

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

João Paulo Calembo Batista Menezes

**APLICAÇÃO DE MODELOS DE MACHINE LEARNING EM OPERAÇÕES DE
FUSÕES E AQUISIÇÕES NA SAÚDE SUPLEMENTAR**

Belo Horizonte
2023

João Paulo Calembo Batista Menezes

**APLICAÇÃO DE MODELOS DE MACHINE LEARNING EM OPERAÇÕES DE
FUSÕES E AQUISIÇÕES NA SAÚDE SUPLEMENTAR**

Trabalho de Conclusão de Curso apresentado
ao Curso de Especialização em Ciência de
Dados e Big Data como requisito parcial à
obtenção do título de especialista.

Belo Horizonte

2023

LISTA DE TABELAS

Tabela 1 – Composição da base de dados.....	10
Tabela 2 – Resultados dos modelos.....	38
Tabela 3 – Variáveis mais importantes entre os modelos (Top 5)	39

LISTA DE QUADROS

Quadro 1 – Relação de empresas objeto de fusões e aquisições, entre 2007 e 2016	20
--	-----------

LISTA DE FIGURAS

Figura 1 – Importação dos dados e visualização	12
Figura 2 – Outputs dos dados – Parte 1	12
Figura 3 – Outputs dos dados – Parte 2	13
Figura 4 – Estatística descritiva	14
Figura 5 – Confirmação do desbalanceamento da amostra.....	15
Figura 6 – Análise das variáveis com dados faltantes.....	15
Figura 7 – Análise das variáveis com dados faltantes.....	16
Figura 8 – Verificação da inclusão dos dados faltantes	17
Figura 9 – Bibliotecas utilizadas para reamostragem	18
Figura 10 – Exclusão de variáveis não numéricas.....	18
Figura 11 – Aplicação da técnica SMOTE	18
Figura 12 – Carregamento da biblioteca e treinamento do modelo (XGBoost).....	27
Figura 13 – Carregamento da biblioteca e teste do modelo (XGBoost)	28
Figura 14 – Análise dos resultados do modelo (XGBoost)	29
Figura 15 – Verificação das variáveis importantes (XGBoost)	29
Figura 16 – Demonstração das variáveis importantes (XGBoost)	30
Figura 17 – Carregamento da biblioteca e treinamento do modelo (LightGBM).....	31
Figura 18 – Teste do modelo (LightGBM).....	32
Figura 19 – Ajuste dos resultados probabilísticos em binário (LightGBM)	32
Figura 20 – Análise dos resultados do modelo (LightGBM)	33
Figura 21 – Verificação das variáveis importantes (LightGBM)	33
Figura 22 – Demonstração das variáveis importantes (LightGBM)	34
Figura 23 – Carregamento da biblioteca e treinamento do modelo (Random Forest)	34
Figura 24 – Teste do modelo (Random Forest)	35
Figura 25 – Análise dos resultados do modelo (Random Forest)	36
Figura 26 – Verificação das variáveis importantes (Random Forest)	36
Figura 27 – Demonstração das variáveis importantes (Random Forest).....	37

LISTA DE GRÁFICOS

Gráfico 1 – Frequência das operações de F&A por operadoras de planos de saúde, entre 2007 e 2016	24
Gráfico 2 – Número de OPSs na amostra entre 2007 e 2016	25

SUMÁRIO

1. Introdução	7
1.1. Objetivos	8
2. Coleta de Dados	9
3. Tratamento de Dados	12
3.1 Análise estatística preliminar	12
3.2 Balanceamento dos dados	17
4. Análise e Exploração dos Dados	20
5. Criação de Modelos de Machine Learning	26
5.1 Aplicação do XGBoost	27
5.2 Aplicação do LightGBM	30
5.3 Aplicação do Random Forest	34
6. Análise e Interpretação dos Resultados.....	38
7. Links	41
REFERÊNCIAS	42

1. Introdução

O setor da saúde suplementar no Brasil foi regulado em 1998 por meio da Lei 9.656, e desde então, a área da saúde suplementar tem sido alvo de diversas operações de fusões e aquisições, principalmente devido à crescente competitividade e à busca por maiores ganhos de escala por parte dos agentes econômicos envolvidos. Contudo, a identificação de empresas propensas a essas operações ainda é um desafio dada a grande complexidade do setor. Nesse sentido, a aplicação de modelos de Machine Learning pode ser uma solução eficiente para identificar potenciais alvos de fusões e aquisições na saúde suplementar.

Os modelos de Machine Learning são capazes de analisar grandes conjuntos de dados e identificar padrões ocultos que não seriam perceptíveis por métodos convencionais. Dessa forma, é possível utilizar esses modelos para prever quais empresas têm maior probabilidade de serem alvos de operações de fusões e aquisições na saúde suplementar, levando em consideração diversos fatores, como desempenho financeiro, tamanho, características da carteira de beneficiários, dentre outros.

No entanto, é importante ressaltar que a escolha do modelo de Machine Learning adequado para esse tipo de análise pode ser crucial para uma escolha satisfatória. É necessário também considerar a complexidade dos dados a serem analisados, o tamanho do conjunto de dados, a precisão e a interpretabilidade dos resultados obtidos, além de outros fatores relevantes.

Portanto, este estudo tem como objetivo identificar o melhor modelo de Machine Learning para identificar empresas propensas a uma operação de fusões e aquisições na saúde suplementar. A partir da análise de diferentes modelos e de um conjunto de dados representativo, pretende-se avaliar a eficiência e a acurácia de cada modelo na previsão de operações de fusões e aquisições no setor da saúde suplementar, contribuindo assim para a tomada de decisão das empresas do setor e para o desenvolvimento de estratégias mais eficazes.

1.1. Objetivos

Objetivo Geral:

Identificar o melhor modelo de Machine Learning para a previsão de empresas propensas a uma operação de fusões e aquisições na saúde suplementar.

Objetivos Específicos:

- Analisar dados da Agência Nacional de Saúde Suplementar (ANS) e do Conselho Administrativo de Defesa Econômica (CADE), contemplando variáveis relevantes para a análise das Fusões e Aquisições;
- Selecionar modelos de Machine Learning para previsão de operações de fusões e aquisições;
- Avaliar a eficiência e a acurácia de diferentes modelos de Machine Learning na previsão de operações de fusões e aquisições no setor da saúde suplementar, considerando a precisão, a interpretabilidade e a complexidade dos modelos.

2. Coleta de Dados

Os dados utilizados nesse trabalho foram obtidos a partir do estudo realizado por Menezes (2019), que utilizou para a obtenção da amostra duas fontes de informações. Essas fontes de dados utilizadas nesta pesquisa incluem o Conselho Administrativo de Defesa Econômica (CADE) e a Agência Nacional de Saúde Suplementar (ANS).

No sítio do CADE, foram coletados dados sobre as operações de fusões e aquisições na área da saúde que foram objeto de análise no período de 1º de janeiro de 2000 a 31 de dezembro de 2016. Os dados brutos foram obtidos em duas etapas: em 31 de março de 2016, foram coletados os dados até 2015 e em 27 de fevereiro de 2017, foram coletados os dados referentes a 2016. Para obter as informações relevantes, foram utilizados os termos "hospital" e "saúde" na sessão "Pesquisa Processual" do sítio do CADE, selecionando os itens processos, documentos gerados e documentos externos do tipo de processo "finalístico: ato de concentração sumário" e "finalístico: ato de concentração ordinário". Os dados brutos foram analisados e organizados até 28 de abril de 2017, utilizando o software Microsoft Excel. Além disso, informações relevantes foram obtidas junto à ANS.

Os dados obtidos junto à Agência Nacional de Saúde Suplementar (ANS) foram coletados por meio de pesquisa realizada no sítio da internet e por meio do Sistema Eletrônico do Serviço de Informação ao Cidadão, em conformidade com as disposições da Lei nº 12.527, de 2011. Para acessar as demonstrações contábeis das Operadoras de Planos de Saúde (OPS), foram obtidos arquivos eletrônicos no sítio da ANS em 07/06/2018, referentes ao período de 2007 a 2016. A partir de 2007, as operadoras da modalidade seguradora especializada em saúde passaram a ser obrigadas a enviar informações via Documento de Informações Periódicas das Operadoras de Plano de Saúde (DIOPS) e não mais por meio de Formulários de Informação Periódica (FIP), conforme estabelecido pela Resolução Normativa (RN) nº 136, de 31 de outubro de 2006. Como resultado, os dados obtidos no CADE anteriores a 2007 foram descartados e, considerando que os dados dos atos de concentração obtidos no CADE estavam limitados até o ano de 2016, as bases de dados do CADE e da ANS foram delimitadas para o período de 2007 a 2016, ou seja, 10 anos.

Após a obtenção dessas informações, houve a integração das bases do CADE e da ANS, uma vez que nem todas as operações de fusões e aquisições estão sujeitas ao controle do CADE.

A partir da integração dessas bases, foi possível mapear as operações de fusões e aquisições realizadas no setor da saúde suplementar e identificar a nacionalidade dos seus investidores, se nacional ou estrangeiro. Para isso, foram levadas em consideração as informações constantes nos processos de análise do CADE e as demonstrações contábeis obtidas junto à ANS, que incluem a informação do capital social estrangeiro.

