respiratórias, perdigotos provindos do trato respiratório ou pela saliva, de pacientes infectados na forma sintomática ou assintomática. A transmissão também acontece através do contato direto com esses indivíduos contaminados, e indireto pelas mãos, objetos ou superfícies com o vírus. Além dessas formas, pode ocorrer a infecção por meio de aerossóis em pacientes que necessitam de procedimentos relacionados às vias aéreas, como aspiração das vias aéreas (SILVA et al., 2020; ANVISA, 2020).

Em novembro de 2020, havia cerca de 6 milhões de casos confirmados e 170 mil óbitos confirmados (MINISTÉRIO DA SAÚDE, 2020). Alguns estudos mostraram que pessoas acima de 60 anos e a população com uma imunidade funcional fraca, diabetes, doenças cardiovasculares, doenças respiratórias crônicas, câncer, problemas renais e disfunção hepática estão com um maior risco para contrair a forma grave do COVID-19 (HAYAT et al., 2020).

Um outro estudo recente realizado no Brasil mostra que quanto mais desenvolvido for o município, maiores serão os números de infectados e mortos. Um dos motivos é que nestes locais há uma alta taxa de interações sociais entre a população, aumentando a capacidade e velocidade do vírus se espalhar (ROUBAUD, 2020). Porém, observou-se também que os municípios mais pobres também foram muito atingidos. A prevalência dos casos de COVID-19 na parte da população socioeconomicamente vulnerável é quase o dobro dos mais ricos. A baixa condição econômica tem efeito direto no PIB per capita do país, mostrando que quanto maior a desigualdade, maior a taxa de contaminação e óbitos (HALLAL *et al*, 2020).

Além disso, a cor da pele é uma das variáveis que apresenta forte relação à COVID-19. Sob as mesmas condições, as cidades com mais pessoas brancas na sua população foram as que tiveram menos casos. Isso é reflexo de fatores que a população negra está a mercê em sua maioria: educação, pobreza, informalidade, moradia, etc. (ROUBAUD, 2020).

Com isso, fez-se necessário realizar pesquisas em todo o mundo a fim de investigar possíveis padrões no comportamento do vírus e sua propagação.Uma das técnicas para a análise desses dados é a de *Machine Learning* (ou Aprendizagem de Máquina). De acordo com Monard

e Baranauskas (2003), o aprendizado de máquina é uma área da inteligência artificial que tem como objetivo desenvolver técnicas computacionais sobre o aprendizado, além de possibilitar a criação de sistemas capazes de adquirir conhecimento de maneira automática. Cada sistema de aprendizado possui suas próprias características, não havendo, portanto, um algoritmo universal, capaz de solucionar todos os problemas. Nestas circunstâncias, a Aprendizagem de Máquina tem sido notável, podendo ser utilizada em diversas áreas em nosso cotidiano. No entanto, é importante compreender o poder e a limitação dos diversos algoritmos de Aprendizagem de Máquina utilizando alguma metodologia que permita avaliar os conceitos induzidos por esses algoritmos em determinados problemas.

Diante deste contexto, o objetivo deste trabalho é analisar a influência de fatores socioeconômicos e de mobilidade na incidência e mortalidade do COVID-19 em municípios brasileiros usando técnicas de aprendizagem de máquina. Mais especificamente, pretende-se prever os níveis de incidência e mortalidade da COVID-19 em municípios brasileiros com base em seus fatores socioeconômicos e de mobilidade usando algoritmos classificação. Com base nas classificações feitas pelos algoritmos e a eficácia de suas predições, serão analisados os fatores que mais influenciaram na classificação dos municípios brasileiros em relação à incidência e à mortalidade da COVID-19 para identificar possíveis padrões.

Para atingir os objetivos, primeiramente serão extraídos dados socioeconômicos, de mobilidade e de casos e óbitos da COVID-19 para os municípios brasileiro, sendo feito posteriormente o cruzamento desses dados. Em seguida, será necessário agrupar os municípios brasileiros com base em suas incidências e mortalidades da COVID-19, dividindo-os em 3 níveis (baixo, médio e alto) para que sejam aplicados e avaliados os algoritmos de classificação.

Destaca-se a importância desse trabalho através do auxílio no entendimento das diferenças socioeconômicas dos municípios brasileiros e o impacto delas nos números de incidência e mortalidade da COVID-19, expondo a realidade atual do Brasil. Além disso, a classificação com base na aprendizagem de máquina irá ajudar outros pesquisadores a intervir nos possíveis números da COVID-19 no país, já que ao prever a incidência e mortalidade e identificar os fatores mais relevantes, pode-se agir mais rapidamente para evitar o aumento de casos. Por fim, o fato do tema ser muito recente e ainda haver uma escassez na literatura da área requer que mais pesquisas sejam desenvolvidas para melhor respaldar a comunidade acadêmica e os tomadores de decisão.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão explanados os fundamentos e os conceitos básicos de algoritmos de aprendizagem de máquina — supervisionados e não-supervisionados — de classificação e clusterização usados neste trabalho. Também serão discutidos trabalhos relacionados.

2.1 Aprendizagem de máquina

Segundo Brunialti, Peres, Freire (2015) a teoria de Aprendizagem de Máquina (AM) é baseada nos princípios do aprendizado indutivo (AI), ou seja, modelos são determinados a partir de um conjunto de dados ou representações de experiências. Normalmente, o aprendizado indutivo é implementado por algoritmos que processam um conjunto de dados e extraem um modelo capaz de explicar ou representar os dados sob algum aspecto.

Neste trabalho foram utilizadas três técnicas de aprendizagem de máquina supervisionada, juntamente com uma técnica não-supervisionada, trabalhando em conjunto para a classificação.

2.1.1 Classificação com o método de Árvores de decisão

A Árvore de Decisão é um modelo estatístico que utiliza um treinamento supervisionado para classificação com base em dados de entrada. Em sua composição é usado um conjunto de treinamento formado pelos dados de entrada (nomeados de previsores) e, posteriormente, os de saída (as classes).

Estes modelos utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em subproblemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema (Gama, 2004).

2.1.2 Classificação com o método Naive Bayes

Modelos bayesianos como o Naive Bayes geralmente são algoritmos robustos, simples e que possuem um bom desempenho sem a necessidade de grande número de dados para obter classificações com um bom grau de assertividade.

É do tipo supervisionado, já que são fornecidas ao algoritmo de AM as instâncias juntamente com seus rótulos (ou seja, as classes). Seguindo o paradigma estatístico, o algoritmo faz uso de fórmulas estatísticas e cálculos de probabilidades para realizar a classificação (Mitchell, 1997).

2.1.3 Classificação com o método SVM

O SVM consiste em uma técnica computacional de aprendizado para problemas de reconhecimento de padrão. Os algoritmos de SVM têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre classes por meio da minimização dos erros (Vapnik, 1995).

2.1.4 Agrupamento com o método K-means

O k-means é um algoritmo de aprendizagem não-supervisionada, que através da clusterização busca por padrões e gera agrupamentos através de um conjunto de dados. De acordo com Karypis (2002), este algoritmo divide os dados em grupos úteis ou significativos, chamados clusters, nos quais a similaridade intra-cluster é maximizada e a similaridade inter-cluster é minimizada. Estes clusters descobertos podem ser usados para explicar as características da distribuição dos dados subjacentes e assim servir como base para várias técnicas de análise e mineração de dados

2.2 Validação cruzada com K-Fold

A validação cruzada K-fold é uma técnica computacional intensiva, que usa todas as amostras disponíveis como amostras de treinamento e teste (Duchesne, 2005). Com isso, o conjunto de dados de treino não é executado e testado uma única vez, e sim K vezes..

2.3 Métricas de desempenho dos classificadores

As métricas são usadas para medir o grau de assertividade de cada classificador, através do resultado de seus testes e de sua predição. Neste trabalho foram utilizadas as métricas abaixo para avaliação dos algoritmos de classificação.

2.3.1 Matriz de confusão

Dado um classificador e uma instância, existem quatro possíveis resultados. Se a instância for positiva e for classificada como positiva, é considerado um verdadeiro positivo (TP). Se é classificada como negativa, é considerado um falso negativo (FN). Se a instância é negativa e é classificada como negativa, é contada como um verdadeiro negativo (TN). Se for classificada como positiva, é contada como um falso positivo (FP) (FAWCETT, 2005). Já a somatória do resultado de todas as instâncias é o total (P).

Na tabela 1, tem-se o exemplo de uma matriz de confusão de ordem 2, que identifica os quatro possíveis resultados de um classificador e suas instâncias

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Tabela 1: Exemplo de matriz de confusão 2x2

De acordo com Fawcett (2005), métricas como acurácia, precisão, revocação e f1-score utilizam valores de ambas as colunas da matriz de confusão e são usadas para os cálculos abaixo citados.