As informações financeiras e não financeiras obtidas nas demonstrações contábeis foram organizadas possibilitando a construção das variáveis utilizadas na pesquisa. A amostra final selecionada foi composta por 12.972 observações, sendo 595 relativas a OPSs que foram objeto de operações de fusões e aquisições, conforme tabela 1.

Tabela 1 – Composição da base de dados

(continua)

Nome da variável	Frequência	Descrição	Dtype
REG ANS	12.972	Código de registro na ANS	int64
DATA	12.972	Data da operação	object
IED	12.972	Investimento estrangeiro	int64
FEA	12.972	Fusões e Aquisições	int64
Fatur	12.972	Faturamento	float64
EVA	12.972	Valor econômico adicionado - <i>Economic Value Added®</i>	float64
EBITDATRAD	12.972	Lucro antes dos juros, impostos, depreciação e amortização (<i>EBITDA</i>)	float64
RSV	12.886	Resultado sobre vendas	float64
ROA	12.965	Retorno sobre o ativo total	float64
ROE	12.957	Retorno sobre o patrimônio líquido	float64
MEBITDA	12.915	Margem EBITDA	float64
PMRE	12.687	Prazo médio de recebimento de eventos	float64
PMPE	12.687	Prazo médio de pagamento de eventos	float64
Ciclo	12.687	Ciclo operacional	float64
Imob	12.972	Imobilização do ativo total	float64
PCT	11.866	Índice de endividamento	float64
CE	12.970	Composição do endividamento	float64
LC	12.773	Liquidez corrente	float64
LG	12.774	Liquidez geral	float64
ISTR	12.939	Índice de despesas assistenciais	float64
IDC	12.960	Índice de despesas comerciais	float64

(conclusão)

Nome da variável	Frequência	Descrição	Dtype
IDA	12.825	Índice de despesas administrativas	float64
TAM	12.972	Tamanho (Logaritmo natural)	float64
HHI	12.972	Herfindahl-Hirschman Index	float64
CR4	12.972	Rácio de concentração - 4 Maiores	float64
CR8	12.972	Rácio de concentração - 8 Maiores	float64
CR12	12.972	Rácio de concentração - 12 Maiores	float64
CRn	12.972	Rácio de concentração	float64
Benefassist	11.796	Beneficiários em planos de assistência médica	float64
Benefexcl	11.796	Beneficiários em exclusivamente odontológicos	float64
Beneftot	11.796	Total de beneficiários de planos de saúde	float64
TKM	11.704	Tíquete médio	float64
CMg	11.378	Custo marginal	float64
Lerner	4.609	Índice de Lerner	float64
Reclam	12.972	Reclamações de beneficiários	float64
Ireclam	11.796	Índice de reclamações de beneficiários	float64
Segmentacao	12.972	Segmento de atuação - ANS	object
Modalidade	12.972	Modalidade da OPS	object
Estado	12.972	Estado brasileiro	object

Fonte: Dados da pesquisa

Uma vez obtida a base de dados, deu-se início ao tratamento dos dados, cujo procedimento é apresentado na próxima seção.

3. Tratamento de Dados

3.1 Análise estatística preliminar

Uma vez que a base de dados estava consolidada em um arquivo .csv (*Comma-separated values*) iniciou-se a fase de processamento tratamento dos dados, para tanto foi utilizado o google Colaboratory (versão sem custo financeiro). Inicialmente, de acordo com a figura 1, foi feita a importação dos dados e visualização para análise inicial.

Figura 1 – Importação dos dados e visualização

```

import pandas as pd

# Leitura do arquivo CSV
dadosops = pd.read_csv('BaseOPS.csv', sep=";", decimal=',')
display(dadosops)
print(dadosops.info())

```

Fonte: Dados da pesquisa

Os outputs foram os seguintes, parte 1 na Figura 2.

Figura 2 – Outputs dos dados – Parte 1


	REG ANS	DATA	IED	Fatur	EVA	EBITDATRAD	RSV	ROA	ROE	MEBITDA	...	Benefitot	TKM	CMg	Lerner	Rei
0	27	01/10/2007	0	4.606816e+06	-1497034.12	4.040939e+04	0.56	0.06	0.07	0.49	...	2366.0	162.26	76.36	0.53	
1	43	01/10/2007	1	4.039453e+09	17316018.84	1.812340e+08	0.28	0.14	0.27	0.12	...	760438.0	436.36	317.80	0.27	
2	51	01/10/2007	0	0.000000e+00	-559338.91	-2.541484e+04	0.00	-0.01	-0.01	0.00	...	545311.0	0.00	0.00	NaN	
3	361	01/10/2007	0	3.738584e+08	-2585007.95	1.405217e+07	0.36	0.13	0.30	0.15	...	83703.0	372.21	287.08	0.23	
4	477	01/10/2007	0	5.089409e+08	-275577.18	1.392222e+07	0.19	0.11	0.29	0.09	...	155874.0	272.09	195.10	0.28	
...
12967	420654	01/10/2016	0	0.000000e+00	0.00	-1.895670e+04	0.00	-0.05	-0.05	0.00	...	NaN	NaN	NaN	NaN	
12968	420662	01/10/2016	0	1.000000e+00	0.00	0.000000e+00	0.00	0.00	0.00	0.00	...	NaN	NaN	NaN	NaN	
12969	420671	01/10/2016	0	0.000000e+00	0.00	-8.001000e+01	0.00	0.00	0.00	0.00	...	NaN	NaN	NaN	NaN	
12970	420689	01/10/2016	0	1.000000e+00	0.00	0.000000e+00	0.00	0.00	0.00	0.00	...	NaN	NaN	NaN	NaN	
12971	420743	01/10/2016	0	0.000000e+00	-2892.96	1.494260e+03	0.00	0.03	0.03	0.00	...	NaN	NaN	NaN	NaN	

12972 rows x 39 columns

Fonte: Dados da pesquisa

E na figura 3, a parte 2.

Figura 3 – Outputs dos dados – Parte 2



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12972 entries, 0 to 12971
Data columns (total 39 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   REG ANS                12972 non-null  int64
1   DATA                  12972 non-null  object
2   IED                    12972 non-null  int64
3   Fatur                  12972 non-null  float64
4   EVA                    12972 non-null  float64
5   EBITDATRAD             12972 non-null  float64
6   RSV                    12886 non-null  float64
7   ROA                    12965 non-null  float64
8   ROE                    12957 non-null  float64
9   MEBITDA                12915 non-null  float64
10  PMRE                   12687 non-null  float64
11  PMPE                   12687 non-null  float64
12  Ciclo                  12687 non-null  float64
13  Imob                   12972 non-null  float64
14  PCT                    11866 non-null  float64
15  CE                     12970 non-null  float64
16  LC                     12773 non-null  float64
17  LG                     12774 non-null  float64
18  ISTR                   12939 non-null  float64
19  IDC                    12960 non-null  float64
20  IDA                    12825 non-null  float64
21  TAM                    12972 non-null  float64
22  HHI                    12972 non-null  float64
23  CR4                    12972 non-null  float64
24  CR8                    12972 non-null  float64
25  CR12                   12972 non-null  float64
26  CRn                    12972 non-null  float64
27  Benefassist            11796 non-null  float64
28  Benefexcl              11796 non-null  float64
29  Beneftot               11796 non-null  float64
30  TKM                    11704 non-null  float64
31  CMg                    11378 non-null  float64
32  Lerner                 4609 non-null  float64
33  Reclam                 12972 non-null  int64
34  Ireclam                11796 non-null  float64
35  Segmentacao            12972 non-null  object
36  Modalidade             12972 non-null  object
37  Estado                 12972 non-null  object
38  FEA                    12972 non-null  int64
dtypes: float64(31), int64(4), object(4)
memory usage: 3.9+ MB
None
```

Fonte: Dados da pesquisa

Uma primeira análise estatística descritiva foi realizada, os resultados demonstrados na figura 4.

Figura 4 – Estatística descritiva

```

stats = dadosops.describe(include = 'all')
print(stats)

```

	REG ANS	DATA	IED	Fatur	EVA \
count	12972.000000	12972	12972.000000	1.297200e+04	1.297200e+04
unique	NaN	10	NaN	NaN	NaN
top	NaN	01/10/2007	NaN	NaN	NaN
freq	NaN	1438	NaN	NaN	NaN
mean	363530.854456	NaN	0.015187	1.036273e+08	-5.613817e+06
std	52624.989545	NaN	0.122299	6.134611e+08	7.136234e+07
min	27.000000	NaN	0.000000	-1.046248e+07	-3.602216e+09
25%	330264.000000	NaN	0.000000	1.378487e+06	-1.190562e+06
50%	359289.000000	NaN	0.000000	1.046298e+07	-8.372429e+04
75%	411558.000000	NaN	0.000000	4.644742e+07	5.635078e+04
max	420743.000000	NaN	1.000000	2.016843e+10	1.878906e+08

	EBITDATRAD	RSV	ROA	ROE	MEBITDA \
count	1.297200e+04	12886.000000	12965.000000	12957.000000	12915.000000
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	1.060885e+06	0.031632	-0.026063	0.101601	0.090839
std	3.936664e+07	4.661846	1.448109	3.432577	3.534064
min	-2.608769e+09	-92.520000	-79.110000	-87.350000	-90.980000
25%	-2.307747e+04	-0.010000	-0.020000	-0.010000	0.000000
50%	1.220135e+05	0.070000	0.030000	0.110000	0.050000
75%	9.787689e+05	0.270000	0.110000	0.310000	0.130000
max	1.145761e+09	95.820000	20.740000	91.070000	97.050000

	Beneftot	TKM	CMg	Lerner \
count	1.179600e+04	11704.000000	1.137800e+04	4609.000000
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN
mean	5.038758e+04	160.733511	-1.110032e+02	0.486949
std	2.397741e+05	170.260491	2.663961e+04	0.280125
min	1.000000e+00	0.000000	-2.065598e+06	0.000000
25%	3.021500e+03	33.597500	-4.535000e+00	0.250000
50%	1.032012e+04	135.350000	4.133500e+01	0.460000
75%	2.969875e+04	218.402500	2.008525e+02	0.720000
max	5.784163e+06	1980.240000	4.952308e+05	1.000000

	Reclam	Ireclam	Segmentacao	Modalidade \
count	12972.000000	11796.000000	12972	12972
unique	NaN	NaN	2	8
top	NaN	NaN	MÉDICO-HOSPITALAR	Medicina de Grupo
freq	NaN	NaN	8575	3315
mean	43.156337	7.046062	NaN	NaN
std	368.136092	87.251698	NaN	NaN
min	0.000000	0.000000	NaN	NaN
25%	0.000000	0.000000	NaN	NaN
50%	1.000000	0.890000	NaN	NaN
75%	7.000000	4.520000	NaN	NaN
max	15512.000000	9052.550000	NaN	NaN

	Estado	FEA
count	12972	12972.000000
unique	27	NaN
top	SP	NaN
freq	4714	NaN
mean	NaN	0.045868
std	NaN	0.209207
min	NaN	0.000000
25%	NaN	0.000000
50%	NaN	0.000000
75%	NaN	0.000000
max	NaN	1.000000

[11 rows x 39 columns]

Fonte: Dados da pesquisa

Considerando a característica da variável dependente, foi feita uma confirmação do grande desbalanceamento entre os dados das ocorrências em que não houve Fusões e Aquisições com aquelas em que houve, os valores atribuídos foram, respectivamente, 0 e 1, figura 5.

Figura 5 – Confirmação do desbalanceamento da amostra

```
# Verificar se há desbalanceamento
display(dadosops["FEA"].value_counts())
display(dadosops["FEA"].value_counts(normalize=True).map("{:.1%}".format))

0      12377
1        595
Name: FEA, dtype: int64
0      95.4%
1       4.6%
Name: FEA, dtype: object
```

Fonte: Dados da pesquisa

Para além do desbalanceamento, foi verificado também o quantitativo de dados faltantes na amostra, figura 6.

Figura 6 – Análise das variáveis com dados faltantes

[6] dadosops.isnull().sum()			
REG ANS	0	IDC	12
DATA	0	IDA	147
IED	0	TAM	0
Fatur	0	HHI	0
EVA	0	CR4	0
EBITDATRAD	0	CR8	0
RSV	86	CR12	0
ROA	7	CRn	0
ROE	15	Benefassist	1176
MEBITDA	57	Benefexcl	1176
PMRE	285	Beneftot	1176
PMPE	285	TKM	1268
Ciclo	285	CMg	1594
Imob	0	Lerner	8363
PCT	1106	Reclam	0
CE	2	Ireclam	1176
LC	199	Segmentacao	0
LG	198	Modalidade	0
ISTR	33	Estado	0
		FEA	0
		dtype: int64	

Fonte: Dados da pesquisa

Dada a necessidade de preenchimento dos dados faltantes da amostra, se tomou como estratégia a não exclusão de nenhuma variável para manter a fidedignidade dos dados obtidos, para resolução do problema foi aplicada interpolação de dados por meio da técnica foward fill, figura 7.

Figura 7 – Análise das variáveis com dados faltantes

```
dadosops[["RSV", "ROA", "ROE", "MEBITDA", "PMRE", "PMPE", "Ciclo", "PCT", "CE",
"LC", "LG", "ISTR", "IDC", "IDA", "Benefassist", "Benefexcl",
"Beneftot", "TKM", "CMg", "Lerner", "Ireclam"]] = dadosops[["RSV",
"ROA", "ROE", "MEBITDA", "PMRE", "PMPE",
"Ciclo", "PCT", "CE", "LC", "LG", "ISTR", "IDC", "IDA",
"Benefassist", "Benefexcl", "Beneftot", "TKM", "CMg",
"Lerner", "Ireclam"]].fillna(method='ffill')
```

Fonte: Dados da pesquisa

A interpolação é uma técnica comum usada para preencher dados faltantes em amostras com dados financeiros. A interpolação *forward fill*, em particular, é uma técnica que consiste em preencher os dados faltantes com o valor mais recente observado antes do ponto de dados faltante, (ABREU, 2021).

Existem diversas justificativas para o uso da interpolação *forward fill* em amostras financeiras com dados faltantes. Em primeiro lugar, ela é uma técnica simples e fácil de implementar, exigindo pouco processamento computacional. Além disso, essa técnica é particularmente útil em situações em que a tendência dos dados é constante ou com pequenas variações, pois os valores preenchidos são baseados em valores históricos próximos, o que mantém a tendência geral da série.

Outra razão pela qual a interpolação *forward fill* é frequentemente usada em dados financeiros é que a presença de dados faltantes pode prejudicar a análise estatística. A falta de dados pode levar a imprecisões nos cálculos de medidas estatísticas, como a média e o desvio padrão, além de introduzir viés nos resultados. A inclusão dos dados faltantes foi realizada conforme figura 8.

Figura 8 – Verificação da inclusão dos dados faltantes

<pre># Conferir se dados faltantes foram substituídos dadosops.isnull().sum()</pre>		ISTR	0
		IDC	0
		IDA	0
		TAM	0
		HHI	0
		CR4	0
		CR8	0
		CR12	0
		CRn	0
		Benefassist	0
		Benefexcl	0
		Beneftot	0
		TKM	0
		CMg	0
		Lerner	0
		Reclam	0
		Ireclam	0
		Segmentacao	0
		Modalidade	0
		Estado	0
		FEA	0
		dtype:	int64
REG ANS	0		
DATA	0		
IED	0		
Fatur	0		
EVA	0		
EBITDATRAD	0		
RSV	0		
ROA	0		
ROE	0		
MEBITDA	0		
PMRE	0		
PMPE	0		
Ciclo	0		
Imob	0		
PCT	0		
CE	0		
LC	0		
LG	0		

Fonte: Dados da pesquisa

Concluída essa fase, na seção seguinte são apresentados os aspectos relativos ao balanceamento dos dados.

3.2 Balanceamento dos dados

O desbalanceamento de classe é um problema comum em muitos conjuntos de dados de Machine Learning, onde uma ou mais classes são significativamente sub-representadas em comparação com outras. Esse desequilíbrio pode levar a modelos enviesados e imprecisos, já que o modelo tende a favorecer a classe majoritária.

O SMOTE (Técnica de Oversampling Minoritário Sintético) é uma técnica de amostragem para lidar com o problema de classes desbalanceadas em conjuntos de dados (CHAWLA *et al.*, 2002). A técnica gera dados sintéticos para a classe minoritária, através da interpolação de exemplos já existentes. Os novos exemplos são gerados na vizinhança de cada exemplo da classe minoritária, a fim de aumentar o espaço de decisão e melhorar a capacidade de generalização dos classificadores obtidos. Os exemplos sintéticos são gerados aleatoriamente ao longo do segmento de reta que une cada exemplo da classe minoritária a um de seus k-vizinhos mais próximos, selecionados de forma aleatória.

No caso específico, os dados utilizados nessa pesquisa apresentaram um desbalanceamento, uma vez que, somente 4,6% dos casos eram de fusões e aquisições. Portanto foi utilizada a técnica SMOTE para equilibrar as classes.

Foram utilizadas as bibliotecas de Python, conforme figura 9, que permitiram fazer essa reamostragem.

Figura 9 – Bibliotecas utilizadas para reamostragem

```
#Bibliotecas
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.metrics import classification_report
```

Fonte: Dados da pesquisa

Foi necessário também excluir as variáveis não numéricas, figura 10.

Figura 10 – Exclusão de variáveis não numéricas

```
# Exclusão das colunas não numéricas (Strings)
dadosops.drop(["REG ANS", "DATA", "Modalidade", "Segmentacao", "Estado" ],
              axis=1, inplace=True)
```

Fonte: Dados da pesquisa

Os resultados decorrentes da aplicação da técnica, estão apresentados na figura 11. Permitindo, portanto, uma análise balanceada.

Figura 11 – Aplicação da técnica SMOTE

```
[11] # Definição do X e y do modelo
      X = dadosops.drop(['FEA'], axis=1)
      y = dadosops['FEA']

[12] # Aplicação do smote
      smote = SMOTE(sampling_strategy='minority', random_state=42)