Métrica	Descrição	Cálculo
Acurácia	Indica a performance do modelo a ser avaliado. Ou seja, dentre todas as predições quantas realmente foram corretas	FP/P
Precisão	Mostra especificamente quais entre as classificações da classe “Sim” estão corretas.	TP/(TP+FP)
Revocação	Frequência em que o classificador classifica corretamente como positivo os exemplos que são de fato positivos.	TP/(TP+FN)
F1-score	Média harmônica entre precisão e revocação.	2*(precisão/revocação) / (precisão+revocação)

Tabela 2: Cálculo e descrição das métricas

3. METODOLOGIA

Neste trabalho foi realizada uma pesquisa quantitativa, onde através dos sites do Brasil.io do Instituto Brasileiro de Geografia e Estatística (IBGE) e da InLoco, foi realizada a coleta inicial e o cruzamento dos dados.

Com base nesses dados, foram utilizados três algoritmos para classificar os municípios em níveis de incidência e mortalidade da COVID-19, sendo eles: Árvore de Decisão, Naive Bayes e SVM (máquina de vetores de suporte). Em seguida, foi realizada uma análise da eficácia dos algoritmos utilizando as métricas de acurácia, precisão, revocação e f1-score.

3.1 Fontes dos dados

Para a obtenção dos dados sobre a COVID-19 nos municípios brasileiros foi utilizada a base do Brasil.io. Dados estes disponibilizados em formato .csv contendo as seguintes colunas: semana epidemiológica, data da última atualização, dias a partir do 1º caso, UF (estado), município, código IBGE do município, tipo de local (cidade ou estado), confirmados acumulados, confirmados acumulados por 100 mil habitantes, confirmações no dia, óbitos acumulados, óbitos no dia, óbitos/confirmados (taxa de mortalidade), população estimada em 2019 (segundo o IBGE) e mortalidade por 100 mil habitantes.

Para a obtenção dos dados socioeconômicos do IBGE — 71 atributos, extraídos do Censo 2010 — foram utilizadas as bases do site Atlas Brasil[1]. E para a obtenção dos dados de mobilidade foi usada a base de dados da empresa InLoco[2], providenciada através de um convênio para realizar pesquisas no âmbito da UFPB.

3.2 Cruzamento dos dados[2]

Para realizar a classificação dos municípios, previamente foi efetuado o cruzamento dos dados dos três datasets (da COVID-19, socioeconômicos e de mobilidade), selecionando — com o apoio da biblioteca Pandas e o ambiente da plataforma *Google Colab*[3] — as colunas necessárias para a criação do classificador.

Após a obtenção dos dados dos municípios de todo o Brasil, então foi efetuado o tratamento com os seguintes passos:

- Carregamento do dataset
- Pré-processamento dos dados

1 https://atlasbrasil.org.br/
2 https://mapabrasileirodacovid.inloco.com.br/pt/
3 https://colab.research.google.com/drive/1X2VrVYF09_tbvrMOpRFv6IY9FqhsAYuO?usp=sharing

- Cálculo de média de mobilidade durante os meses da pandemia
- Renomeação de colunas para melhor assimilação;
- Seleção do último ano (2010) em que a pesquisa de estimativa de população foi realizada. (Entre três existentes no dataset: 1991, 2000 e 2010);

Em seguida, foi realizado o tratamento dos datasets do Brasil.io, a fim de remover as colunas que **não** seriam utilizadas, visando usar apenas as colunas das classes (incidência e mortalidade), ou seja, as variáveis de saída do classificador. Logo, as duas **únicas** colunas utilizadas foram as que continham: o número de casos confirmados por 100 mil habitantes acumulado e o número de mortes por mil habitantes acumulado para cada município.

Efetuadas essas alterações, o próximo passo foi elaborar um modo de juntar os data frames e gerar um único dataset, associando todos os atributos a seus municípios. Portanto, através do código do IBGE de cada município a junção foi realizada, totalizando 5565 linhas onde cada uma representa um município brasileiro.