[13] X, y = smote.fit_resample(X, y)

[14] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                         stratify=y, random_state=42)

[15] # Verificação do balanceamento com Smote
      print(pd.Series(y_train).value_counts())
      print(pd.Series(y_test).value_counts())

      1    8664
      0    8663
      Name: FEA, dtype: int64
      0    3714
      1    3713
      Name: FEA, dtype: int64
```

Fonte: Dados da pesquisa

Como é possível verificar na figura 11, a partir do balanceamento por meio da técnica SMOTE, os dados também já foram ajustados na proporção 70 (17.327 observações) x 30 (7.427 observações), respectivamente treino e teste, para que fossem aplicados os algoritmos de aprendizagem de máquina que serão explorados na próxima seção.

4. Análise e Exploração dos Dados

Embora desbalanceados, a análise dos dados brutos pode dar alguns *insights* que auxiliem a criação dos modelos de machine learning.

Relativamente as F&A, foram identificadas na amostra 86 operadoras de planos de saúde, que estavam relacionadas a 73 operações de compra e 68 operações de venda, que totalizaram as 595 observações na amostra. A relação das operadoras objeto de operações de F&A identificadas no período da pesquisa é apresentada no quadro 1.

Quadro 1 – Relação de empresas objeto de fusões e aquisições, entre 2007 e 2016

(continua)

Reg. ANS	Razão social
400386	ADCON - ADMINISTRADORA DE CONVENIOS ODONTOLÓGICOS LTDA.
335657	ADVANCE PLANOS DE SAÚDE LTDA.
416452	AFINIDADE ADMINISTRADORA DE BENEFÍCIOS LTDA.
416771	ALIANÇA ADMINISTRADORA DE BENEFÍCIOS DE SAUDE S.A.
327107	AMESP SISTEMA DE SAÚDE LTDA.
306622	AMICO SAÚDE LTDA.
326305	AMIL ASSISTÊNCIA MÉDICA INTERNACIONAL S.A.
412384	AMIL PLANOS POR ADMINISTRAÇÃO LTDA.
411264	ASL - ASSISTÊNCIA À SAÚDE LTDA.
325767	ASSISTÊNCIA MÉDICA SÃO PAULO-SUL S/C LTDA.
419150	ASSOCIAÇÃO MAIS SAÚDE SANTA CASA DE SÃO JOÃO DA BOA VISTA
346411	BIODENT ASSISTÊNCIA ODONTOLÓGICA S/A
000051	BRADERCO DENTAL S.A.
419419	BRASIL DENTAL OPERADORA DE PLANOS ODONTOLÓGICOS S.A.
304590	CAIXA DE ASSISTENCIA A SAUDE DOS EMPREGADOS DO BEG - CASBEG
406406	CARE PLUS DENTAL LTDA.
315516	CENTRAL MÉDICA DE PREVENÇÃO LTDA.

(continua)

Reg. ANS	Razão social
392804	CENTRO CLÍNICO GAÚCHO LTDA.
350117	CLIDEC - CLÍNICA DENTÁRIA ESPECIALIZADA CURA D'ARS LTDA.
301591	DENTAL CENTER SERVIÇOS ODONTOLÓGICOS LTDA.
321991	DENTAL PLAN LTDA.
411159	DENTALCORP ASSISTÊNCIA ODONTOLÓGICA INTERNACIONAL LTDA.
415286	DIVICOM ADMINISTRADORA DE BENEFÍCIOS LTDA.
411051	EXCELSIOR MED S/A
313971	FEDERAÇÃO DAS SOCIEDADES COOPERATIVAS DE TRABALHO MÉDICO DO ACRE, AMAPÁ, AMAZONAS, PARÁ, RONDÔNIA E RORAIMA
302881	FUNDAÇÃO ASSISTENCIAL VIÇOSSENSE
312126	FUNDAÇÃO SAÚDE ITAÚ
319147	FUNDAÇÃO WALDEMAR BARNESLEY PESSOA
409197	GAMA ODONTO S/A.
407011	GAMA SAÚDE LTDA.
403911	GOLDEN CROSS ASSISTÊNCIA INTERNACIONAL DE SAÚDE LTDA.
391727	GRUPO SERVIÇOS DE MEDICINA LTDA.
368253	HAPVIDA ASSISTÊNCIA MÉDICA LTDA.
317501	INTERODONTO - SISTEMA DE SAÚDE ODONTOLÓGICA LTDA.
307408	LIFE SYSTEM ASSISTÊNCIA MÉDICA LTDA.
326933	LINCX SISTEMAS DE SAÚDE LTDA.
414697	MAXI CARE ODONTOLOGIA EMPRESARIAL S.A.
302872	MEDIAL SAÚDE S.A.
322946	MEDICAMP ASSISTÊNCIA MÉDICA LTDA.
348520	MEDISANITAS BRASIL ASSISTÊNCIA INTEGRAL À SAÚDE S/A.
333689	MEDISERVICE OPERADORA DE PLANOS DE SAÚDE S/A
406481	METLIFE PLANOS ODONTOLÓGICOS LTDA.

(continua)

Reg. ANS	Razão social
348732	MULTICARE SAÚDE LTDA.
359017	NOTRE DAME INTERMÉDICA SAÚDE S.A.
006980	NOTRE DAME SEGURADORA S/A
310981	ODONTO EMPRESAS CONVENIOS DENTARIOS LTDA.
360813	ODONTO SERV LTDA.
301949	ODONTOPREV S/A
413631	OPS - PLANOS DE SAÚDE S.A.
403300	ORALGOLD PLANOS ODONTOLÓGICOS S.A.
357294	PREVDONTO PARTICIPAÇÕES LTDA.
413267	PREVENT SENIOR CORPORATE OPERADORA DE SAÚDE LTDA.
407755	PRONTO SOCORRO INFANTIL LUIZ FRANÇA LTDA.
345571	PRONTODENTE - ODONTOLOGIA INTEGRAL LTDA.
417271	PS PADRÃO ADMINISTRADORA DE BENEFÍCIOS LTDA.
417173	QUALICORP ADMINISTRADORA DE BENEFÍCIOS S.A.
355950	SANTA CASA DE MISERICORDIA DONA CAROLINA MALHEIROS
355097	SANTA HELENA ASSISTÊNCIA MÉDICA S/A.
358509	SANTA LUZIA ASSISTENCIA MEDICA S.A.
339245	SANTAMÁLIA SAÚDE S.A.
365319	SÃO FRANCISCO ODONTOLOGIA LTDA.
302091	SÃO FRANCISCO SAÚDE SOCIEDADE EMPRESÁRIA LTDA.
338362	SEISA SERVIÇOS INTEGRADOS DE SAÚDE LTDA.
352942	SEPAO - ASSISTÊNCIA ODONTOLÓGICA EMPRESARIAL LTDA.
340332	SISTEMA IPIRANGA DE ASSISTÊNCIA MÉDICA LTDA.

(conclusão)

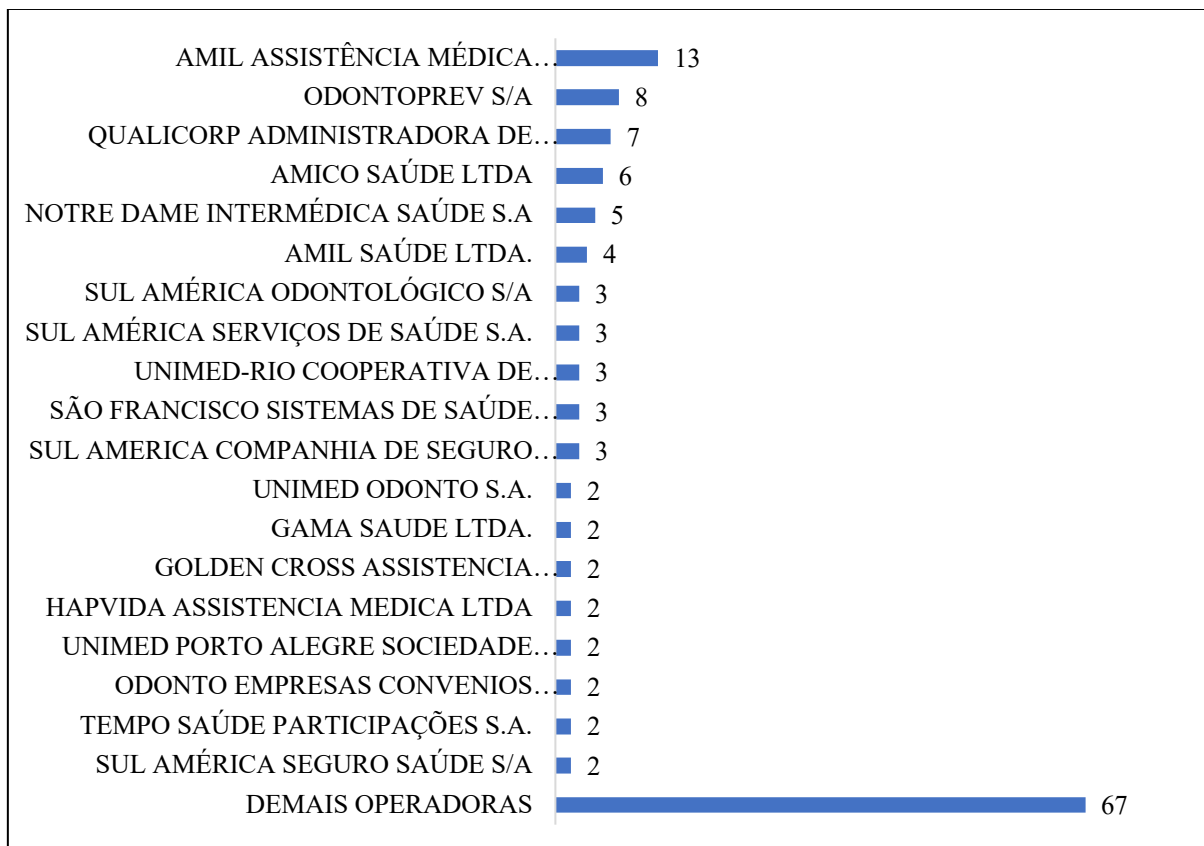
Reg. ANS	Razão social
006246	SUL AMÉRICA COMPANHIA DE SEGURO SAÚDE
417815	SUL AMÉRICA ODONTOLÓGICO S/A
005622	SUL AMÉRICA SAÚDE COMPANHIA DE SEGUROS
000043	SUL AMÉRICA SEGURO SAÚDE S.A
416428	SUL AMÉRICA SERVIÇOS DE SAÚDE S.A.
400289	SUL AMÉRICA SERVIÇOS MÉDICOS S.A.
000361	TEMPO SAÚDE PARTICIPAÇÕES S.A.
343889	UNIMED - BELO HORIZONTE COOPERATIVA DE TRABALHO MÉDICO
306398	UNIMED - COOPERATIVA DE SERVIÇOS DE SAÚDE DOS VALES DO TAQUARI E RIO PARDO LTDA.
361518	UNIMED BETIM COOPERATIVA DE TRABALHO MÉDICO
310964	UNIMED CENTRO SUL - SOCIEDADE COOP. DE TRAB. MÉDICO LTDA.
345270	UNIMED DO ABC - COOPERATIVA DE TRABALHO MÉDICO
311618	UNIMED MISSÕES/RS - COOPERATIVA DE ASSISTÊNCIA À SAÚDE LTDA.
416801	UNIMED ODONTO S.A.
352501	UNIMED PORTO ALEGRE - COOPERATIVA MÉDICA LTDA.
344885	UNIMED RECIFE COOPERATIVA DE TRABALHO MÉDICO
000701	UNIMED SEGUROS SAÚDE S.A.
357391	UNIMED VITÓRIA COOPERATIVA DE TRABALHO MÉDICO
393321	UNIMED-RIO COOPERATIVA DE TRABALHO MÉDICO DO RIO JANEIRO LTDA.
413038	VITALLIS SAÚDE S/A

Fonte: Dados da pesquisa

Sobre a frequência das ocorrências, no gráfico 1 é possível verificar as operadoras que participaram de pelo menos duas operações de F&A. Nesse sentido, das 141 identificadas no período pesquisado. A Amil Assistência Médica Internacional S.A. (326305) foi a empresa com a maior frequência de operações, tendo sido identificadas 13 operações de F&A.

Posteriormente, a Odontoprev S.A. (301949), com oito operações de F&A e Qualicorp Administradora de Benefícios S.A. (417173), com sete operações.

Gráfico 1 – Frequência das operações de F&A por operadoras de planos de saúde, entre 2007 e 2016

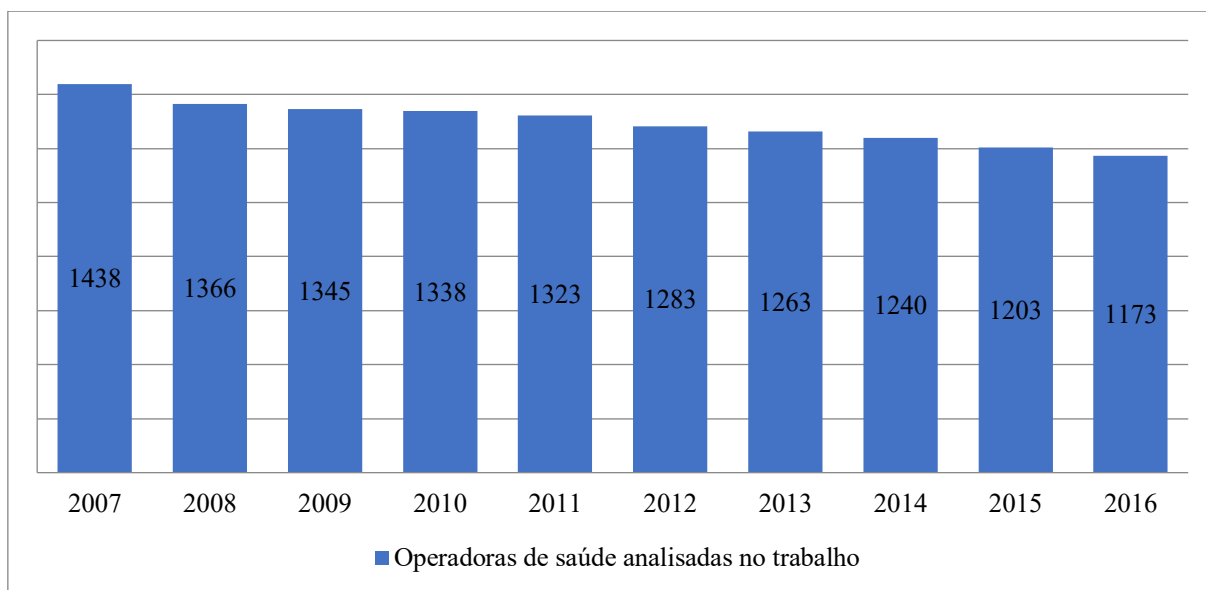


Fonte: Dados da pesquisa

A amostra integral é composta de 12.972 observações, compreendendo 1.769 OPS, entre 2007 e 2016. A distribuição das OPS no período analisado está apresentada no gráfico

2.

Gráfico 2 – Número de OPSs na amostra entre 2007 e 2016



Fonte: Dados da pesquisa

Uma análise que também foi possível refere-se ao movimento de entrada e saída de OPS na base de dados, já que nenhuma operadora foi excluída da amostra original. Assim, entre 2007 e 2016, 524 OPS deixaram de fazer parte do conjunto de dados analisados e 259 OPS foram incluídas na referida base de dados. Esse movimento indica uma redução no número total de OPS. Assim, das 1.769 OPS analisadas, 914 estão presentes em todo o período e foram identificadas 72 que entraram após 2007 e saíram antes de 2016.

Essa análise permite identificar um fator importante, relativo ao de tendência de concentração do setor, com redução do número de OPS, ainda que em um cenário de intensas operações de F&A. A construção de modelos de machine learning pode auxiliar na percepção, por exemplo, de que quais são as variáveis mais importantes na predição, além de ser também ferramenta para identificação de novas empresas *targets* para novos negócios.

Nesse propósito, na próxima seção serão abordados aspectos relativos aos modelos de machine learning que serão utilizados.

5. Criação de Modelos de Machine Learning

Para a criação de modelos de machine learning, foram escolhidos 3 modelos de aprendizado de máquina supervisionado, XGBoost, LightGBM e Random Forest.

Proposto por Chen e Guestrin (2016), o XGBoost é uma biblioteca otimizada de aumento de gradiente distribuído. Com essa biblioteca é possível implementar algoritmos de aprendizagem de máquinas sob a estrutura de Gradient Boosting, resolvendo problemas de ciências de dados de forma precisa e rápida. O LightGBM é um modelo baseado em machine learning que utiliza lógica de aprendizado por Árvore de Decisão baseado em aumento de gradiente que fornece uma nova implementação do Gradient Boosting Decision Tree (GBDT) (KE et al., 2017). Por último, o Random Forest é um método que foi proposto por Leo Breiman e Adele Cutler, (BREIMAN, 2015), e utiliza diversas árvores de classificação e regressão. A aleatoriedade é executada de duas formas. Na primeira, ao treinar cada uma das árvores, a amostra é escolhida estocasticamente de todas as amostras de treinamento por meio da técnica bootstrap e, segundo, pelo fato de que, em cada nó, um subconjunto de todos os recursos é selecionado aleatoriamente e utilizado para calcular a divisão ideal.

Os métodos de aprendizado de máquina supervisionados como XGBoost, LightGBM e Random Forest são amplamente utilizados para resolver problemas de classificação e regressão. Esses algoritmos pertencem à família de modelos de árvores de decisão, onde o conjunto de regras de decisão é derivado a partir dos dados.

Resumidamente, Random Forest é um algoritmo de conjunto de árvores de decisão que cria várias árvores em paralelo e combina as previsões, enquanto o XGBoost e o LightGBM são algoritmos de conjunto de árvores de decisão que utilizam uma técnica de aumento de gradiente para melhorar o desempenho da previsão. O XGBoost utiliza uma abordagem em cascata para criar árvores sucessivas, enquanto o LightGBM utiliza uma abordagem de construção de árvores em nível e uma técnica de amostragem para selecionar os exemplos de treinamento mais importantes. Ambos, XGBoost e LightGBM são conhecidos por sua alta precisão e velocidade, com o LightGBM sendo mais eficiente em termos de memória e velocidade.

5.1 Aplicação do XGBoost

Para a aplicação do XGBoost na amostra rebalanceada com o SMOTE, vide seção 3.2, foi necessário carregar algumas bibliotecas e na sequência executar o treinamento do modelo, conforme figura 12.

Figura 12 – Carregamento da biblioteca e treinamento do modelo (XGBoost)