3.3 Separação de previsores e classe

Após a junção das tabelas citadas anteriormente, foi possível selecionar as colunas que seriam usadas nos dados de entrada e as que seriam usadas nos dados de saída. Desta maneira, usando a biblioteca Pandas de Python, foram separados os atributos entre previsores e classe. Seguindo os seguintes passos:

- Armazenando na variável “previsores” os atributos de entrada;
- Armazenando na variável “classe_inc” a coluna que foi manipulada posteriormente para calcular os níveis de incidência (casos por 100 mil habitantes);
- Armazenando na variável “classe_mort” a coluna que foi manipulada posteriormente para calcular os níveis de mortalidade (mortes por 100 mil habitantes);

3.3.1 Técnica para agrupamento das classes (variáveis de saída)

Em virtude de ser necessária uma clusterização dos dados para dividir a incidência e a mortalidade em classes, através da função k-means da biblioteca “sklearn” foi utilizado o parâmetro “n_clusters” igual a 3, para desmembrar os níveis de incidência e mortalidade da COVID-19 em: baixa, média e alta.

3.4 Criação e treinamento dos modelos

Nesta etapa foram construídos os algoritmos responsáveis pela classificação, são eles: Naive Bayes, SVM e Árvore de Decisão. Cada modelo foi treinado com duas classes diferentes, a de incidência e a de mortalidade.

3.4 Testes dos modelos

Para os testes de cada classificador, buscando visualizar a capacidade de generalização de cada algoritmo, foi utilizado o método de validação cruzada K-fold com 10 iterações, onde 20% da base de dados foram utilizados para teste em cada iteração, enquanto os 80% restantes foram utilizados treino. Além disso, a função Shuffle foi ativada para 'embaralhar' os municípios, desagrupando-os de seus estados ou qualquer outro tipo de ordem prévia.

4. RESULTADOS

Nesta seção, serão apresentados os resultados obtidos a partir do treinamento dos algoritmos de aprendizagem de máquina Árvore de Decisão, SVM e Naive Bayes, sendo utilizado o dataset "covid-idhm", onde foram cruzados dados da pandemia da COVID-19 (incidência e mortalidade) com dados socioeconômicos e de mobilidade para 5565 municípios do Brasil, conforme descrito na metodologia.

4.1 Avaliação dos testes

Abaixo nas tabelas 2 e 3, podemos observar a porcentagem de acerto de cada classificador nas 10 execuções do K-fold para os testes, tanto nos níveis de incidência quanto nos de mortalidade. Em negrito e em vermelho temos o melhor resultado, entre os três classificadores, para cada métrica.

	Árvore de Decisão	**Naive Bayes**	**SVM**
Acurácia	58,2%	59%	**61,8%**
Precisão	47,8%	**50,2%**	45,3%
Revocação	43,5%	**45,6%**	45%
F1-Score	**47,9%**	43,8%	45,1%

Tabela 2 - Avaliação para a classificação da incidência.

	Árvore de Decisão	**Naive Bayes**	**SVM**
Acurácia	53,5%	55,3%	**58%**
Precisão	48,3%	55,8%	**63,1%**

Revocação	48,3%	**50,7%**	48,2%
F1-Score	**48,3%**	**48,3%**	45,1%

Tabela 3 - Avaliação da classificação da mortalidade.

Portanto, em relação à Acurácia o SVM obteve o melhor resultado. O Naive Bayes obteve um resultado estático, se mantendo acima dos 50% de precisão em ambas as classes (Incidência e Mortalidade) e se destacando também na Revocação. E por fim, em relação ao F1-Score a Árvore de Decisão obteve o melhor resultado.

4.3 Predição e matriz de confusão

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Imagem 01: Matriz 3x3

Numa matriz de ordem 3 (3 linhas e 3 linhas) temos que os resultados de valor positivo estão localizados nas posições a11, a22 e a33 da matriz. Vide imagem 01 acima.

- **Árvore de Decisão**

Na matriz de confusão para a incidência da COVID-19 pode-se notar que o algoritmo de Árvore de Decisão foi mais assertivo nos níveis baixo e alto, como mostra a Figura 3.

```
([[1613,  216,  805],
  [ 192,   91,  148],
  [ 803,  172, 1525]])
```

Figura 3: Matriz de confusão da *Decision Tree com a classe de Incidência*

Na matriz de confusão para a mortalidade o comportamento foi similar ao de incidência, como podemos observar na Figura 4.

```
([[1193,  238,  731],
  [ 237,  248,  344],
  [ 699,  376, 1499]])
```

Figura 4: Matriz de confusão da *Decision Tree* com a classe de Mortalidade

- **Naive Bayes**

Na matriz de incidência pode-se notar que o algoritmo Naive Bayes foi mais assertivo nos níveis baixo e alto, porém acertou poucos no nível médio e errou consideravelmente no nível alto, como mostra a Figura 5.

```
([[1062,  177, 1395],
  [  51,   38,  342],
  [ 145,  167, 2188]])
```

Figura 5: Matriz de confusão do Naive Bayes *com a classe de Incidência*

Na matriz de mortalidade o comportamento foi diferente da incidência, havendo muitos 'falsos nível baixo'. Como podemos observar na Figura 6.

```
([[1903,  159,  100],
  [ 537,  226,   66],
  [1289,  335,  950]])
```

Figura 6: Matriz de confusão do Naive Bayes *com a classe de Mortalidade*

- **SVM**

Na matriz de incidência pode-se notar que o algoritmo foi mais assertivo nos níveis baixo e alto, porém o nível médio não pôde ser dividido, como mostra a Figura 7.

```
([[1143,    0, 1491],
  [  69,    0,  362],
  [ 203,    0, 2297]])
```

Figura 7: Matriz de confusão do SVM *com a classe de Incidência*

Na matriz de mortalidade o comportamento foi o mesmo do de incidência, porém com mais 'falso nível baixo' e com algumas predições da classe nível médio. Como podemos observar na Figura 8.

```
([[1953,    3,  206],
  [ 530,   58,  241],
  [1327,   29, 1218]])
```

Figura 8: Matriz de confusão do SVM *com a classe de Mortalidade*

4.4 Métricas de avaliação da predição

Anteriormente foi mostrado a avaliação dos testes do K-Fold, onde foram realizadas 10 execuções com partes diferentes do dataset e obtivemos a avaliação média de todas as execuções. Abaixo nas tabelas 4 e 5, pode-se observar o percentual de acerto na execução da predição **única,** utilizando o dataset completo em cada classificador para uma avaliação geral de capacidade de generalização dos classificadores. Em negrito e em vermelho temos o melhor resultado.

Deve-se destacar que, em razão de obtermos classes desproporcionais, a fórmula mais indicada para avaliação é a do **F1-score**, métrica esta que está sendo priorizada em nossa avaliação.

	Árvore de Decisão	**Naive Bayes**	**SVM**
Acurácia	58%	59%	**61,8%**
Precisão	47%	**50%**	45,3%
Revocação	**47,7%**	45%	45%
F1-Score	**47,6%**	44%	41%

Tabela 4 - Métricas de desempenho da predição da incidência

	Árvore de Decisão	**Naive Bayes**	**SVM**
Acurácia	52,8%	55,3%	**58%**
Precisão	47,6%	55,8%	**62,9%**
Revocação	47,7%	**50,7%**	48,2%
F1-Score	47,7%	**48,4%**	45,1%

Tabela 5 - Métricas de desempenho da predição da mortalidade

4.5 Atributos considerados mais importantes

- **Árvore de Decisão**

Após o treinamento , foi possível classificar também os atributos mais importantes para o algoritmo, atributos estes presentes nos gráficos abaixo para a mortalidade.