```
[15] #Biblioteca
      from xgboost import XGBClassifier

[16] #Carregando o XGBoost
      model = XGBClassifier()

[17] #Treinando o modelo
      model = model.fit(X_train, y_train)
      model
```

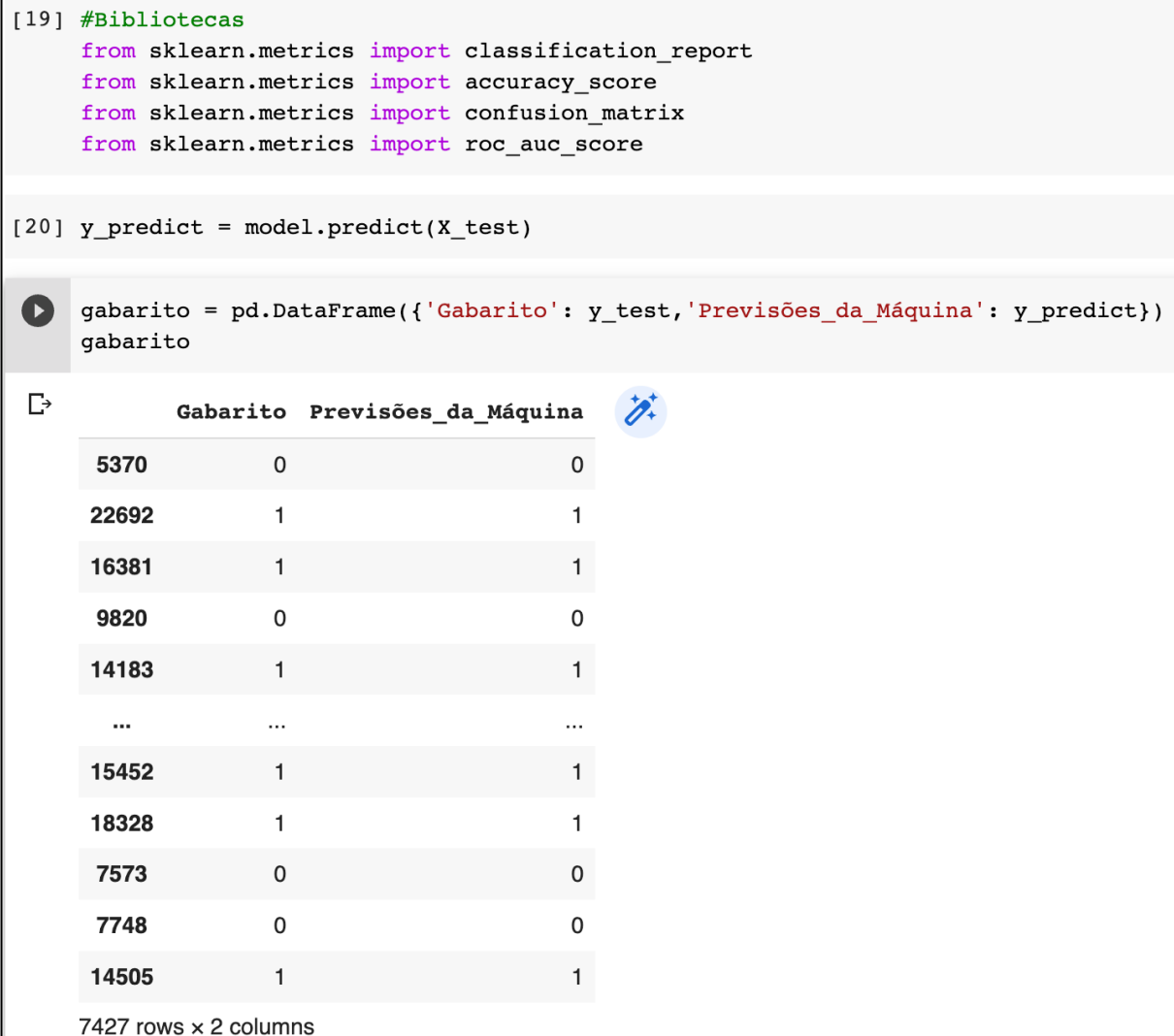
XGBClassifier

XGBClassifier(base_score=None, booster=None, callbacks=None,
 colsample_bylevel=None, colsample_bynode=None,
 colsample_bytree=None, early_stopping_rounds=None,
 enable_categorical=False, eval_metric=None, feature_types=None,
 gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
 interaction_constraints=None, learning_rate=None, max_bin=None,
 max_cat_threshold=None, max_cat_to_onehot=None,
 max_delta_step=None, max_depth=None, max_leaves=None,
 min_child_weight=None, missing=nan, monotone_constraints=None,
 n_estimators=100, n_jobs=None, num_parallel_tree=None,
 predictor=None, random_state=None, ...)

Fonte: Dados da pesquisa

Realizado o treinamento do modelo, o passo seguinte foi a importação de novas bibliotecas do sklearn.metrics, definição da parcela preditora da base e verificação por meio de gabarito dos resultados alcançados, figura 13.

Figura 13 – Carregamento da biblioteca e teste do modelo (XGBoost)



Fonte: Dados da pesquisa

Para análise dos resultados, figura 14, foram verificadas as métricas do modelo, sua acurácia, matriz de confusão e resultado da curva ROC.

Figura 14 – Análise dos resultados do modelo (XGBoost)

```

print('Classification metrics: \n', classification_report(y_test,y_predict))
print('Acurácia: \n', accuracy_score(y_test,y_predict))
print('Confusion Matrix: \n', confusion_matrix(y_test,y_predict))
print('Curva ROC: \n', roc_auc_score(y_test,y_predict))

```

Classification metrics:

	precision	recall	f1-score	support
0	1.00	0.99	1.00	3714
1	0.99	1.00	1.00	3713
accuracy			1.00	7427
macro avg	1.00	1.00	1.00	7427
weighted avg	1.00	1.00	1.00	7427

Acurácia:
0.9950181769220412

Confusion Matrix:
[[3695 19]
[18 3695]]

Curva ROC:
0.9950181949606972

Fonte: Dados da pesquisa

Por último, uma análise da importância das variáveis, o que permite uma compreensão macro da aplicação do modelo, figura 15.

Figura 15 – Verificação das variáveis importantes (XGBoost)

```

# obter as importâncias relativas das variáveis de entrada
importances = model.feature_importances_

# criar um DataFrame com as importâncias das variáveis
df_importances = pd.DataFrame({'feature': X_train.columns,
                              'importance': importances})

# ordenar o DataFrame por importância
df_importances = df_importances.sort_values('importance', ascending=False)

# imprimir o DataFrame
print(df_importances)

```

Fonte: Dados da pesquisa

Os resultados, figura 16, demonstram que a variável Beneficiários total (Beneftot) foi a mais importante na determinação do modelo preditivo.

Figura 16 – Demonstração das variáveis importantes (XGBoost)

	feature	importance			
27	Beneftot	0.268112	11	Imob	0.013372
0	IED	0.069959	7	MEBITDA	0.012148
23	CR12	0.069466	15	LG	0.010535
21	CR4	0.069442	8	PMRE	0.010350
32	Ireclam	0.069279	19	TAM	0.010237
20	HHI	0.047209	3	EBITDATRAD	0.009417
26	Benefexcl	0.037734	30	Lerner	0.009013
31	Reclam	0.036649	28	TKM	0.008335
22	CR8	0.034911	16	ISTR	0.008331
17	IDC	0.031973	4	RSV	0.008237
13	CE	0.025439	2	EVA	0.007994
9	PMPE	0.022365	12	PCT	0.007371
24	CRn	0.020621	5	ROA	0.006871
25	Benefassist	0.016678	10	Ciclo	0.005516
1	Fatur	0.015545	6	ROE	0.004280
18	IDA	0.015485	29	CMg	0.003395
14	LC	0.013734			

Fonte: Dados da pesquisa

Finalizada as análises do XGBoost, na próxima seção será analisado o LightGBM.

5.2 Aplicação do LightGBM

Na análise do LightGBM, foi utilizada também a amostra rebalanceada com o SMOTE. O carregamento das bibliotecas, LightGBM, necessárias e o treinamento do modelo são apresentados na figura 17.

Figura 17 – Carregamento da biblioteca e treinamento do modelo (LightGBM)

```
[24] #Biblioteca
import lightgbm as lgb
from lightgbm import LGBMClassifier

[25] # Light GBM
train_data=lgb.Dataset(X_train,label=y_train)
#Parametros
param = {'num_leaves':1000, 'objective':'binary', 'max_depth':7,
         'learning_rate': .01, 'max_bin':200}
param ['metric'] = ['auc', 'binary_biggloss']