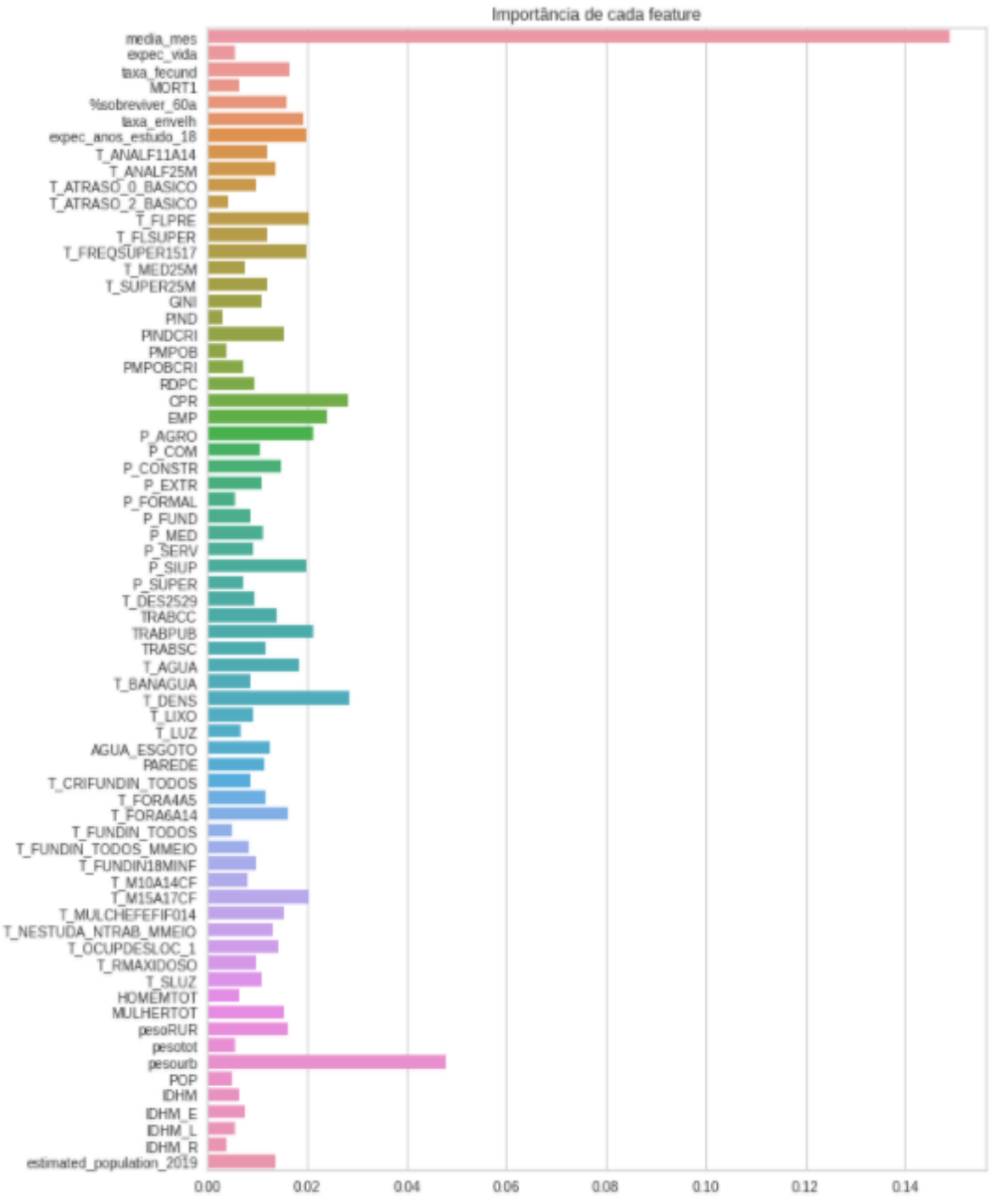

Figura 1: Atributos mais importantes para o classificador *Decision Tree com a classe de Mortalidade*

Percebe-se que o atributo de ***media_mes*** (média mensal de isolamento social de cada município de março até outubro) se destaca nesse classificador, além do ***pesourb***, que se refere à população residente na área urbana, o ***t_dens***, referente ao percentual da população que vive em domicílios com densidade superior a 2 pessoas por dormitório e o ***CPR*** que diz respeito à porcentagem de trabalhadores por conta própria com mais de 18 anos.

Enquanto em relação à incidência, os atributos abaixo se destacaram no classificador.

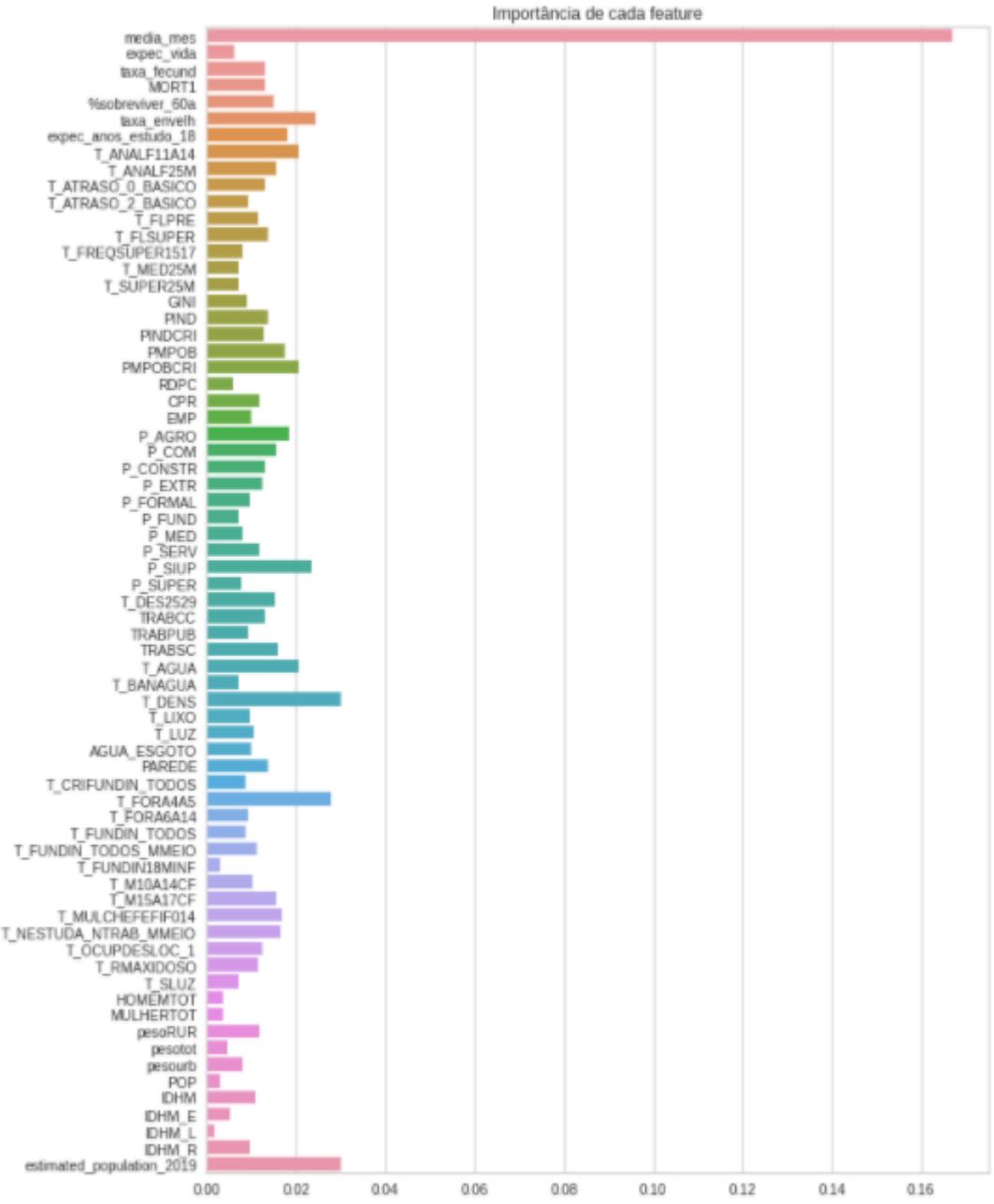

Figura 2: Atributos mais importantes para o classificador *Decision Tree com a classe de Incidência*[5]

Na classificação de incidência pode-se observar que os atributos de ***isolamento social*** e ***t_dens*** **—** descritos anteriormente — obtiveram a maior importância além dos atributos de: ***população*** estimada em 2010, ***T_FORA4A5*** (percentual de crianças de 4 a 5 anos fora da escola) e ***taxa_envelh*** (razão entre a população de mais de 65 anos e a população total multiplicado por 100).

- **Naive Bayes**

Assim como a árvore de decisão, o algoritmo bayesiano possui uma propriedade para classificação de seus atributos mais importantes. Os quais podemos visualizar na imagem abaixo.

Os atributos que se destacam na aprendizagem Bayesiana são:

1. Percentual da população que vive em domicílios com banheiro e água encanada
2. % de crianças que vivem em domicílios em que nenhum dos moradores tem o ensino fundamental completo.
3. % dos ocupados com médio completo - 18 anos ou mais
4. % de pessoas vulneráveis à pobreza e que gastam mais de uma hora até o trabalho.
5. % de pessoas de 15 a 24 anos que não estudam nem trabalham e são vulneráveis à pobreza.

- **SVM**

No treinamento o Kernel utilizado foi a Função de Base Radial (RBF), que transforma os dados em outro espaço, diferente do espaço de entrada, o que impossibilitou a obtenção dos atributos mais importantes, pois o RBF não possui compatibilidade com o atributo *_coef*, responsável pela listagem dos atributos mais importantes.

O Kernel Linear foi testado, porém o *Google Colab* possui uma limitação de tempo e processamento, o que impossibilitou o treinamento do algoritmo com essas configurações.

5. CONCLUSÃO

Como foi mostrado, os algoritmos de classificação de incidência e mortalidade possuem um bom desempenho, cumprindo o seu objetivo trabalhando em conjunto com o *dataset* gerado neste trabalho.