#Treinando o modelo com Light GBM
num_round=50
lgbm=lgb.train(param,train_data,num_round)
```

Fonte: Dados da pesquisa

Após a realização do treinamento do modelo, foi realizada a definição da parcela preditora da base e verificação por meio de gabarito dos resultados alcançados, figura 18.

Figura 18 – Teste do modelo (LightGBM)

```
[26] # Fazendo o teste com o Light GBM
      y_predict = lgbm.predict(X_test)
```

```
# Comparando os gabaritos
gabarito = pd.DataFrame({'Gabarito': y_test, 'Previsões_da_Máquina': y_predict})
gabarito
```

	Gabarito	Previsões_da_Máquina
5370	0	0.445792
22692	1	0.675030
16381	1	0.412674
9820	0	0.457490
14183	1	0.644302
...
15452	1	0.683846
18328	1	0.690305
7573	0	0.307539
7748	0	0.307539
14505	1	0.683061

7427 rows x 2 columns

Fonte: Dados da pesquisa

Considerando o fato do output do LightGBM ser probabilístico, foi feita a conversão dos resultados, com base no tamanho da amostra, figura 19.

Figura 19 – Ajuste dos resultados probabilísticos em binário (LightGBM)

```
[28] #Conferir o tamanho da amostra
      y_predict.size
```

7427

```
#Convertendo as probabilidades em 0 e 1
for i in range(0,7427):
    if y_predict[i]>=.5: y_predict[i]=1
    else:
        y_predict[i]=0
```

Fonte: Dados da pesquisa

Para análise dos resultados, figura 20, foram verificadas as métricas do modelo, sua acurácia, matriz de confusão e resultado da curva ROC.

Figura 20 – Análise dos resultados do modelo (LightGBM)

```
# Avaliando o modelo
print('Classification metrics: \n', classification_report(y_test,y_predict))
print('Acurácia: \n', accuracy_score(y_test,y_predict))
print('Confusion Matrix: \n', confusion_matrix(y_test,y_predict))
print('Curva ROC: \n', roc_auc_score(y_test,y_predict))
```

Classification metrics:

	precision	recall	f1-score	support
0	0.91	0.95	0.93	3714
1	0.95	0.91	0.93	3713
accuracy			0.93	7427
macro avg	0.93	0.93	0.93	7427
weighted avg	0.93	0.93	0.93	7427

Acurácia:
0.9316009155782954

Confusion Matrix:
[[3543 171]
[337 3376]]

Curva ROC:
0.9315979049290641

Fonte: Dados da pesquisa

Como última etapa, foi feita a análise da importância das variáveis, o que permite uma compreensão macro da aplicação do modelo, figura 21.

Figura 21 – Verificação das variáveis importantes (LightGBM)

```
# Obtendo as importâncias relativas das variáveis de entrada
importances = lgbm.feature_importance()

# Criando um DataFrame com as importâncias das variáveis
df_importances = pd.DataFrame({'feature': X_train.columns,
                              'importance': importances})

# Ordenando o DataFrame por importância
df_importances = df_importances.sort_values('importance', ascending=False)

# Imprimindo o DataFrame
print(df_importances)
```

Fonte: Dados da pesquisa

Os resultados, figura 22, demonstram que a variável Beneficiários em exclusivamente odontológicos (Benefexcl) foi a mais importante na determinação do modelo preditivo.

Figura 22 – Demonstração das variáveis importantes (LightGBM)

	feature	importance			
		4	RSV	114	
26	Benefexcl	301	24	CRn	98
17	IDC	274	15	LG	92
1	Fatur	201	19	TAM	88
32	Ireclam	183	12	PCT	83
27	Beneftot	179	29	CMg	83
18	IDA	178	10	Ciclo	63
8	PMRE	178	7	MEBITDA	60
23	CR12	169	9	PMPE	59
28	TKM	162	30	Lerner	55
31	Reclam	153	5	ROA	51
16	ISTR	148	3	EBITDATRAD	51
13	CE	145	25	Benefassist	49
21	CR4	142	14	LC	49
20	HHI	136	0	IED	46
11	Imob	130	6	ROE	36
22	CR8	117	2	EVA	12

Fonte: Dados da pesquisa

Finalizada as análises do LightGBM, na próxima seção será analisado o Random Forest.

5.3 Aplicação do Random Forest

O carregamento da biblioteca para análise com o Random Forest foi realizado em conjunto com o treinamento do modelo, figura 23.

Figura 23 – Carregamento da biblioteca e treinamento do modelo (Random Forest)

```
[32] from sklearn.ensemble import RandomForestClassifier

[33] model = RandomForestClassifier()

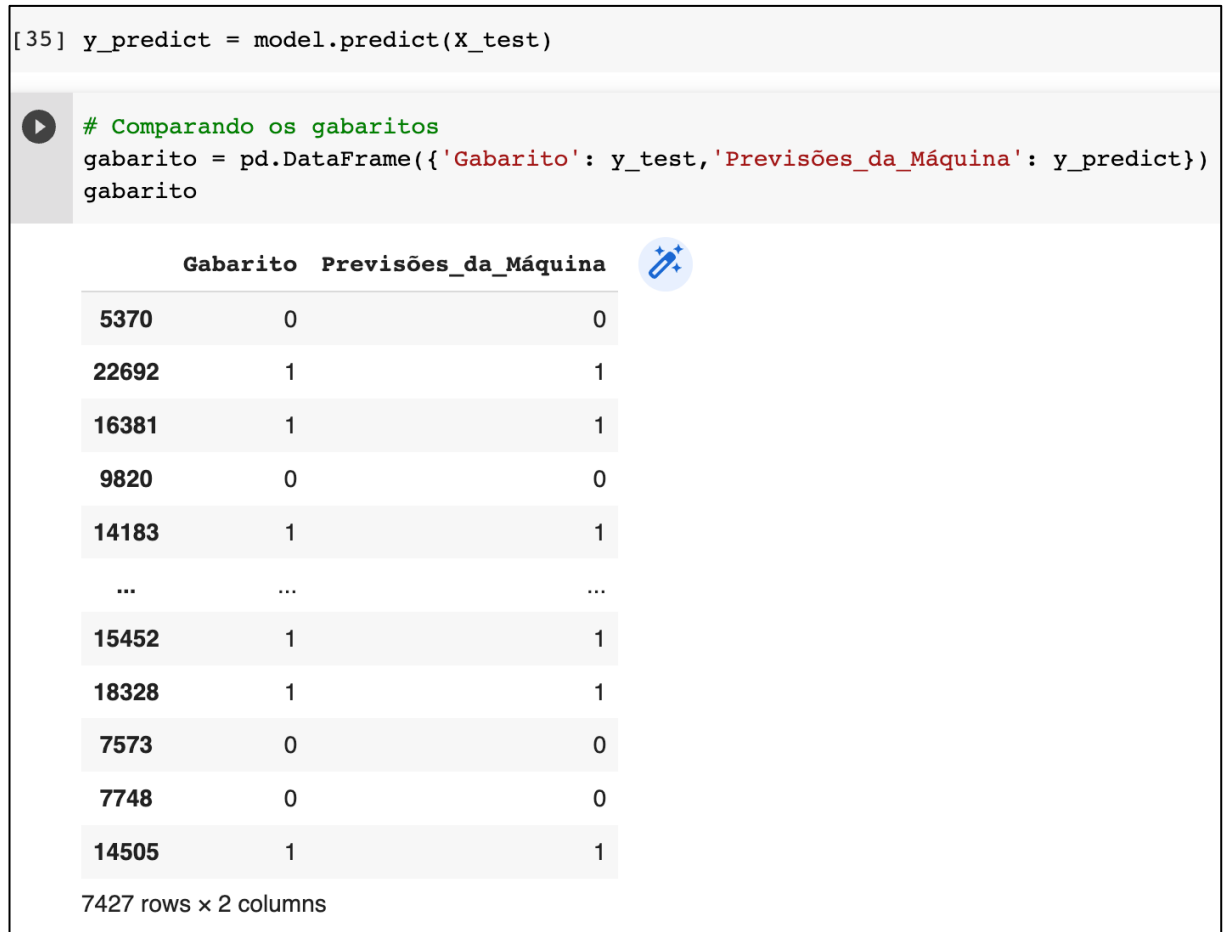
[34] model = model.fit(X_train, y_train)
      model

▼ RandomForestClassifier
RandomForestClassifier()
```

Fonte: Dados da pesquisa

Realizado o treinamento do modelo, o passo seguinte foi a definição da parcela preditora da base e verificação por meio de gabarito dos resultados alcançados, figura 24.

Figura 24 – Teste do modelo (Random Forest)



Fonte: Dados da pesquisa

Da mesma forma que nos modelos anteriores, foram verificadas as métricas do modelo, sua acurácia, matriz de confusão e resultado da curva ROC, figura 25.

Figura 25 – Análise dos resultados do modelo (Random Forest)