Portanto, os classificadores[4] criados possuem um potencial para melhorar a cada dia, desde que, através de uma boa análise, novas melhorias sejam implementadas. Como é o caso

4 https://github.com/jorgewellison/classifiers-tcc

de alguns que ainda podem ser otimizados, como o SVM, buscando uma maior capacidade de generalização.

Assim conclui-se que os resultados obtidos nos alertam sobre os impactos dentre diferentes classes sociais em nosso país e sobre a importância do isolamento social — como observamos nos atributos mais importantes: média mensal de isolamento, população residente em área urbana etc — , para que o controle de incidência e de mortalidade sejam realizados de forma eficaz por nossos governantes e por nós.

Por fim, foi possível perceber que no tocante à eficácia, na classe de **incidência** a Árvore de Decisão obtém os melhores resultados. No entanto, em relação à **mortalidade** o **Naive Bayes** obteve o melhor resultado. Pois, já que as classes (incidência e mortalidade) possuem valores desproporcionais, a métrica considerada mais importante é a F1-Score.

REFERÊNCIAS

ANVISA. Nota Técnica GVIMS/GGTES/ANVISA Nº 07/2020. **Orientações para a prevenção da transmissão de COVID-19 dentro dos serviços de Saúde**. (Complementar à Nota Técnica GVIMS/GGTES/ANVISA Nº 04/2020). 2020

BARBOSA FILHO, F. de H. (2017). **A crise econômica de 2014/2017**. Estudos avançados, 31(89), 51-60.

BRUNIALTI, Lucas; PERES, Sarajane; FREIRE, Valdinei; **Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática**. XI Brazilian Symposium on Information System, 2015.

DEPARTAMENTO DE MEDICINA DA PUC-RJ. **Sistema auxilia na padronização de análise de dados COVID-19**. Disponível em: https://www.med.puc-rio.br/notcias/2020/4/15/sistema-auxilia-na-padronizao-de-anlise-de-dados-covid-19

DUCHESNE, Pierre; RÉMILLARD, Bruno (Ed.). **Statistical modeling and analysis for complex data problems**. Springer Science & Business Media, 2005.

FAWCETT, Tom. **An introduction to ROC analysis**. Institute for the Study of Learning and Expertise, 2009.

GAMA, J. **Árvores de Decisão**, 2000. Disponível em: http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html. Acesso em: 14 ago. 2004.

HALLAL, Pedro C. et al. **SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys**. The Lancet Global Health, v. 8, n. 11, p. e1390-e1398, 2020.

HAYAT et al., **"The Pathogenesis of Coronavirus Disease 2019 (COVID-19): Evaluation and Prevention"**, Journal of Immunology Research, vol. 2020, Article ID 1357983, 7 pages, 2020.

KARYPIS, G. CLUTO – **A Clustering Toolkit. Department of Computer Science and Engineering**. University of Minnesota. Minneapolis, 2002.

MAGALHÃES S. S. A., MACHADO, C. J. (2014). **Conceitos epidemiológicos e as pandemias recentes: novos desafios**. Cadernos Saúde Coletiva, 22(1), 109-110.

MINISTÉRIO DA SAÚDE. **Coronavírus Brasil**. Disponível em: https://covid.saude.gov.br/

MITCHELL, T.M. (1997). **Machine Learning**. McGraw Hill, New York.

MONARD, M. C. AND BARANAUSKAS, J. A. **Conceitos de aprendizado de máquina**. In: REZENDE, S.O. Sistemas inteligentes: fundamentos e aplicações. Barueri, SP: Manole, cap.4, p.89-94, 2003.

ROUBAUD, François et al. **The municipios facing COVID-19 in Brazil: socioeconomic vulnerabilities, transmisssion mechanisms and public policies**. 2020.

SILVA FEM et al.. SARS-COV-2 - **Aspectos relacionados à biologia, propagação e transmissão da doença emergente COVID-19**. Revista desafios [periódicos na internet]. 7(Especial-3).2020

S. M. Peres, T. Rocha, M. R. C. B. Bíscaro, H. H., and C. Boscarioli. **Tutorial sobre fuzzy-c-means e fuzzy learning vector quantization: Abordagens híbridas para tarefas de agrupamento e classificação**. 19(1):120–163, 2012.

VAPNIK, V. **The Nature of Statistical Learning Theory**. New York: Springer-Verlag, 1995.

ZHU N, et al. **A novel coronavirus from patients with pneumonia in China**, 2019. New England Journal of Medicine 2020.