```

# Avaliando o modelo
print('Classification metrics: \n', classification_report(y_test,y_predict))
print('Acurácia: \n', accuracy_score(y_test,y_predict))
print('Confusion Matrix: \n', confusion_matrix(y_test,y_predict))
print('Curva ROC: \n', roc_auc_score(y_test,y_predict))

```

Classification metrics:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	3714
1	0.99	1.00	0.99	3713
accuracy			0.99	7427
macro avg	0.99	0.99	0.99	7427
weighted avg	0.99	0.99	0.99	7427

Acurácia:
0.9928638750504915

Confusion Matrix:
[[3675 39]
[14 3699]]

Curva ROC:
0.9928643281454019

Fonte: Dados da pesquisa

Como última etapa, realizou-se uma análise da importância das variáveis, o que permite uma compreensão macro da aplicação do modelo, figura 26.

Figura 26 – Verificação das variáveis importantes (Random Forest)

```

# Obter as importâncias relativas das variáveis de entrada
importances = model.feature_importances_

# Criar um DataFrame com as importâncias das variáveis
df_importances = pd.DataFrame({'feature': X_train.columns,
                              'importance': importances})

# Ordenar o DataFrame por importância
df_importances = df_importances.sort_values('importance', ascending=False)

# Imprimir o DataFrame
print(df_importances)

```

Fonte: Dados da pesquisa

Os resultados, figura 27, assim como no modelo XGBoost, demonstram que a variável Beneficiários total (Beneftot) foi a mais importante na determinação do modelo preditivo.

Figura 27 – Demonstração das variáveis importantes (Random Forest)

	feature	importance			
27	Beneftot	0.100518	18	IDA	0.016328
1	Fatur	0.080485	28	TKM	0.015628
26	Benefexcl	0.069545	9	PMPE	0.014981
31	Reclam	0.061190	15	LG	0.014517
25	Benefassist	0.059343	14	LC	0.013692
32	Ireclam	0.055166	12	PCT	0.013522
23	CR12	0.054732	16	ISTR	0.013477
20	HHI	0.042412	8	PMRE	0.013364
17	IDC	0.041124	7	MEBITDA	0.012955
19	TAM	0.040892	0	IED	0.012903
21	CR4	0.040250	10	Ciclo	0.011899
24	CRn	0.036018	5	ROA	0.011021
13	CE	0.026928	6	ROE	0.010794
2	EVA	0.024056	4	RSV	0.010675
22	CR8	0.022879	30	Lerner	0.009967
3	EBITDATRAD	0.022112	29	CMg	0.007979
11	Imob	0.018647			

Fonte: Dados da pesquisa

Finalizada a análise dos 3 modelos, na próxima seção os resultados serão discutidos para compreensão global das operações de Fusões e Aquisições no setor da Saúde Suplementar.

6. Análise e Interpretação dos Resultados

Nos três modelos analisados foi possível observar que os resultados apresentaram níveis satisfatórios de acurácia, uma vez que todos foram superiores a 90%, e ainda tanto para f1-score e recall os valores foram também superiores a 90%.

Tabela 2 – Resultados dos modelos

Métrica	XGBoost	LightGBM	Random Forest
Acurácia	0,995018	0,931600	0,992863
f1-score	1,00	0,93	0,99
Recall (0)	0,99	0,95	0,99
Recall (1)	1,00	0,91	1,00

Fonte: Dados da pesquisa

A acurácia é uma métrica que mede a proporção de previsões corretas em relação ao total de previsões. Neste caso, o XGBoost apresenta a maior acurácia, seguido pelo Random Forest e, por último, o LightGBM. Isso indica que o XGBoost tem melhor desempenho geral na classificação das observações.

O F1-score é uma métrica que combina a precisão e o recall em uma única medida, oferecendo uma melhor visão do equilíbrio entre ambos. Valores próximos a 1 indicam um melhor desempenho do modelo. Neste caso, o XGBoost alcançou a pontuação máxima, seguido pelo Random Forest e LightGBM. Isso sugere que o XGBoost possui um bom equilíbrio entre precisão e recall.

O recall é a proporção de verdadeiros positivos em relação à soma dos verdadeiros positivos e falsos negativos. Neste caso, os modelos XGBoost e Random Forest apresentam recall igual ou próximo a 1 para ambas as classes (0 e 1), indicando um excelente desempenho na identificação das classes. O LightGBM apresenta recall mais baixo para ambas as classes, sugerindo que pode estar enfrentando dificuldades em identificar corretamente as classes, principalmente a classe 1.

Em resumo, o XGBoost se destaca como o modelo com melhor desempenho geral, com a maior acurácia e F1-score, e apresentando excelente recall para ambas as classes. O Random Forest apresenta desempenho semelhante ao XGBoost, porém com acurácia e F1-score ligeiramente menores. O LightGBM apresenta desempenho inferior em comparação aos outros dois modelos nas métricas analisadas.

Outro ponto que pode ser analisado é o da importância das variáveis. Na tabela 3 temos as 5 primeiras variáveis mais importantes de cada modelo.

Tabela 3 – Variáveis mais importantes entre os modelos (Top 5)

Classificação	XGBoost	Importância	LightGBM	Importância	Random Forest	Importância
1	Beneftot	0.268112	Benefexcl	301	Beneftot	0.132633
2	IED	0.069959	IDC	274	Fatur	0.076986
3	CR12	0.069466	Fatur	201	Benefexcl	0.067074
4	CR4	0.069442	Ireclam	183	Ireclam	0.056564
5	Ireclam	0.069279	Beneftot	179	Reclam	0.055261

Fonte: Dados da pesquisa

No modelo XGBoost, a variável 'Beneftot' apresenta a maior importância, sendo significativamente maior que as demais variáveis. As outras variáveis (IED, CR12, CR4 e Ireclam) apresentam importâncias similares entre si.

Já no modelo LightGBM, a variável 'Benefexcl' é a mais importante, seguida por 'IDC' e 'Fatur'. As variáveis 'Ireclam' e 'Beneftot' apresentam importâncias menores em comparação às três primeiras. É importante notar que as escalas de importância são diferentes entre os modelos, portanto, as comparações devem ser feitas apenas dentro do mesmo modelo.

O modelo Random Forest apresentou a variável 'Beneftot' como a de maior importância, seguida por 'Fatur' e 'Benefexcl'. As variáveis 'Ireclam' e 'Reclam' apresentam importâncias menores em comparação às outras três.

Em resumo, a variável 'Beneftot' é a mais importante para os modelos XGBoost e Random Forest, enquanto 'Benefexcl' é a mais importante para o modelo LightGBM. Algumas variáveis, como 'Ireclam', também são comuns entre os modelos como importantes. Vale lembrar que a importância das variáveis pode ser específica para cada modelo e conjunto de dados, e que a análise dessas importâncias pode ajudar a entender as principais características que os modelos estão usando para fazer previsões.

Com base nos resultados da pesquisa, foi possível perceber que o melhor modelo para predição de empresas propensas a serem alvo de uma operação de Fusões e Aquisições é o XGBoost. Outro resultado importante é que a variável Total de beneficiários de planos de saúde (Beneftot) é relevante para as análises e o índice de reclamações dos beneficiários (Ireclam).

Dessa forma, foi possível a partir desse trabalho analisar os dados da ANS e do CADE, trabalhar com 3 modelos de machine learning da família de árvore de decisão e ainda analisar as métricas de cada um desses.

Como dica de investigação futura pode-se sugerir a atualização da base de dados para além de 2016 e a aplicação de outros modelos de machine learning.

Esse trabalho possui algumas limitações, dentre elas destacamos a utilização de todo o conjunto de variáveis disponíveis na base de dados, um vez que poderiam ter sido

trabalhadas variáveis selecionadas e também por não ter sido possível identificar outros trabalhos que tivessem também utilizado técnicas de aprendizagem de máquina no setor da saúde suplementar para além do trabalho de Menezes (2019).

7. Links

Conforme orientação, nessa seção são disponibilizados o link do vídeo de apresentação do trabalho e o repositório dos dados.

Link para o vídeo: <https://youtu.be/1-dFQPrRs9w>

Link para o repositório: <https://github.com/Calemb082/301007>

REFERÊNCIAS

ABREU, André Fidelis Figueiredo De. *Aplicação de machine learning na pré-seleção de ativos para portfólios de investimento*. 2021. 2021.

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS). 136. . RESOLUÇÃO NORMATIVA - RN N. 136, DE 31 DE OUTUBRO DE 2006. , 2006. Disponível em: <<http://www.ans.gov.br/component/legislacao/?view=legislacao&task=TextoLei&format=raw&id=MTEwNg==>>. Acesso em: 9 out. 2018.

BRASIL. Lei 9.656. , 6 mar. 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L9656.htm>. Acesso em: 21 dez. 2015.

BRASIL. Lei n. 12.527. . Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. , nov. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 16 maio 2017.

BREIMAN, Leo. Random forests leo breiman and adele cutler. *Random Forests-Classification Description*, v. 106, 2015.

CHAWLA, Nitesh V. *et al.* SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. 2016, [S.l: s.n.], 2016. p. 785–794.

KE, Guolin *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017.

MENEZES, JPCB. *Fusões e aquisições, concorrência e concentração: investimento estrangeiro em saúde suplementar no Brasil*. 2019. Tese (Doutorado em Administração)-Faculdade de Ciências Econômicas ..., 2019.